Modern Acoustics and Signal Processing

Geoffrey Stewart Morrison
Peter F. Assmann *Editors*

Vowel Inherent Spectral Change

Modern Acoustics and Signal Processing

Springer

Modern Acoustics and Signal Processing

For further volumes:
http://www.springer.com/series/3754

Series Preface for Modern Acoustics and Signal Processing

In the popular mind, the term "acoustics" refers to the properties of a room or other environment—the acoustics of a room are good or the acoustics are bad. But as understood in the professional acoustical societies of the world, such as the highly influential Acoustical Society of America, the concept of acoustics is much broader. Of course, it is concerned with the acoustical properties of concert halls, classrooms, offices, and factories—a topic generally known as architectural acoustics, but it is also concerned with vibrations and waves too high or too low to be audible. Acousticians employ ultrasound in probing the properties of materials, or in medicine for imaging, diagnosis, therapy, and surgery. Acoustics includes infrasound—the wind-driven motions of skyscrapers, the vibrations of the earth, and the macroscopic dynamics of the sun.

Acoustics studies the interaction of waves with structures, from the detection of submarines in the sea to the buffeting of spacecraft. The scope of acoustics ranges from the electronic recording of rock and roll and the control of noise in our environments to the inhomogeneous distribution of matter in the cosmos.

Acoustics extends to the production and reception of speech and to the songs of humans and animals. It is in music, from the generation of sounds by musical instruments to the emotional response of listeners. Along this path, acoustics encounters the complex processing in the auditory nervous system, its anatomy, genetics, and physiology—perception and behavior of living things.

Acoustics is a practical science, and modern acoustics is so tightly coupled to digital signal processing that the two fields have become inseparable. Signal processing is not only an indispensable tool for synthesis and analysis, it informs many of our most fundamental models about how acoustical communication systems work.

Given the importance of acoustics to modern science, industry, and human welfare Springer presents this series of scientific literature, entitled Modern Acoustics and Signal Processing. This series of monographs and reference books is intended to cover all areas of today's acoustics as an interdisciplinary field. We expect that scientists, engineers, and graduate students will find the books in this series useful in their research, teaching, and studies.

July 2012

William M. Hartmann
Series Editor-in-Chief

Geoffrey Stewart Morrison
Peter F. Assmann
Editors

Vowel Inherent Spectral Change

Springer

Editors
Geoffrey Stewart Morrison
Forensic Voice Comparison Laboratory,
School of Electrical Engineering
and Telecommunications
University of New South Wales
Sydney, NSW
Australia

Peter F. Assmann
School of Behavioral and Brain Sciences
University of Texas at Dallas
Richardson, TX
USA

ISBN 978-3-642-14208-6 ISBN 978-3-642-14209-3 (eBook)
DOI 10.1007/978-3-642-14209-3
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012953375

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Introduction

Peter F. Assmann and Geoffrey Stewart Morrison

Abbreviations

CVC	Consonant–vowel-consonant
DCT	Discrete cosine transform
F1	First formant
F2	Second formant
L1	First language
L2	Second language
q_1	Scaling coefficient of the first mode (parameter in Story and Bunton articulatory synthesizer)
q_2	Scaling coefficient of the second mode (parameter in Story and Bunton articulatory synthesizer)
VISC	Vowel inherent spectral change

1 Introduction

The term *vowel inherent spectral change* (VISC) was coined in Nearey and Assmann (1986). It refers to the changes in spectral properties over the time course of a vowel which are characteristic of vowel-phoneme identity. It refers not only to the widely-recognized spectral changes found in diphthongs and triphthongs, but also to the less-well-recognized spectral changes which are characteristic of

P. F. Assmann
School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, USA

G. S. Morrison (✉)
Forensic Voice Comparison Laboratory, School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia
e-mail: geoff-morrison@forensic-voice-comparison.net

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_1,

vowel-phonemes which have traditionally been called monophthongs in some dialects of some languages, particularly in North American English. Although the importance of VISC in so-called monophthongs has been recognized at least since Joos (1948, p. 101, see also his Fig. 25) and Potter and Steinberg (1950), static spectral models of vowels have predominated in phonetics research. We think, however, that this situation is changing as a larger number of researchers in more sub-branches of phonetics research are now explicitly investigating VISC or incorporating VISC as a component of their wider experimental design. In order to draw attention to and encourage wider understanding and consideration of VISC, we felt it was time to produce a volume dedicated to the topic.[1] The present volume incorporates ten chapters (besides this introduction) written by 18 authors, some of whom have been working on VISC for decades and others for whom it is newer either as a focus of investigation or as an adjunct to their established research interests. We divide the volume into four parts: *VISC Perception*, *VISC Production*, *VISC in Different Populations of Speakers*, and *VISC Applied*. Below we briefly describe these sections and the chapters within them. We hope that the chapters give a representative coverage of the breadth of work being conducted on VISC at the present time and that they will also supply readers with relatively in-depth knowledge of historical, theoretical, and empirical work on VISC.

Our initial inspiration to work on this topic came from our former PhD supervisor, Terrance M. Nearey (Fig. 1), and we dedicate this volume to him. We thank the following individuals who reviewed earlier versions of the chapters: Robert A. Fox, James M. Hillenbrand, Michael Kiefte, Keith Kluender, Richard S. McGowan, Philip Rose, Christian E. Stilp, Stephen A. Zahorian.

Postscript: As this volume was going to press in November 2012 we received the sad news that James Jenkins (one of the chapter authors) had passed away. He had a long and distinguished career, and his contributions to science will be missed. On behalf of the authors of all the other chapters in this volume, we extend our sympathies to Winifred and to his other family and friends.

2 Summary of Sections and Chapters

The first and longest section in this volume deals with *VISC perception*, and particularly theories relating the dynamic spectral properties of vowels and the perception of vowel phonemes.

Hillenbrand (2013 Chap. 2) begins by providing a historical overview of research indicating that VISC exists in many nominal monophthongs, that it is important for the perception of vowel-phoneme identity (static targets are neither

[1] We also organized a special session on VISC at the 157th Meeting of the Acoustical Society of America in Portland, Oregon, in May 2009, at which early versions of a number of the chapters from this volume were presented.

Fig. 1 Terry and Salvador at the International Congress of Phonetic Sciences 2003 Barcelona

necessary nor sufficient), and that models of dynamic spectral properties result in higher correct-classification rates and higher correlation with human-listeners' responses than models based on static measurements.

Morrison (2013a Chap. 3) then reviews studies addressing the classic hypotheses as to the perceptually relevant aspects of VISC: onset + offset, onset + slope, and onset + direction. He concludes that models based on the onset + offset hypothesis outperform the other two in terms of correct-classification rates and correlation with human-listeners' perceptual responses, including their pattern of vowel confusions.

Nearey (2013 Chap. 4) first summarizes the patterns of VISC that occur in several dialects of North American English. He then uses statistical models of formant trajectories, including in consonant–vowel-consonant (CVC) contexts, to address the question of whether VISC can be distinguished from the effects of coarticulation with adjacent consonants. Results indicate that more complex models than earlier two-point onset + offset models are needed, but that VISC is an independent aspect of vowel identity, not just a reflex of consonantal coarticulation.

Strange and Jenkins (2013 Chap. 5) review the studies that led them to posit their *Dynamic Specification* model of vowel perception. They observed that correct-identification rates are higher for vowels in consonant context than for vowels in isolation, and for silent-center/edge-only vowels than for center-only/silent-edge vowels. This led them to propose that the most important cues to vowel identity were in the spectro-temporal patterns of the consonant-vowel and vowel-consonant formant transitions. They also review a number of cross-language studies and propose that languages with more crowded vowel spaces use dynamic spectral patterns to maintain contrasts (adaptive dispersion).

Finally in this section on VISC perception, Kluender, Stilp, and Kiefte (2013 Chap. 6) discuss VISC within a broader information-theoretic efficient-coding framework for understanding speech perception. With information measured as cochlea-scaled spectral entropy (unpredictable change), the relatively low-frequency components of vowels (e.g., first and second formants, F1 and F2) provide high psychoacoustic potential information, which is further increased when there is spectral change (e.g., VISC). They also discuss VISC as a means of adaptive dispersion, helping to maintain psychoacoustic contrasts between vowel phonemes—if sensorineural systems optimize sensitivity to change then phonemes are better described in terms of how they contrast with other phonemes in the inventory rather than in terms of prototypes.

The second section, *VISC production*, consists of one chapter by Story and Bunton (2013 Chap. 7). They use a software-implemented articulatory synthesizer based on three-dimensional magnetic resonance imaging of human vocal tracts. The model is parsimonious, requiring three functions of distance from the glottis to describe the vocal tract of a particular speaker, and two parameters (q_1 and q_2) to describe the instantaneous configuration of the vocal tract. There is a nearly one-to-one mapping between these parameters and F1 and F2. Making the values of q_1 and q_2 time dependent leads to an articulatory synthesizer which can map vowel-inherent changes in vocal tract shape over the time course of a vowel to vowel-inherent changes in the acoustic spectrum. Human listeners' correct-identification rates for the vowels synthesized with dynamic vocal-tract patterns (and hence dynamic spectral patterns) were much higher than for those with static vocal tracts.

The third section looks at *VISC in different populations of speakers*, in speakers of different dialects of American English, in adult speakers versus child speakers of different ages, and in second-language versus first-language speakers of American English.

VISC has been largely neglected in dialectology. Jacewicz and Fox (2013 Chap. 8) examine cross-dialectal differences in the nominal monophthongs /ɪ/, /ɛ/, and /æ/ in three dialects of American English, central Ohio (Midland), southern Wisconsin (Inland North), and western North Carolina (South), and diachronic changes in those dialects. They conclude that VISC plays a central part in the differentiation of American English dialects and that there are systematic cross-generational changes in VISC patterns associated with ongoing sound changes (vowel shifts).

Assmann et al. (2013 Chaps. 1 and 9) summarize the results of a developmental study of VISC in children ranging in age from five through 18 years. They report reliable patterns of VISC throughout the age range studied, and note that these patterns are largely preserved across a wide range of formant frequency variation associated with age and sex differences. Statistical pattern recognition tests indicate that vowels are well-classified when formant frequency measurements are taken at two sample points around 20 and 70 % of the vowel duration. The optimum locations for onset + offset sampling of the formant trajectory do not vary substantially as a function of sex and age, and adding a third sample point does not lead to significantly better classification scores.

Rogers et al. (2013 Chap. 10) examine the production and perception of VISC by second-language (L2) learners of English (primarily first-language, L1, American Spanish learners of American English). They summarize a series of experiments that use natural and modified syllables to compare vowel identification performance by monolingual English speakers with early and late L2 learners of English. Compared to L1 listeners, both early and late L2 listeners' perception is more sensitive to manipulations that disrupt time-varying formant changes in vowels, and these difficulties may contribute to the overall greater difficulties they face when attending to speech in noisy environments. VISC enhances phonetic contrast, and in production L1 and L2 speakers may adopt different strategies in using VISC to maintain separation between otherwise neighboring vowels.

The last section, *VISC applied*, consists of one chapter by Morrison (2013b Chap. 3) in which he reviews the use of VISC in forensic voice comparison. He provides theoretical motivation as to why measurements of VISC might be good features for forensic voice comparison, and critically reviews procedures which have been used to extract information from VISC. He concludes that parametric curve fitting is the best procedure, many of the others having theoretical or practical flaws. He also empirically demonstrates that a procedure based on fitting parametric curves to formant trajectories outperforms an onset + offset model.

References

Assmann, P.F., Nearey, T.M., Bharadwaj, S.V.: Developmental patterns in children's speech: Patterns of spectral change in vowels. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 9. Springer, Heidelberg (2013)

Hillenbrand, J.M.: Static and dynamic approaches to understanding vowel perception. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 2. Springer, Heidelberg (2013)

Jacewicz, E., Fox, R.A.: Cross-dialectal differences in dynamic formant patterns in American English vowels. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 8. Springer, Heidelberg (2013)

Joos, M.A.: Acoustic phonetics. Language Supplement. **24**, 1–136. Stable http://www.jstor.org/stable/522229 (1948)

Kluender, K.R., Stilp, C.E., Kiefte, M.: Perception of vowel sounds within a biologically realistic model of efficient coding. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 6. Springer, Heidelberg (2013)

Morrison, G.S.: Theories of vowel inherent spectral change: A review. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 3. Springer, Heidelberg (2013a)

Morrison, G.S.: Vowel-inherent spectral change in forensic voice comparison. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 11. Springer, Heidelberg (2013b)

Nearey, T.M.: Vowel inherent spectral change in the vowels of North American English. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 4. Springer, Heidelberg (2013)

Nearey, T.M., Assmann, P.F.: Modeling the role of vowel inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986). doi:10.1121/1.394433

Potter, R.K., Steinberg, J.C.: Toward the specification of speech. J. Acoust. Soc. Am. **22**, 807–820 (1950). doi:10.1121/1.1906694

Rogers, C.L., Glasbrenner, M.M., DeMasi, T.M., Bianchi, M.: Vowel-inherent spectral change and the second-language learner. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 10. Springer, Heidelberg (2013)

Strange, W., Jenkins, J.J.: Dynamic specification of coarticulated vowels: Research chronology, theory, and hypotheses. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 5. Springer, Heidelberg (2013)

Story, B.H., Bunton, K.: Simulation and identification of vowels based on a time-varying model of the vocal tract area function. In: Morrison, G.S., Assmann, P.F. (eds.), Vowel Inherent Spectral Change, Chap. 7. Springer, Heidelberg (2013)

Part I
VISC Perception

Static and Dynamic Approaches to Vowel Perception

James M. Hillenbrand

Abstract The goal of this chapter is to provide a broad overview of work leading to the view that vowel inherent spectral change (VISC) plays a significant role in vowel perception. The view that implicitly guided vowel perception research for many years was the idea that nearly all of the information that was needed to specify vowel quality was to be found in a cross section of the vowel spectrum sampled at a reasonably steady portion of the vowel. A good deal of evidence shows that this static view is incomplete, including: (1) measurement data showing that most nominally monophthongal English vowels show significant spectral change throughout the course of the vowel; (2) pattern recognition studies showing that vowel categories are separated with far greater accuracy by models that take spectral change into account than otherwise comparable models using features sampled at steady-state; (3) perceptual experiments with "silent center" vowels showing that vowel steady-states can be removed with little or no effect on vowel intelligibility; and (4) perceptual experiments with both naturally spoken and synthetic speech showing that vowels with stationary spectral patterns are not well identified.

Abbreviations

VISC	Vowel inherent spectral change
F0	Fundamental frequency
F1	First formant
F2	Second formant
F3	Third formant

J. M. Hillenbrand (✉)
Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, US
e-mail: james.hillenbrand@wmich.edu

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_2,

1 Introduction

This chapter provides a broad overview of work examining the role of vowel inherent spectral change (VISC) in the recognition of vowel identity. Although seldom explicitly stated, the view that guided vowel perception research for many years was the idea that nearly all of the information that was needed to specify vowel quality was to be found in a single cross section of the vowel spectrum sampled at a reasonably steady portion of the vowel. There is now a considerable body of evidence, most of it based on the study of North American English vowels, showing that VISC plays an important role in the recognition of vowel identity. Evidence comes from a variety of experimental techniques, including: (1) measurement data showing that many nominally monophthongal English vowels show significant spectral change throughout the course of the vowel; (2) statistical pattern recognition studies showing that vowel categories can be separated with greater accuracy, and better agreement is seen with labeling data from human listeners, when the underlying measurements incorporate spectral change; (3) perceptual experiments with "silent center" vowels showing that static vowel targets can be removed or obscured by noise with little or no effect on vowel intelligibility; and (4) perceptual experiments with both naturally spoken and synthetic speech signals showing that vowels with stationary spectral patterns are not well identified.

The starting point in this discussion will be familiar to most readers. Figure 1 shows first and second formant (F1 and F2) measurements for vowels in /hVd/ syllables recorded by Peterson and Barney (1952; hereafter PB) from 33 men, 28 women, and 15 children. The formants were sampled at steady-state, i.e., "… a part of the vowel following the influence of the [h] and preceding the influence of the [d] during which a practically steady state is reached …" (p. 177). The /hVd/ syllables were presented in random order to 70 listeners with no training in phonetics. Listeners were asked to circle one of ten key words corresponding to the monophthongal vowels /i, ɪ, ɛ, æ, ɑ, ɔ, ʊ, u, ʌ, ɚ/. The listening-test results were quite simple: signals were identified as the vowel intended by the talker just over 94 % of the time. The difficulty, of course, is reconciling the excellent intelligibility of these signals with the extensive crowding and overlap that is seen in the F1−F2 measurements of Fig. 1. It is obvious, then, that listeners must be attending to features other than (or in addition to) F1 and F2 at steady-state. Many possibilities have been explored over the years, but the one that is central to the topic of this book is related to the fact that the measurements were made at a single time slice. Figure 2, from PB, shows spectrograms and spectral slices from ten /hVd/ syllables. The arrows show the times at which the acoustic measurements were made. The decisions that were made about locating the steadiest point in the vowel seem reasonable enough, but it can also be seen that several of these vowels show quite a bit of spectral movement (see especially /ɪ/, /ɛ/, /æ/, /ʊ/, and /ʌ/). It turns out that PB, along with many of their contemporaries, were well aware of the limitations of representing vowels with a single time slice. The passage below is very nearly the last thing that PB say in their paper:

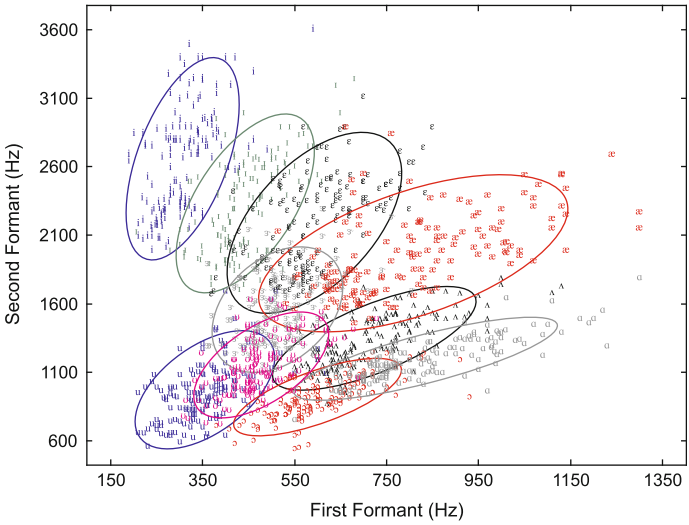


Fig. 1 Formant frequency measurements from Peterson and Barney (1952)

> It is the present belief that the complex acoustical patterns represented by the words are not adequately represented by a single section, but require a more complex portrayal. The initial and final influences often shown in the bar movements of the spectrogram are of importance here. The evaluation of these changing bar patterns … is, of course, a problem of major importance in the study of the fundamental information bearing elements of speech. (p. 184)

Below is a very similar comment from Tiffany (1953):

> It has been commonly assumed or implied that the essential physical specification of a vowel phoneme could be accomplished in terms of its acoustic spectrum as measured over a single fundamental period, or over a short interval including at most a few cycles of the fundamental frequency. That is to say, each vowel has been assumed to have a unique energy vs. frequency distribution, with the significant physical variables all accounted for by an essentially cross-sectional analysis of the vowel's harmonic composition. (p. 290)

Tiffany goes on to argue that this single-cross-section view is almost certainly too simplistic. Similar remarks can be found in Potter and Steinberg (1950) and Stevens and House (1963). Looking back at these comments, which appeared in widely read and influential papers going back to 1950, it is curious that the potential role of spectral change did not receive any steady attention for roughly another 30 years.

2 Measurement Data

Figure 3, from Nearey and Assmann (1986), shows F1 and F2 values measured from the beginnings and ends of ten Western Canadian vowels spoken in isolation by five men and five women. The initial formant measurements were taken from

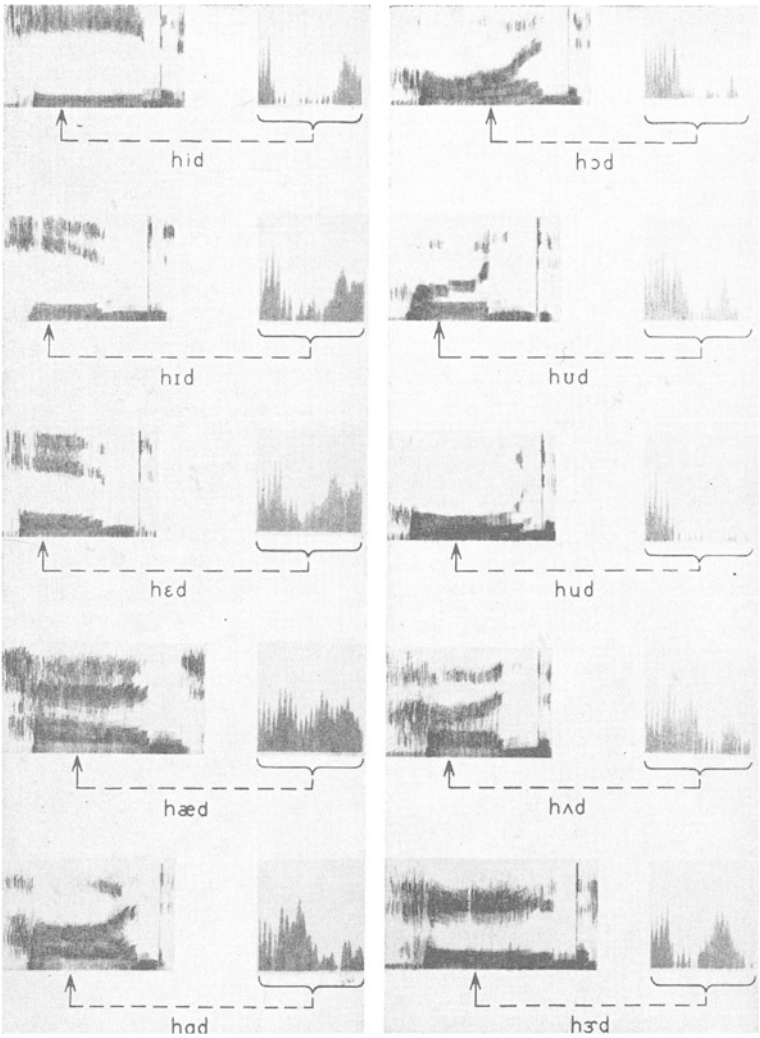


Fig. 2 Spectrograms and spectral slices from Peterson and Barney (1952)

the first frame with measurable formants and overall amplitude within 15 dB of peak amplitude, and the final formant measurements were taken from the last frame with measurable formants and overall amplitude within 15 dB of peak amplitude. It can be seen that there are a few vowels that do not show much spectral movement, especially /i/ and /u/. The phonetic diphthongs /e/ and /o/ show exactly the kinds of offglide patterns one would expect. The main thing to notice, though, is that the amount of spectral change for /e/ and /o/ is no greater than it is for the nominally monophthongal vowels /ɪ/, /ɛ/, and /æ/. Figure 4 shows similar data for speakers from the Upper Midwest, predominantly Southern Michigan

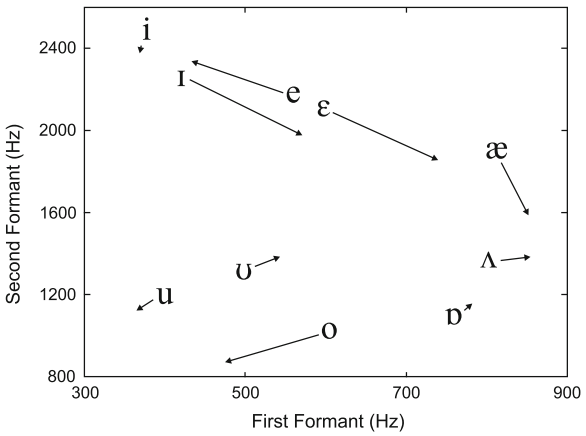


Fig. 3 Formant frequencies measured from the beginnings and ends of ten Western Canadian vowels spoken in isolation by five men and five women (Nearey and Assmann 1986). *Arrow heads* indicate vowel offsets

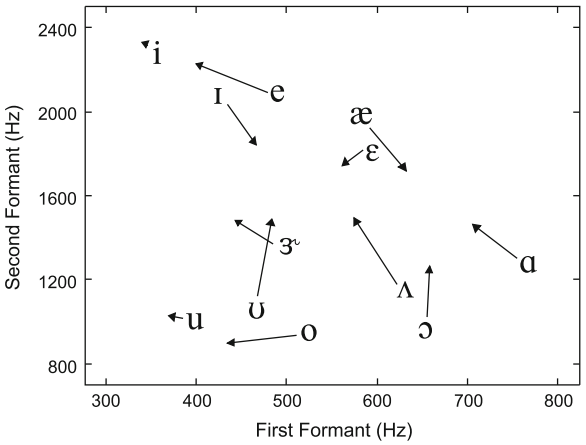


Fig. 4 Formant frequency measurements measured at 20 and 80 % of vowel duration for vowels in /hVd/ syllables spoken by 45 men. *Arrow heads* indicate vowel offsets. Speakers are from the Upper Midwest, predominantly Southern Michigan. From Hillenbrand et al. (1995)

(Hillenbrand et al. 1995, hereafter H95). The figure shows formants sampled at 20 and 80 % of vowel duration. The data are from /hVd/ syllables spoken by 45 men, but spectral change patterns for women and children are quite similar. There are some differences from the Western Canadian vowels, but there are quite a few features in common: /i/ and /u/ are more-or-less stationary, and there are several nominally monophthongal vowels that show roughly as much spectral movement as /e/ and /o/. The /æ/-/ɛ/ pair is of special interest: /æ/ is raised and fronted in this dialect, placing the two vowels almost on top of one another (see Fig. 5). Listeners in this study, however, rarely confused /æ/ and /ɛ/, with intelligibility at ∼94–95 % for the two vowels. High intelligibility is maintained in spite of the large overlap in part because of duration differences between the two vowels (see Sect. 5), but differences in spectral change patterns may also play a role. Figure 5 also shows extensive overlap between /e/ and /ɪ/, yet listeners almost never confused the two vowels with one another. But note the highly distinctive spectral

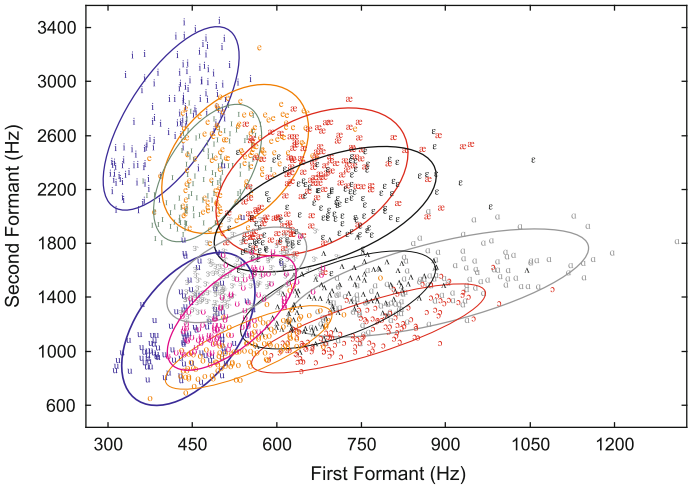


Fig. 5 Formant frequencies measured at steady state from Hillenbrand et al. (1995)

change patterns for /e/ and /ɪ/ in Fig. 4. These differences in spectral-change patterns are even more exaggerated in the Western Canadian data in Fig. 3.

3 Pattern Recognition

Evidence from pattern recognition studies shows that vowels can be classified with greater accuracy, and better agreement is seen between recognition-model output and human listener data, when the recognition model takes spectral movement into account. Zahorian and Jagharghi (1993) recorded CVC syllables with nine vowel types, nine initial consonants, and eight final consonants (not fully crossed) from 10 men, 10 women, and 10 children. The signals were analyzed using both formants and a cepstrum-like spectral-shape representation based on the discrete cosine transform (see Morrison 2013a Chap. 3, for details on this method). For both formant and spectral-shape representations, classification accuracy was consistently higher when feature trajectories were used as the basis for categorization as compared to static feature vectors. The authors also reported better agreement between model outputs and confusion matrices derived from human listeners when classification was based on dynamic rather than static features.

Hillenbrand et al. (1995) used a quadratic discriminant classifier to categorize 12 vowel types (the ten PB vowels plus /e/ and /o/) in /hVd/ syllables spoken by 45 men, 48 women, and 46 children using various combinations of fundamental frequency (F0) and the three lowest formant frequencies. The recognition model was trained on spectral features sampled once at steady-state, twice (at 20 and 80 % of vowel duration), and three times (at 20, 50 and 80 % of vowel duration).

Table 1 Accuracy in categorizing vowels using a quadratic discriminant classifier trained on one, two, or three samples of various combinations of F0 and formant frequencies

Parameters	1 samples	2 samples	3 samples
F1, F2	76.1	90.3	90.4
F1-F3	84.6	92.7	93.1
F0, F1, F2	82.0	92.5	92.6
F0, F1-F3	87.8	94.1	94.8

From Hillenbrand et al. (1995)

A sampling of the results is shown in Table 1. A substantial improvement was seen in classification accuracy from a single sample to two samples. However, adding a third sample at the center of the vowel produced little additional benefit, which might suggest that vowel identity is associated primarily with information in the onsets and offsets of vowels (see Morrison 2013a Chap. 3, for further discussion of these findings).

Hillenbrand et al. (2001) made recordings from six men and six women producing eight vowel types (/i, ɪ, ɛ, æ, ɑ, ʊ, u, ʌ/) in isolation and in CVC syllables consisting of all combinations of seven initial consonants (/h, b, d, g, p, t, k/) and six final consonants (/b, d, g, p, t, k/). As with an earlier study by Stevens and House (1963), there were many highly reliable effects of phonetic environment on vowel formants. While most of the context effects were small to modest in absolute terms, a few were quite large, especially an upward shift in F2 of ~500 Hz for men and nearly 700 Hz for women for /u/ in the environment of initial alveolars, and an upward shift in F2 of ~200 Hz in men and ~250 Hz in women for /ʊ/ in the environment of initial alveolars. Despite these context effects, vowel intelligibility was quite good in all phonetic environments. For example, the full range of variation in average intelligibility across the 42 consonant environments was only ~6 percentage points (~91–97 %), with a standard deviation of just 1.7. The listeners, then, were clearly not bothered much by variation in consonant environment. The question is whether spectral movement would continue to aid in the separation of vowel categories in spite of the acoustic complications introduced by variation in phonetic environment. In tests with a quadratic discriminant classifier, consistently better category separability was found for a variety of different combinations of F0, formants, and vowel duration when using two samples of the formant pattern as compared to a single sample at steady-state. It was also found that many aspects of the human listener data could be modeled reasonably well with a simple classifier incorporating F0, duration, and two discrete samples of the formant pattern.

Finally, Hillenbrand and Houde (2003) evaluated a spectral-shape model of vowel recognition that is quite different from the discrete cosine model described by Zahorian and Jagharghi (1993). Briefly, each vowel type is represented as a sequence of smoothed spectral-shape templates derived empirically by averaging narrow band spectra of tokens spoken by different talkers at similar times

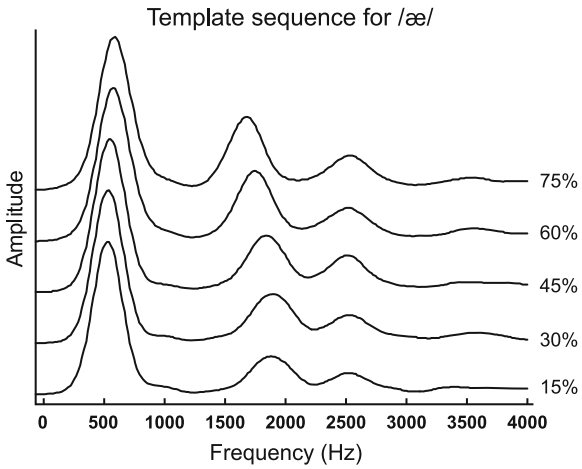


Fig. 6 Formant template sequence for /æ/ in the Hillenbrand and Houde (2003) vowel recognition model. For clarity, the templates are offset vertically. From Hillenbrand and Houde (2003)

throughout the course of the vowel. The standard version of the model represents each vowel category as a sequence of five spectral-shape templates sampled at equally spaced intervals between 15 and 75 % of vowel duration (Fig. 6). Input signals are represented by a sequence of five narrow band spectra at these same time points (15, 30, 45 %, etc.). A simple city-block distance measure is used to compare the sequence of narrow-band input spectra with template sequences for each vowel category (see Fig. 7). The vowel is recognized as the template type that produces the smallest accumulated difference over the sequence of templates. The model was trained and tested on the H95 vowels, using separate template sets for men, women, and children. Of special relevance to the present discussion, recognition performance for the standard five-slice version of the model was compared to models using: (a) a single slice (five separate tests at each of the time points; i.e., 15, 30, 45 %, etc.), (b) two slices (at 15 and 75 % of vowel duration), and (c) three slices (at 15, 45 and 75 % of vowel duration). Vowel recognition accuracy varied between 75.5 and 80.4 % for single-slice versions of the model, with performance being somewhat better for slices taken near the center of the vowel rather than at the margins. Performance improved quite sharply to 90.6 % for two slices, but little further improvement was seen for three (91.6 %) and five slices (91.4 %). Consistent with the H95 pattern recognition results, information about vowel onsets and offsets seems to be the most critical. However, both sets of tests were carried out with citation-form syllables in the fixed and neutral /hVd/ environment. The situation may not be this simple with connected speech and more complex phonetic environments.

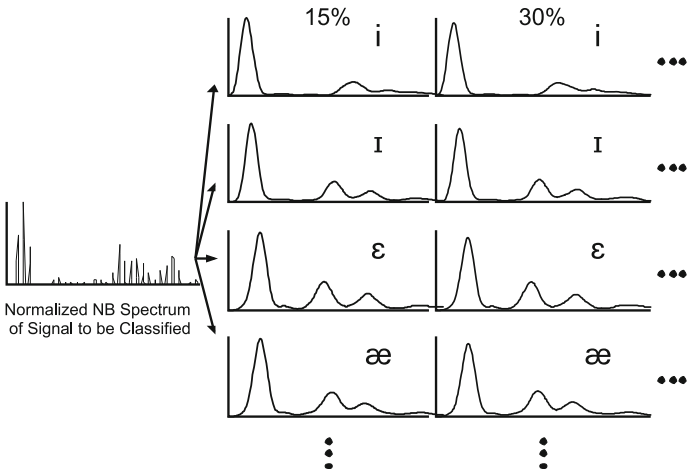


Fig. 7 Illustration of the recognition algorithm used in the Hillenbrand and Houde (2003) model. The normalized *narrow* band spectrum computed at 15 % of vowel duration (shown to the *left*) is compared to templates for each vowel category computed at the same time point (only four of twelve vowel categories are shown here). The *narrow* band spectrum at 30 % of vowel duration (not shown) is then compared to all vowel templates computed at 30 % of vowel duration, and so on. From Hillenbrand and Houde (2003)

4 Listening Studies 1: Static Targets are Not Necessary

The evidence discussed to this point is based on data analysis methods of one sort or another. The obvious question is whether listeners make use of these spectral movements in judging vowel identity. Evidence from several sources indicates that static vowel targets are neither necessary nor sufficient for the recognition of vowel identity. Figure 8, which addresses the first of these points, is from Jenkins et al. (1983), one of several silent-center studies carried out at the University of Minnesota. The original, naturally spoken /bVb/ syllables are shown in the top row. For the silent center signals, the middle 50 % has been replaced by silence for short vowels, and the middle 65 % has been replaced by silence for long vowels. The last row shows the vowel centers alone, with the onglides and offglides replaced by silence. The main finding is that the silent center signals were identified just as well (92.4 %) as the full syllables (93.1 %). The vowel centers were also well identified (86.8 %; error rates for the three conditions shown in Fig. 8 were statistically indistinguishable), but the main point is that the acoustic segments associated with presumed vowel targets are not needed. A second experiment showed, among other things, that initial segments presented in isolation and final segments presented in isolation were not nearly as well identified as the silent center stimuli.

Figure 9 is from a follow-up study by Nearey and Assmann (1986), who created test signals from 30 ms Hamming-windowed segments excised from the beginnings and ends of vowels spoken in isolation (A = 'nucleus' at 24 % of vowel

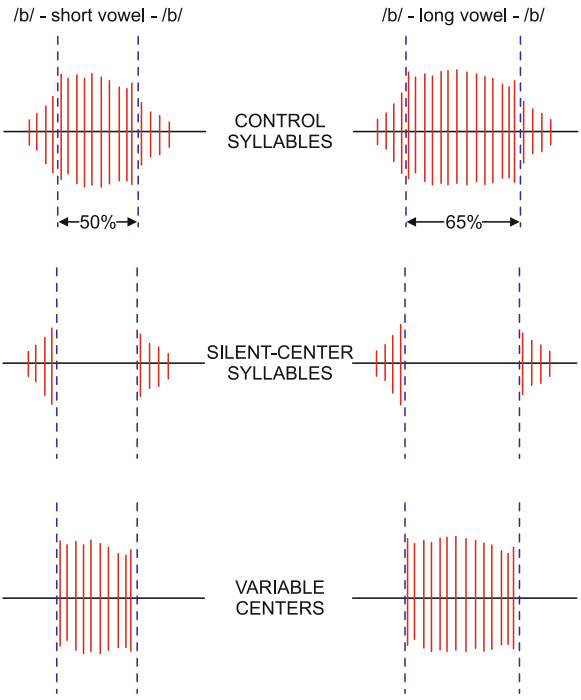


Fig. 8 Control, silent-center, and variable-center conditions from Jenkins et al. (1983)

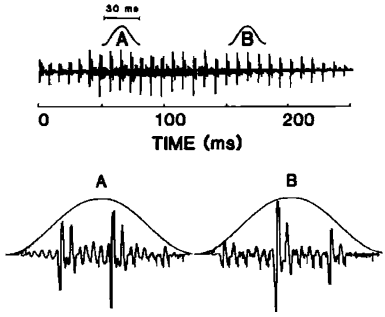


Fig. 9 Stimulus conditions from Nearey and Assmann (1986). Waveform A is a 30 ms Hamming-windowed segment of an isolated vowel centered at 24 % of vowel duration while waveform B is a 30 ms windowed segment centered at 64 % of vowel duration

duration, and B = 'offglide' at 64 % of vowel duration). Listeners heard these segments played back-to-back, with 10 ms of silence in between. Subjects were asked to identify the full syllables, the two segments played in their original order, the initial segment repeated, and the two segments played in reverse order. They found that the segments played in natural order were as intelligible as the full syllables (14.4 % versus 12.4 % error rate); however, error rates were more than twice as high for the repeated nucleus (31.0 %) and for the segments played in reverse order (37.5 %).

There is a good deal more to the literature on edited signals of these kinds (see especially Andruski and Nearey 1992; Jenkins and Strange 1999; Parker and

Diehl 1984; Strange 1989; Strange et al. 1983, 1994), but the primary conclusion to be drawn from this work is that static vowel targets are not necessary, which is fortunate since static targets do not exist for most American English vowels. Brief clips taken from the start and end of a vowel seem to be sufficient for good intelligibility.

5 Listening Studies 2: Static Targets are Not Sufficient

Evidence shows that static vowel targets are not merely unnecessary, they are also insufficient to support the very high levels of intelligibility reported in studies such as PB. One of the more compelling pieces of evidence comes from a study by Fairbanks and Grubb (1961) that received relatively little attention even in its day. The authors recorded nine sustained, static monophthongal vowels (/i, ɪ, ɛ, æ, ʌ, ɑ, ɔ, ʊ, u/)—vowels which they believed could be "… produced without ambiguity in the steady state" (p. 203). The talkers were seven men, all of them faculty in the Speech and Hearing Department at the University of Illinois. The authors went to great lengths to record high quality examples of each vowel. Reproduced below is a brief excerpt from a lengthy set of instructions to the talkers that extended for more than a full journal page.

> Essentially what we are trying to do is to collect samples of each vowel that are as nearly typical or representative of that vowel as possible. More specifically, we are interested in samples that depict the central tendency of each vowel … Another way of putting the problem is to say what we want you to do is to imagine the target on the basis of your experience in listening to speech, and then demonstrate what the target is by producing a vowel of your own that hits the target as you imagine it. You will understand from this that we are trying to get samples that are something more than merely acceptable and identifiable. (p. 204)

Two examples of each vowel were recorded by each talker. The experimenter and the talker listened to each vowel immediately after it was recorded. Talkers typically made several attempts before accepting the recording. The full set of recordings was then auditioned and the speaker was invited to make yet another attempt at any recording that was unsatisfactory for any reason. From these recordings, which were initially ~1–2 s in length, 300 ms segments were excised and presented to listeners for identification. The listeners were eight phonetically trained graduate students from the same department.

With that rather detailed setup, the results were simple: in spite of the elaborate steps that were taken to obtain the highest quality and most representative vowel samples, the average intelligibility was just 74 %. Intelligibility varied sharply across vowel categories. Intelligibility was the highest for /i/ and /u/ (91–92 %), but much lower for /ɪ/, /æ/, /ʌ/, and /ɛ/ (53–66 %). It is probably not a coincidence that the most intelligible vowels were those that show the least spectral movement among Upper Midwest speakers, and the least intelligible vowels were those that

typically show the most spectral movement (see Fig. 4). Further, it is likely that the absence of duration variability also played a role.

It is hard to overstate how striking these findings are. The test signals were spoken by just seven talkers, all of them men, and all phonetically trained. The listeners too were phonetically trained. Of greatest importance, prior to the listening tests all of the signals had been certified by two phonetically trained listeners as not merely identifiable but as representative examples of the vowel category. The listening experiment, then, was not really an identification task in the traditional sense. In effect, listeners were being asked to provide confirmation of labeling judgments that had already been made by two trained listeners. Despite all of this several of the vowel types were misidentified on anywhere from about third to a half of the presentations, and *six of the nine vowel types elicited error rates greater than* 25 %. By contrast, PB, using untrained listeners and untrained talkers, asked subjects to identify ten vowel types spoken by 33 men, 28 women, and 15 children, with formants varying *on average* by a factor of about 1.35 from men to children, and fundamental frequencies varying, again on average, approximately an octave from men to children. Yet the highest error rates reported by PB were 13.0 % (/ɑ/) and 12.3 % (/ɛ/). Similarly, in H95, with 12 vowel types spoken by 139 talkers, about evenly divided among men, women, and children, the only vowel category to elicit a double-digit error rate was /ɔ/, at 13.0 %. The most obvious explanation for the much poorer identifiably of the Fairbanks and Grubb vowels as compared to the vowels recorded in these two /hVd/ studies is the absence of cues related to spectral change and vowel duration.

There is evidence from other sources indicating that vowels with stationary formant patterns elicit much higher error rates than those with natural spectral change patterns. For example, Hillenbrand and Gayvert (1993) used a formant synthesizer to generate 300 ms static versions of all 1,520 signals in the PB database using the original measurements of F0 and F1−F3. One set of signals was synthesized with monotone pitch and a second set was synthesized with a sigmoid-shaped falling contour. Seventeen speech and hearing undergraduates with some training in phonetics served as listeners. Identification results are summarized in Table 2. For comparison, intelligibility figures reported by PB for the original signals are also shown. It can be seen that the flat-formant synthesized signals are roughly as intelligible as the Fairbanks and Grubb vowels, with intelligibility averaging ∼73 % for the monotone versions and ∼75 % for the versions with falling F0 contours. The improvement with falling pitch, shown by all 17 listeners, was highly significant but quite small in absolute terms. The main finding, though, is that the intelligibility of the flat-formant vowels is some 20 percentage points below that of the original signals from which they were synthesized. As with Fairbanks and Grubb, there were very large differences in intelligibility across different vowel types. The most intelligible vowels were /i/, /u/, and /ɚ/ (∼87–96 %) and the least intelligible were /ɑ/, /ʊ/, /æ/, and /ɛ/ (∼55–66 %). Again, vowels that typically show the least spectral change were the most intelligible, and vice versa. Although the limited intelligibility of the static resynthesized PB vowels is almost certainly due at

Table 2 Percent correct identification of flat-formant resynthesized versions of the Peterson and Barney (1952) vowels with either flat or falling F0 contours

Vowel	Flat F0	Falling F0	Original signals
/i/	96.2	95.4	99.9
/ɪ/	67.0	76.8	92.9
/ɛ/	65.8	60.9	87.7
/æ/	63.2	64.2	96.5
/ɑ/	55.0	51.0	87.0
/ɔ/	67.2	71.6	92.8
/ʊ/	62.0	72.8	96.5
/u/	89.1	84.6	99.2
/ʌ/	74.7	79.0	92.2
/ɚ/	86.6	91.7	99.7
Mean	72.7	74.8	94.4

For comparison, identification results are shown for the original stimuli. From Hillenbrand and Gayvert (1993)

least in part to the lack of spectral change cues, it is very likely that the absence of durational information also played a role.[1]

A follow up study by Hillenbrand and Nearey (1999) addressed the same basic question using a 300-signal subset of the H95 /hVd/ syllables. The 300 signals were selected at random from the full database, but with the following constraints: (1) the 12 vowel types were equally represented, and (2) signals were excluded that showed either an unmeasurable formant (e.g., a merger of F1 and F2 for vowels such as /ɔ/ or /u/, or a merger of F2 and F3 for vowels such as /i/), or with an identification error rate of 15 percentage or greater. Listeners identified three versions of each utterance: (1) the naturally spoken syllable, (2) a formant-synthesized version generated from the original measured formant contours, and (3) a flat-formant version with the formants fixed at their steady-state value (see Fig. 10). Note that, unlike Fairbanks and Grubb and the PB resynthesis study discussed above, duration information was preserved in the resynthesized signals (within the limits of the 5 ms synthesis frame rate).

Results showed 95.5 % intelligibility for the naturally spoken signals, 88.5 % for the synthesized versions generated from the original formant contours, and 73.8 % for the flat-formant versions. The ~7 % point drop in intelligibility from the original signals to the original-formant signals is important and almost certainly reveals, in part at least, the loss of information relevant to vowel quality that occurs when speech signals are reduced to a formant representation (see Bladon 1982; Bladon and Lindblom 1981). However, the finding that is most directly relevant to the present topic is the very large drop in intelligibility of nearly 15 percentage points from the original-formant to the flat-formant signals. As was the case with Fairbanks and Grubb (1961) and Hillenbrand and Gayvert (1993),

[1] PB did not measure vowel duration, making it impossible to run the obvious follow-up study using static vowels resynthesized at their measured durations.

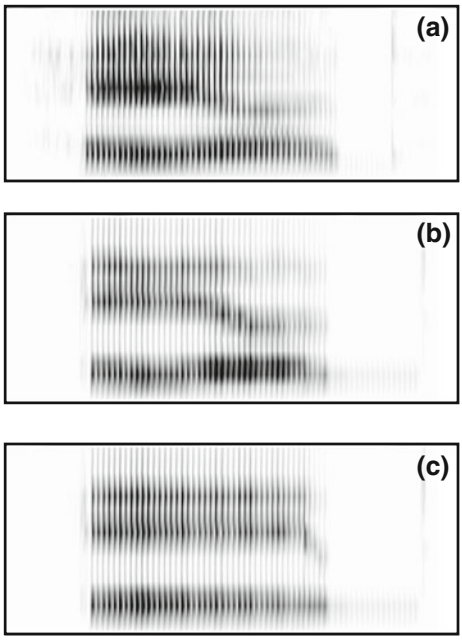


Fig. 10 Three stimulus conditions from Hillenbrand and Nearey (1999). **a** A naturally spoken signal. **b** A formant synthesized version of the same signal generated from the original measured formant contours. **c** A formant synthesized signal generated with formants fixed at the value measured at steady-state

vowels that tend to show relatively large amounts of spectral change were more strongly affected by formant flattening. The main lesson is that spectral change matters a great deal to listeners, even when duration cues are preserved.

6 The Role of Vowel Duration

It is difficult to understand the role of VISC without also considering the influence that vowel duration might have on the perception of vowel identity. As shown in Table 3, English has many pairs of spectrally similar vowels that differ in typical duration. The vowel pairs in the table are listed in order of the size of the difference in typical duration for stressed vowels based on connected speech data from Crystal and House (1988) (see generally similar connected speech data from van Santen 1992 and data from isolated syllables in Black 1949; Peterson and Lehiste 1960; and Hillenbrand et al. 1995). The obvious question is whether listeners make use of duration in identifying vowels.

Several experiments have examined the role of duration in vowel identification. For example, Ainsworth (1972) synthesized two-formant vowels with static formant frequencies covering the English vowel space. The vowels were generated in isolation and in /hVd/ syllables and varied in duration from 120 to 600 ms. Results indicated that listeners were influenced by duration in a manner that was generally consistent with observed durational differences between vowels; for example,

Table 3 Pairs of adjacent American English vowels differing in typical duration shown in parentheses are average duration ratios for vowels in stressed syllables based on connected speech data from Crystal and House (1988)

/e/	>	/ɪ/	1.81
/i/	>	/ɪ/	1.59
/æ/	>	/ɛ/	1.50
/u/	>	/ʊ/	1.48
/ɑ/	>	/ʌ/	1.36
/e/	>	/ɛ/	1.28
/ɔ/	>	/ɑ/	1.06

(Speakers were from the Mid-Atlantic region.)

a vowel in the /u/-/ʊ/ region was more likely to be identified as /u/ if long and /ʊ/ if short, and a vowel in the /ɑ/-/ʌ/ region was more likely to be heard as /ɑ/ if long and /ʌ/ if short. Ainsworth also reported that high vowels (e.g., /i/-/ɪ/, /u/-/ʊ/) were less affected by duration than other spectrally similar vowel pairs that differ in inherent duration. Similar results were reported by Tiffany (1953); Bennett (1968); and Stevens (1959).

Other experiments, however, have produced more equivocal findings. For example, Huang (1986) presented listeners with nine-step continua contrasting a variety of spectrally similar vowel pairs at durations from 40–235 ms. While the expected duration-dependent boundary shifts occurred, duration differences much larger than those observed in natural speech were needed to move the boundaries. For duration differences that approximate those found in natural speech, boundary shifts were small or nonexistent. Somewhat equivocal duration results were also reported by Strange et al. (1983). Listeners were presented with three kinds of silent center stimuli: (1) durational information retained (i.e., onglides and offglides separated by an amount of silence equal to the duration of the original vowel center), (2) durational information neutralized by setting the silent intervals for all stimuli equal to the shortest vowel, and (3) durational information neutralized by setting the silent intervals for all stimuli equal to the longest vowel. Results were mixed: shortening the silent interval to match the shortest vowels did not increase error rates relative to the natural duration condition, but lengthening the intervals to match the longest vowels produced a significant increase in error rates. Strange et al. speculated that the results for the lengthened signals may have been "… due to the disruption of the integrity of the syllables, rather than misinformation about vowel length; that is, subjects may not have perceived a single syllable with a silent gap in it, but instead, heard the initial and final portions as two discrete utterances" (Strange 1989, p. 2140).

Hillenbrand et al. (2000) used a synthesizer to linearly stretch or contract 300 /hVd/ signals drawn from H95 (the same 300 signals used in Hillenbrand and Nearey 1999). A sinusoidal synthesizer similar in conception to McAuley and Quatieri (1986) was used to generate three versions of each syllable: (1) an original-duration version, (2) a short version with duration set two standard

Table 4 Changes in vowel identity resulting from vowel shortening by human listeners and by a pattern recognition model trained on F0, two samples of the formant pattern, and duration (see text)

	Listeners	Model
/ɔ/ → /ɑ/ or /ʌ/	43.0	54.2
/æ/ → /ɛ/	20.7	25.0
/ɑ/ / → /ʌ/	9.4	8.0
/e/ → /ɪ/ or /ɛ/	2.2	4.0
/i/ → /ɪ/	0.6	0.0
/u/ → /ʊ/	0.0	0.0

The percentages reflect the number of shifts in vowel identity (original duration to short duration) divided by the number of opportunities for such shifts to occur
The three rows toward the top of the table show the most common shifts in vowel category involving adjacent vowels differing in inherent duration, and the three rows toward the bottom show shifts in vowel category that rarely occurred in spite of systematic differences in intrinsic duration. From Hillenbrand et al. (2000)

deviations below the grand mean across all vowels, and (3) a long version with duration set two standard deviations above the mean.[2]

The original-duration signals were highly intelligible (96.0 %), showing that the synthesis method did an adequate job of preserving information relevant to vowel identity. Intelligibility dropped to 91.4 % for the short signals and to 90.9 % for the long signals. This nearly symmetrical drop in intelligibility resulting from shortening or lengthening is relatively modest overall, but it turns out that this ~5 % point figure is arrived at by averaging some cases in which duration does not matter at all with other cases in which it matters a good deal.

Table 4 shows the effects of vowel shortening. (See the column labeled "Listeners"; the "Model" column will be discussed below.) Shown toward the top of the table are the vowels that were most likely to change identity from the original-duration to the short-duration version of the same token: /ɔ/ shifting to /ɑ/ or /ʌ/, /æ/ shifting to /ɛ/, and /ɑ/ shifting to /ʌ/. In the bottom of the table are some shifts that might have been expected based on production data but hardly ever occurred. For example, of the 350 presentations of the short version of each vowel (14 listeners × 25 tokens of each vowel type), there were just seven cases of short /e/ shifting to either /ɪ/ or /ɛ/, only two cases of short /i/ shifting to /ɪ/, and not a single case of short /u/ shifting to /ʊ/. Results for lengthened vowels are not shown in the table, but these effects are generally in line with the findings in Table 4, with the obvious exception that the arrows are facing in the opposite direction.

Why do listeners show sensitivity to duration for some groups and pairs of vowels that differ systematically in production (/ɔ/-/ɑ/-/ʌ/ and /æ/-/ɛ/) but not others (/e/-/ɪ/-/ɛ/, /i/-/ɪ/, and /u/-/ʊ/)? There is no obvious relationship between the size of the

[2] A neutral-duration condition was also run in which duration was fixed at the grand mean calculated across all utterances. Results from this condition do not add a great deal to the story, so to simplify the presentation these findings will be omitted.

perceptual effect and the size of the duration differences that are observed in production data. The /i/-/ɪ/ and /u/-/ʊ/ pairs in particular show production differences that are quite large, yet listeners were almost entirely unaffected by large changes in the durations of these vowels. By contrast, observed duration differences in production among the /ɔ/-/ɑ/-/ʌ/ cluster are more modest (particularly /ɔ/-/ɑ/), yet listeners showed a good deal of sensitivity to duration for these vowels.

We believe that there is a reasonably straightforward explanation for these apparently counterintuitive findings. In our view, listeners give little weight to duration for vowel contrasts such as /i/-/ɪ/ and /u/-/ʊ/ that can be distinguished with little ambiguity based entirely on spectral properties. On the other hand, vowel contrasts such as /ɔ/-/ɑ/-/ʌ/ and /æ/-/ɛ/ show a greater degree of overlap in their spectral characteristics, causing listeners to rely on duration to a greater degree in identifying these vowels. In support of this idea are the results of pattern recognition tests showing that findings broadly similar to the perceptual results described above can be modeled with a simple discriminant classifier trained on duration, F0, and formant trajectories. Using measurements from the H95 database, a quadratic classifier was trained on duration, F0, and F1–F3 sampled at 20 and 80 % of vowel duration. To simulate listening tests using shortened and lengthened signals, the pattern recognizer was then tested on the same spectral measurements that were used in training the model but with duration set to either: (1) the original measured value, (2) the value used to generate the short signals, or (3) the value used to generate the long signals. Overall, the recognition model showed somewhat more sensitivity to duration than the listeners, but there were several important features in common with the perceptual findings. As was the case for the listeners, the most frequent changes in vowel classification for the shortened vowels were /ɔ/ shifting to /ɑ/ or /ʌ/, /æ/ shifting to /ɛ/, and /ɑ/ shifting to /ʌ/ (see right-most column of Table 4), and the most frequent changes in vowel classification for the lengthened vowels were the mirror image: /ʌ/ shifting to /ɑ/ or /ɔ/ and /ɛ/ shifting to /æ/ (again, in that order). Finally, the classifier output showed no duration-dependent shifts involving either /i/-/ɪ/ or /u/-/ʊ/, and a relatively small number of shifts involving /ɪ/-/e/-/ɛ/.

In summary, evidence from several listening experiments suggests that vowel duration plays a modest role overall in vowel identification. However, the perceptual influence of vowel duration varies substantially across individual vowel categories. The relative weight that listeners give to vowel duration seems to be influenced by the degree to which a given vowel can be distinguished from its neighbors based on spectral characteristics. It should be noted that all of the evidence that has been discussed here is based on vowels either in isolation or in citation-form syllables. In connected speech there is an exceedingly large, diverse, and often competing set of demands imposed on segment durations in addition to the intrinsic duration of the vowel (see Klatt 1976 for a review). It is not clear what role duration might play in vowel perception when this feature is controlled not just by vowel identity but also by factors such as speaking rate, emphatic stress, word- and phrase-final lengthening, lexical stress, and the phonetic characteristics of neighboring speech sounds.

7 Conclusions

In a preliminary analysis of the PB vowels, Potter and Steinberg (1950) provided one of the few explicit descriptions of the static view of vowel perception, noting that the acoustic information specifying vowel identity, "… is expressible in two dimensions, frequency and amplitude. When samples of a given vowel are identified by ear as the same vowel, there must be some frequency-amplitude relationship that enables a correct identification. The problem is one of recognizing these relationships … for the different vowels." (p. 809). However, later in the same paper the authors comment, "It should be noted … that we are representing a vowel by a single spectrum taken during a particular small time interval in its duration. Actually, a vowel in the word situation ∧ undergoes transitional movements from initial to final consonant. Not only does the spectrum of the vowel change with time, but the ear in identifying the word has the benefit of all of the changes." (p. 815). The primary conclusion to be drawn from the evidence that has been reviewed here is that the cues to vowel identity for North American English vowels are not, in fact, expressible in a single time slice. The transitional movements referred to by Potter and Steinberg do, in fact, play a critical role in the recognition of vowel identity, as does the duration of the vowel. The evidence shows: (1) all but a few nominally monophthongal vowels show a significant amount of spectral movement throughout the course of the vowel, even when those vowels are spoken in isolation; (2) those spectral change patterns aid in the statistical separation of vowels in both fixed and variable phonetic environments; (3) static vowel targets are not necessary for vowel identification, nor are they sufficient to explain the very high levels of vowel intelligibility reported in studies such as PB and H95; and (4) vowel duration plays an important secondary role in vowel perception, although the influence of this feature appears to be quite uneven across individual vowels. The long tradition of representing vowels in terms of their static cross-sectional characteristics and, more importantly, of conceptualizing vowels as static points in phonetic/acoustic/perceptual space (rather than trajectories through that space), remains in widespread use. This static view is a convenient simplification that is useful for some purposes; however, it is a simplification with some important liabilities that are not always properly appreciated.

It would be easy to get the impression from this review that nearly all problems related to the role of VISC in vowel perception have been worked out. There remain a number of aspects of this problem that are not well understood. For example, an issue that has received attention only recently has to do with the role of spectral change in characterizing dialect differences. It is well known that the most prominent phonetic variations across English dialects have to do with differences in vowel production. As shown in Fig. 11, a common approach to characterizing differences in monophthong production across dialects involves plotting F1 and F2 at steady-state. This figure shows measurements from three dialects of American English: the (predominantly) Mid-Atlantic dialect from PB, the Michigan vowels from H95, and unpublished data from /hVd/ syllables spoken

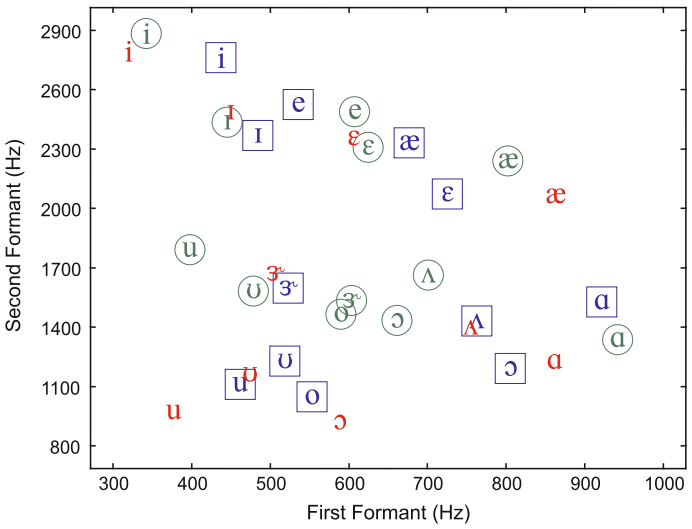


Fig. 11 Formant frequencies at steady-state for three dialects of American English: **a** Mid-atlantic (*non-enclosed* phonetic symbols), **b** Michigan (phonetic symbols enclosed in *squares*), and **c** Memphis, Tennessee (phonetic symbols enclosed in *circles*). Measurements are shown for women only

by 19 talkers from Memphis, Tennessee. (For simplicity, measurements are shown for women only. The patterns for men are quite similar.) Differences in vowel production across the three dialect groups are quite easy to see in this static representation. The description in Fig. 11 is not wrong or misleading so much as it is incomplete. Figure 12 shows the spectral change patterns for the Michigan and Memphis talkers (measurements for the PB vowels were made at steady-state only). It can be seen that some vowels show quite similar spectral change patterns and appear to be simply shifted in phonetic space relative to one another (e.g., /i/, /ɪ/, /o/, and /u/). However, other vowels (e.g., /e/, /ɔ/, /ʌ/, /ʊ/, /æ/, and /ɑ/) appear to show quite different patterns of spectral movement as well (see also Fox and McGory 2007). These kinds of issues are discussed in much greater detail in Nearey (2013 Chap. 4) and Jacewicz and Fox (2013 Chap. 8).

Perhaps the most pressing remaining problem has to do with understanding how both talkers and listeners negotiate the competing demands of VISC and coarticulation. A few conclusions seem reasonable based on current evidence. First, although coarticulation produces a large number of reliable effects on the spectral characteristics of vowels (most of them modest in size, but a few quite large—Stevens and House 1963; Hillenbrand et al. 2001), listeners have little difficulty accurately identifying vowels that are spoken in a wide variety of phonetic environments (Hillenbrand et al. 2001). Second, spectral change patterns of the kinds that can be seen in Figs. 3 and 4 cannot be attributed entirely to coarticulation since these patterns are observed in isolated vowels as well as CVCs (see Nearey 2013 Chap. 4), and because listeners have been found to rely on these

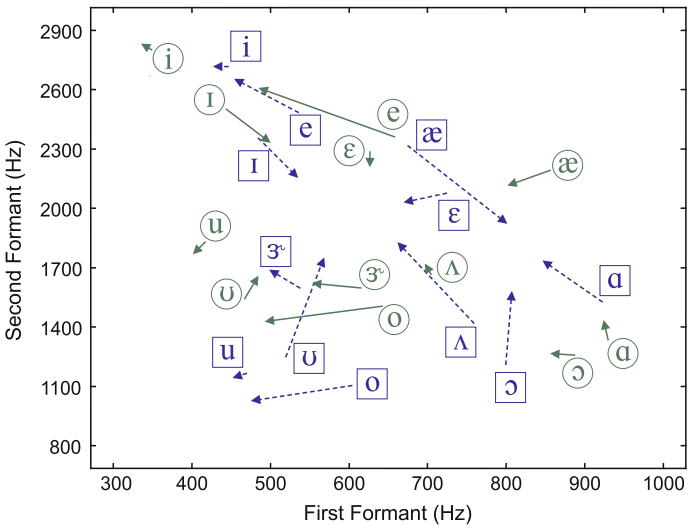


Fig. 12 Spectral change patterns for adult female speakers from two dialects: Southern Michigan (*dashed lines*) and Memphis, Tennessee (*solid lines*)

spectral change patterns in identifying vowels in both CVCs and in isolation (e.g., Nearey and Assmann 1986). Third, while coarticulation effects complicate the relationships between vowel type and VISC patterns, they do not serve to entirely obscure these relationships. Vowel category separability is improved with the incorporation of spectral change information even when there is considerable variation in phonetic environment (Zahorian and Jagharghi 1993; Hillenbrand et al. 2001). Having said all that, it needs to be noted that a very large share of the evidence on these questions comes from citation-form utterances that are much longer and, in all likelihood, show less pronounced coarticulatory effects than are seen in connected speech. Exploring these effects using the far more rapid speaking rates that are observed in conversational speech would be a useful avenue for further work on this problem.

Acknowledgments This work was supported by a grant from the National Institutes of Health (R01-DC01661). Thanks to Kelly Woods for her extensive work in recording and analyzing the Memphis vowels, and to Geoffrey Stewart Morrison, Michael Kiefte, Michael Clark, and Stephen Tasko for comments on earlier versions of this chapter.

References

Ainsworth, W.A.: Duration as a cue in the recognition of synthetic vowels. J. Acoust. Soc. Am. **51**, 648–651 (1972). doi:10.1121/1.1912889

Andruski, J.E., Nearey, T.M.: On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables. J. Acoust. Soc. Am. **91**, 390–410 (1992). doi:10.1121/1.402781

Bennett, D.C.: Spectral form and duration as cues in the recognition of English and German vowels. Lang. Speech **11**, 65–85 (1968)
Black, J.W.: Natural frequency, duration, and intensity of vowels in reading. J. Speech Hear. Disorders **14**, 216–221 (1949)
Bladon, A.: Arguments against formants in the auditory representation of speech. In: Carlson, R., Granstrom, B. (eds.) The Representation of Speech in the Peripheral Auditory System, pp. 95–102. Elsevier Biomedical Press, Amsterdam (1982)
Bladon, A., Lindblom, B.: Modeling the judgment of vowel quality differences. J. Acoust. Soc. Am. **69**, 1414–1422 (1981). doi:10.1121/1.385824
Crystal, T.H., House, A.S.: Segmental durations in connected-speech signals: Current results. J. Acoust. Soc. Am. **83**, 1553–1573 (1988). doi:10.1121/1.395911
Fairbanks, G., Grubb, P.: A psychophysical investigation of vowel formants. J. Speech Hear. Res. **4**, 203–219 (1961)
Fox, R.A., McGory, J.T.: Second language acquisition of a regional dialect of American English by native Japanese speakers. In: Bohn, O.-S., Munro, M.J. (eds.) Language Experience in Second Language Speech Learning: In Honor of James Emil Flege, pp. 117–134. John Benjamins, Amsterdam (2007)
Hillenbrand, J.M., Clark, M.J., Houde, R.A.: Some effects of duration on vowel recognition. J. Acoust. Soc. Am. **108**, 3013–3022 (2000). doi:10.1121/1.1323463
Hillenbrand, J.M., Clark, M.J., Nearey, T.M.: Effect of consonant environment on vowel formant patterns. J. Acoust. Soc. Am. **109**, 748–763 (2001). doi:10.1121/1.1337959
Hillenbrand, J.M., Gayvert, R.T.: Identification of steady-state vowels synthesized from the Peterson-Barney measurements. J. Acoust. Soc. Am. **94**, 668–674 (1993). doi:10.1121/1.406884
Hillenbrand, J.M., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**, 3099–3111 (1995). doi:10.1121/1.411872
Hillenbrand, J.M., Houde, R.A.: A narrow band pattern-matching model of vowel perception. J. Acoust. Soc. Am. **113**, 1044–1055 (2003). doi:10.1121/1.1513647
Hillenbrand, J.M., Nearey, T.M.: Identification of resynthesized /hVd/ syllables: Effects of formant contour. J. Acoust. Soc. Am. **105**, 3509–3523 (1999). doi:10.1121/1.424676
Huang, C.B.: The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp 893–896 (1986)
Jacewicz, E., Fox, R.A.: Cross-dialectal differences in dynamic formant patterns in American English vowels. In Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change, Chap. 8. Springer, Heidelberg (2013)
Jenkins, J.J., Strange, W.: Perception of dynamic information for vowels in syllable onsets and offsets. Percept Psychophysics **61**, 1200–1210 (1999)
Jenkins, J.J., Strange, W., Edman, T.R.: Identification of vowels in 'vowelless' syllables. Percept Psychophys **34**, 441–450 (1983)
Klatt, D.H.: Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. J. Acoust. Soc. Am. **59**, 1208–1221 (1976). doi:10.1121/1.380986
McAuley, R.J., Quatieri, T.F.: Speech analysis/synthesis based on sinusoidal representation. IEEE Trans. Acoust. Speech Signal Process. **34**, 744–754 (1986). doi:10.1109/TASSP.1986.1164910
Morrison, G.S.: Theories of vowel inherent spectral change: A review. In Morrison G.S., Assmann P.F. (eds.) Vowel Inherent Spectral Change, Chap. 3. Springer, Heidelberg (2013a)
Nearey, T.M.: Vowel inherent spectral change in the vowels of North American English. In Morrison G.S., Assmann P.F. (eds.) Vowel Inherent Spectral Change, Chap. 4. Springer, Heidelberg (2013)
Nearey, T.M., Assmann, P.F.: Modeling the role of vowel inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986). doi:10.1121/1.394433
Parker, E.M., Diehl, R.L.: Identifying vowels in CVC syllables: Effects of inserting silence and noise. Percept Psychophys **36**, 80–369 (1984)

Peterson, G., Barney, H.L.: Control methods used in a study of the vowels. J. Acoust. Soc. Am. **24**, 175–184 (1952). doi:10.1121/1.1906875

Peterson, G., Lehiste, I.: Duration of syllable nuclei in English. J. Acoust. Soc. Am. **32**, 693–703 (1960). doi:10.1121/1.1908183

Potter, R.K., Steinberg, J.C.: Toward the specification of speech. J. Acoust. Soc. Am. **22**, 807–820 (1950). doi:10.1121/1.1906694

Stevens, K.N.: The role of duration in vowel identification. Quarterly Progress Report 52, Research Laboratory of Electronics, MIT (1959)

Stevens, K.N., House, A.S.: Perturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Hear. Res. **6**, 111–128 (1963)

Strange, W.: Dynamic specification of coarticulated vowels spoken in sentence context. J. Acoust. Soc. Am. **85**, 2135–2153 (1989). doi:10.1121/1.389855

Strange, W., Jenkins, J.J., Johnson, T.L.: Dynamic specification of coarticulated vowels. J. Acoust. Soc. Am. **74**, 695–705 (1983). doi:10.1121/1.389855

Strange, W., Jenkins, J.J., Miranda, S.: Vowel identification in mixed-speaker silent-center syllables. J. Acoust. Soc. Am. **95**, 1030–1043 (1994). doi:10.1121/1.410014

Tiffany, W.: Vowel recognition as a function of duration, frequency modulation and phonetic context. J. Speech Hear. Disorders **18**, 289–301 (1953)

van Santen, J.P.H.: Contextual effects on vowel duration. Speech Commun. **11**, 513–546 (1992)

Zahorian, S.A., Jagharghi, A.J.: Spectral-shape features versus formants as acoustic correlates for vowels. J. Acoust. Soc. Am. **94**, 1966–1982 (1993). doi:10.1121/1.407520

Theories of Vowel Inherent Spectral Change

Geoffrey Stewart Morrison

Abstract In many dialects of North-American English, in addition to vowels which are traditionally described as true and phonetic diphthongs, several vowels traditionally described as monophthongs also have substantial formant movement. Vowel inherent spectral change (VISC) has also been found to be an important factor in the perception of vowel-phoneme identity. This chapter reviews literature pertinent to theories of the perceptually relevant aspects of VISC. Three basic hypotheses have been proposed, onset + offset, onset + slope, and onset + direction; each taking the position that initial formant values are relevant but then differing as to the relevant aspect of formant movement. Of these, the weight of evidence indicates that the onset + offset hypothesis is superior in terms of leading to higher correct-classification rates and higher correlation with listeners' vowel identification responses. Models which fit curves to whole formant trajectories have, as yet, not been found to outperform simple models based on formant measurements taken at two points (onset and offset) in formant trajectories. A popular curve-fitting model (first-order discrete cosine-transform, DCT) is interpretable as a parameterization of the onset + offset hypothesis.

Abbreviations

DCT	Discrete cosine transform
F	Formant
F1	First formant

G. S. Morrison (✉)
Forensic Voice Comparison Laboratory, School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia
e-mail: geoff-morrison@forensic-voice-comparison.net

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_3,

F2 Second formant
F3 Third formant
t Time
VISC Vowel inherent spectral change

1 Introduction

The English vowel system traditionally comprises true diphthongs, e.g., /aɪ, aʊ, ɔɪ/, so called phonetic diphthongs, /e, o/ (often transcribed as /eɪ, oʊ/), and nominal monophthongs, e.g., /i, ɪ, ɛ, æ/. However, as discussed in Hillenbrand (2013 Chap. 2), in at least some dialects of English several nominal monophthongs are in fact diphthongized. Figure 1 illustrates the extent of *vowel inherent spectral change* (VISC) from the beginning to the end of phonetic diphthongs and nominal monophthongs produced by English speakers from Alberta in western Canada. Note that the /ɪ/, /ɛ/, and /æ/ vowel phonemes, traditionally described as monophthongs, actually have substantial formant movement. Although the details may differ, nominal monophthongs having substantial VISC has been documented in a number of North-American dialects of English, for example, dialects from Alabama (Fox and McGory 2007); Alberta (Andruski and Nearey 1992; Assmann et al. 1982; Nearey and Assmann 1986), Indiana (Hargus Ferguson and Kewley-Port 2002), southern Michigan (Hillenbrand et al. 1995, 2001; Hillenbrand and Nearey 1999), North Carolina (Jacewicz and Fox 2013 Chap. 8), Ohio (Fox 1983, 1989; Fox and McGory 2007; Jacewicz and Fox 2013 Chap. 8), Texas (Assmann and Katz 2000); and Wisconsin (Jacewicz and Fox 2013 Chap. 8). This also appears to be true for Dutch (Adank et al. 2004, 2007). Unless stated otherwise, subsequent references to English in this chapter are references to North-American (Canadian or US) English, and it is the pan-dialectal importance of VISC rather than inter-dialect differences in VISC which are of concern.

Given that traditional monophthongs can have substantial VISC, the present chapter simply treats all traditional monophthongs, phonetic diphthongs, and true diphthongs in a unified manner as vowels which can have some characteristic pattern of spectral change (including the possibility of no or negligible spectral change).

As already discussed in Hillenbrand (2013 Chap. 2), VISC has been found to play an important rôle in speech perception: Listeners' identification responses change when they are presented with vowel stimuli that have natural formant trajectories versus flat formant trajectories or reversed formant trajectories (Nearey and Assmann 1986; Nearey 1995; Hillenbrand and Nearey 1999; Assmann and Katz 2000, 2005). For example, when formant trajectories are reversed, /e/ stimuli may be perceived as /ɪ/, and /ɪ/ stimuli perceived as /e/ (Nearey and Assmann 1986, see Fig. 1). Listeners also give higher goodness ratings to synthetic versions of nominal monophthongs that include VISC (Nearey 1995). In addition, when pattern recognition models are provided with information about formant trajectories in nominal

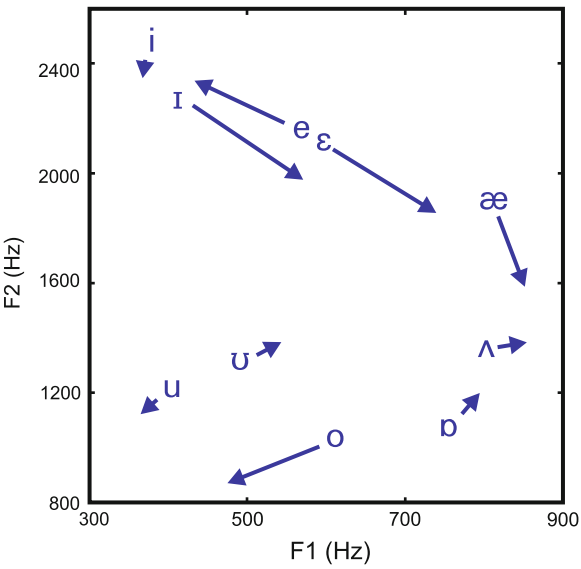


Fig. 1 Diphthongization of traditional monophthongs and phonetic diphthongs in Alberta English (vowels produced in isolation). *Arrow tails* Mean formant values measured at the beginning of the vowel. *Arrow heads* Mean formant values measured at the end of the vowel. Adapted from Nearey and Assmann (1986)

monophthongs, as compared to formant measurements from a single point, higher correct-classification rates are obtained, and there is higher correlation with listeners' perception patterns (see additional discussion in Sect. 3.1).

Three basic hypotheses have been advanced as to the aspects of VISC which are perceptually relevant for vowel identification. This chapter reviews previously published work with the primary aim of determining which of the three hypotheses best fits the empirical data. It also considers whether more complex models of VISC may outperform models based on the basic hypotheses.

2 The Three Basic Hypotheses of VISC Perception

Three basic hypotheses have been advanced as to the aspects of VISC which are relevant for the perception of vowel identity (Gottfried et al. 1993; Nearey and Assmann 1986; Pols 1977). All three hypotheses agree that the initial formant frequencies are perceptually relevant to vowel identification (for supporting evidence see Gay 1970; Bladon 1985, Nábělek et al. 1993; Nearey 1995), but disagree on what additional cues are relevant in VISC perception[1]:

[1] The terminology used here is based on Gottfried et al. (1993) "onset + offset", "onset + slope", and "onset + direction", which will frequently be abbreviated to *offset*, *slope*, and *direction*. Nearey and Assmann's (1986) "dual-target", "target-plus-slope", and "target-plus-direction" represent the same hypotheses. Gottfried et al. only tested F2 slope for their slope hypothesis, but both F1 and F2 slopes were tested in studies conducted by Nearey and colleagues. In contrast to Lehiste and Peterson's (1961) use of the term *target*, Nearey and Assmann's term

- The onset + *offset* hypothesis states that the relevant perceptual cues are the formant values at the end of the vowel. This may be expressed as the relative change in formant values from onset to offset, i.e., [ΔF1, ΔF2] or **ΔF**.
- The onset + *slope* hypothesis states that the relevant perceptual cues are the "velocities" of formant change, i.e., whether the change in the frequency of each formant is positive or negative and the rate of change over time. This may be expressed as **ΔF**/Δt.
- The onset + *direction* hypothesis states that the only relevant factor is the direction of formant movement in an F1–F2 (or similar) space. This may be expressed as ∠**ΔF** or **ΔF**/‖**ΔF**‖.

Under the offset hypothesis, the final formant values achieved (and by implication the direction) are relevant, but the rate of formant change is irrelevant. Under the slope hypothesis, the direction and rate of formant change are relevant, but the final formant values achieved are irrelevant. Under the direction hypothesis, the direction of formant change is relevant, but the rate of formant change and the final formant values achieved are irrelevant.

The differences between the hypotheses can be illustrated using the stylized formant vectors which appear in Fig. 2:

- Vectors A, B, C, and D have the same direction in the F1–F2 plane, and under the direction hypothesis they should therefore all be perceived as the same vowel category.
- Vectors A, B, and D have the same direction and rate of change of formant values and under the slope hypothesis should therefore all be perceived as the same vowel category. The rate of change of vector C is half that of vectors A, B, and D, and under the slope hypothesis it may therefore potentially be perceived as a different vowel category.
- Vectors A, C, and D have the same end-point in the F1–F2 plane, and under the offset hypothesis they should therefore all be perceived as the same vowel category. Vector B is twice as long in the F1–F2 plane as vectors A, C, and D, and under the offset hypothesis it may therefore potentially be perceived as a different vowel category.

A requirement for all three hypotheses is that formant change be at least detectable by the listener. Tokens of a vowel category with negligible VISC may have random fluctuations in formant movement which are not perceived (Nearey and Assmann 1986). Perception of a vowel as a diphthong, as opposed to a monophthong, may also require the time during which formant movement is detectable to exceed some minimum duration (see Nábělek et al. 1994). Kewley-Port and Goodman (2005) reported that, under near optimal conditions, the mean perceptual threshold for the magnitude of second-formant movement in front vowels ranged from 16 to 51 Hz for

(Footnote 1 continued)
dual-*target* does not imply that there must be steady states at the beginning and end of the diphthongs.

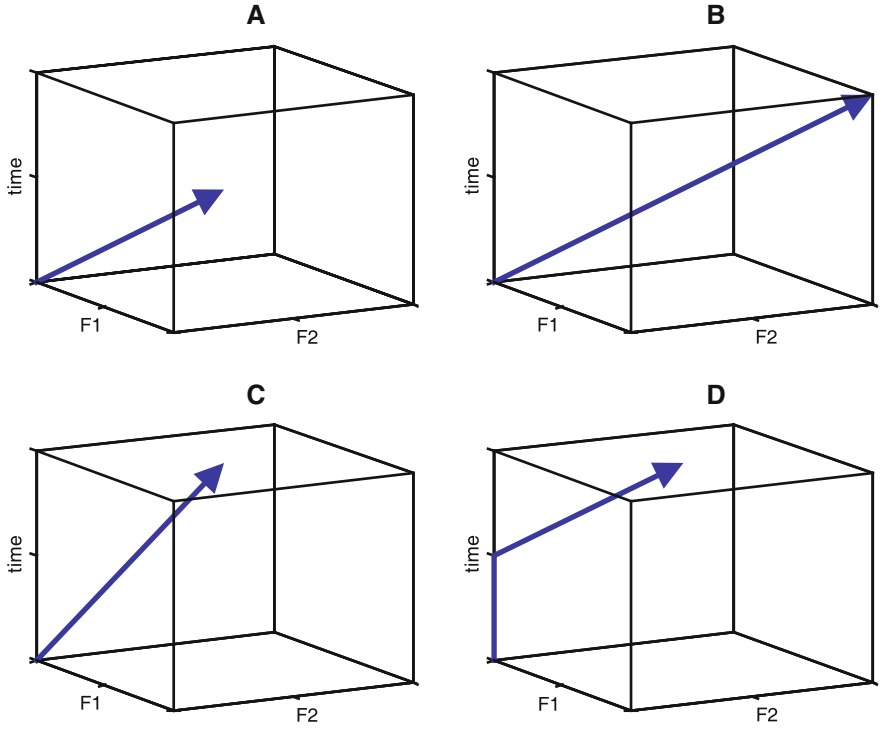


Fig. 2 Stylized formant vectors [ΔF1, ΔF2, Δt] used to illustrate the differences between the three basic hypotheses, see text. Vector A: [1, 1, 1]. Vector B: [2, 2, 2]. Vector C: [1, 1, 2]. Vector D: piecewise [0, 0, 1] + [1, 1, 1]

rising F2, and 24 to 66 Hz for falling F2 (initial F2 values were 2,068, 2,272, and 2,525 Hz, and vowel durations were 110 and 165 ms). These values were at least a factor of four smaller than the magnitude of F2 movement produced in natural English /i, ɪ, e, ɛ, æ/ vowels, leading Kewley-Port and Goodman to conclude that the formant movement in these vowels would be detectable by listeners.

3 Tests of the Basic Hypotheses of VISC Perception

3.1 Studies Comparing One of the VISC Hypotheses with Static Spectral Properties

Several pattern recognition studies have found that models which include parameterization of VISC outperform models which are based on static spectral properties. These results have often been interpreted as lending support to one or

other of the VISC hypotheses without necessarily considering the alternative hypotheses.

Some studies have found evidence which could be interpreted as supporting the slope hypothesis. Gay (1970) claimed that slope was the primary cue for distinguishing between different diphthongs, e.g., /ɔɪ/–/aɪ/; however, his synthetic stimuli confounded either offset and slope or duration and slope, and his set of experiments did not allow full separation of the effects of slope from its covariates.[2] Assmann et al. (1982) applied pattern recognition models to measurements of formant values of English nominal monophthongs and phonetic diphthongs. They found that when they included formant slope values in addition to midpoint formant values they obtained higher correct-classification and, more importantly, higher correlation with human listeners' response patterns.

Other studies have found evidence which could be interpreted as supporting the offset hypothesis. Hillenbrand et al. (1995), Andruski and Nearey (1992), Hillenbrand and Nearey (1999), and Hillenbrand et al. (2001) obtained higher correct-classification rates or higher correlation with listeners' response patterns when they used two-point (onset + offset) versus one-point (mean or midpoint) parameterizations of English nominal monophthongs and phonetic diphthongs (see also Adank et al. 2004 for Dutch vowels). Andruski and Nearey (1992) conducted experiments using silent-center natural /bVb/ stimuli (using only short portions extracted from the beginning and end of natural productions), silent-center natural isolated vowel stimuli, and synthetic /bVb/ stimuli in which the vowel formants were linear interpolations from initial to final formant values.[3] Since similar perceptual results were obtained for all three stimulus types, they argued that the perceptually relevant cues were those shared by all three, i.e., the onset and offset values (this is also a possible interpretation of the results of Strange et al. 1983).[4] Using a different methodology with natural English true

[2] The interpretation of Gay's (1970) results is hindered by contradictions between the description of his stimuli and the discussion of the results. Discussion and graphical results suggest that, in his Experiment II, F2 offset did not covary with duration so as to maintain a fixed slope, rather, F2 offset stepped up at a slower rate than duration. For example, for /ɔ/–/ɔɪ/ stimuli with an F2 onset of 840 Hz, the first three duration steps of 100, 110, and 120 ms all had an F2 offset of 1,320 Hz, and thus progressively shallower slopes of 4.80, 4.36, and 4.00 Hz/ms; the next two duration steps of 130 and 140 ms both had an F2 offset of 1,440 Hz, and thus slopes of 4.62 and 4.29 Hz/ms; etc.

[3] Although the classical description of a diphthong includes an initial steady state, a glide, and a final steady state (Lehiste and Peterson 1961), there is usually no second steady state (see Holbrook and Fairbanks 1962), the first steady state may disappear at fast speaking rates (Gay 1968), and intelligible diphthongs can be synthesized using only a glide (Gay 1970).

[4] There has been an ongoing disagreement between Strange and colleagues and Nearey and colleagues over the interpretation of silent-center results. Strange and Jenkins (2013 Chap. 5) review one of their earlier studies (Jenkins and Strange 1999) in which a silent-center condition (first 3 pitch periods and last 4 pitch periods of the vowel) resulted in listeners having high correct-classification rates for vowel-phoneme identity even when vowel-duration information was neutralized, but playing the last 8 pitch periods resulted in low correct-classification rates. They claim that these results argue "against the hypothesis that nucleus + offglide direction information provides the critical dynamic spectral information for AE vowels in continuous

diphthongs, phonetic diphthongs, and nominal monophthongs in /hVd/ context, Fox (1983) also obtained results consistent with the offset hypothesis. In a multidimensional scaling experiment, Fox extracted four perceptual dimensions: the first dimension was most highly correlated with F2 formant values measured at the end of the vowels, and the third dimension with F2 formant values measured at the beginning of the vowels (the second dimension was not dominated by any measurement relevant to VISC).

Other studies have found evidence which could be interpreted as supporting an onset + midpoint + offset hypothesis; however, this has not been proven to be superior to the onset + offset hypothesis. Huang (1992) for American English /i/, /ɪ/, /e/, /ɛ/, and /ʌ/, and Harrington and Cassidy (1994) for Australian English diphthongs and nominal monophthongs, found that pattern classifiers based on three-point models (e.g., measurements taken at 25, 50, and 75 % of vowel duration) outperformed one-point models (e.g., measurements taken at 50 % of vowel duration). The authors did not claim that the three-point model was the absolute correct parameterization, only that more than a one-point model was necessary. As already discussed in Hillenbrand (2013 Chap. 2), Hillenbrand et al. (1995) compared one-point, two-point, and three-point parameterizations of English nominal monophthongs. Substantially higher correct-classification rates were obtained for two-point models compared to one-point models, but three-point models offered little or no improvement over two-point models.

Neel (2004) investigated the perception of synthetic 1, 2, 3, 5, and 11 point versions of English phonetic diphthongs and nominal monophthongs. Each stimulus was based on the formant tracks from a single /dVd/ production from one of two speakers. Two-point stimuli based on formant measurements at 10 and 90 % of duration were poorly identified, typically at rates substantially worse than one-point stimuli based on formant measurements at 50 % of duration. Averaged over all vowels and both speakers the correct-classification rate was 64 % for one-point stimuli and 43 % for two-point stimuli (although correct-classification rates

(Footnote 4 continued)
speech contexts". On the contrary, the silent-center results are exactly what the onset + offset hypothesis would predict, and from their citations it appears that they are referring to what I am calling the onset + offset hypothesis. They claim that the last 8 pitch periods included "target plus offglide", but unless the whole vowel was about 8 pitch periods long this cannot be the case (again interpreting their "target plus offglide" as equivalent to what I am calling onset + offset). In fact, the two shortest vowel-tokens they tested, tokens of /ɪ/ and /ʊ/ with durations of 12–14 pitch periods, had better correct-classification rates in the last-8-pitch periods condition than in the silent-center condition. The situation was reversed once the vowel tokens were about twice as long as 8 pitch periods (for all vowel phonemes with tokens averaging more than 15 pitch periods long). Correct-classification rates were very poor for the longest vowels, tokens of /æ/ and /a/ were 19–21 pitch periods long, and worst for tokens of /e/ and /o/ which were 18–21 pitch periods long and had greater formant movement than /æ/ and /a/. What Jenkins and Strange appear to have tested for the longer vowels is an offset-only hypothesis—I am not aware of anyone ever having seriously advocated such a hypothesis.

from some stimuli actually improved). A possible reason for the general poor-performance on the two-point versus the one-point stimuli is that the 10 and 90 % points may actually have been in the consonant transitions and were therefore not representative of the vowel categories' characteristic onset and offset values. Identification rates were generally high for three-point stimuli based on formant measurements at 10, 50, and 90 % of duration (71 % correct averaged over all vowels and both speakers), five-point stimuli based on formant measurements at 10, 30, 50, 70 and 90 % of duration (81 % correct averaged over all vowels and both speakers), and eleven-point stimuli based on formant measurements at 10, 20, 30, 40, 50, 60, 70, 80 and 90 % of duration (85 % correct averaged over all vowels and both speakers). The correct-classification rate for the original vowels was 89 % averaged over all vowels and both speakers. The increasing improvement beyond three points, suggests that the entire trajectory of the formants (possibly including consonant transitions, see Nearey 2013 Chap. 4) may include additional vowel identity information over and above that captured in a simple two- or three-point model. This is further explored in Sect. 4.

3.2 Studies Comparing Two of the VISC Hypotheses

Much of the work on theories of VISC has presented arguments against one of the hypotheses in an attempt to falsify it and leave one of the alternatives as the survivor. In these studies only two of the three hypotheses were considered.

Contra the offset hypothesis and in support of the slope hypothesis, Gay (1968) found substantial speaking-rate dependent differences in final formant values and more consistency in slope (see also Borzone de Marique 1979 for slope consistency in Spanish, and Pols 1977 for direction consistency in Dutch); however, it could be argued that listeners are able to compensate for target undershoot, and that substantial variability in target may be unproblematic if there are only a few widely separated targets, and thus little chance of confusion between them (Bladon 1985).

Contra the slope hypothesis and in support of the offset hypothesis, Bond (1978, 1982) found that changing the duration of the glide between onset and offset had little effect on vowel identification, and in some cases even deleting the glide completely (without leaving a gap) had no effect (for glide deletion see also Wise 1964; Bladon 1985; Nearey and Assmann 1986; Andruski and Nearey 1992). In Kewley-Port and Goodman's (2005) study on the perceptual threshold for F2 movement, stimuli included long and short synthetic vowels (165 v 110 ms), where a long vowel had the same slope as one of the shorter vowels but the same offset value as a different shorter vowel. They found no significant effect for duration, i.e., no difference in ΔF threshold values for long and short stimuli with the same end points but different slopes, leading them to conclude that their results supported the offset hypothesis over the slope hypothesis.

Contra the direction hypothesis and in support of the offset hypothesis, Bladon (1985) found that phonetically trained listeners transcribed truncated diphthongs with pairs of symbols appropriate for monophthongs at the initial and final formant values of the stimuli. The second symbol varied with the final formant values and changed even though the direction of formant movement did not. However, the generalizability of Bladon's (1985) results is not clear: He removed the latter portions of /ia/, /iɛ/, and /ie/, all three have similar initial formant values and a similar direction, but different final targets. It is not clear that /ia/, /iɛ/, and /ie/ are perceived holistically as single vowels rather than as sequences of two vowels, and the results may therefore not be generalizable to the perception of true diphthongs, phonetic diphthongs, and nominal monophthongs (Divenyi 2009 investigated the perception of formant transitions between a sequence of two vowels). Jacewicz et al. (2003) found that listeners' responses shifted from /a/ to /aɪ/ as F2 offset was increased, and from /aɪ/ to /ɛɪ/ as F2 onset was increased. Although they argued that the initial and final formant values did not characterize the diphthong, their interpretation of the results amounts to a thresholded version of the offset hypothesis: Once formant movement in a synthetic vowel has achieved roughly half the formant movement found in natural /aɪ/ productions, listeners fairly reliably identify the synthetic vowel as the diphthong /aɪ/. The threshold was much greater than the just noticeable difference for F2 movement (Kewley-Port and Goodman 2005), hence my interpretation of this as a thresholded version of the offset hypothesis rather than a thresholded version of the direction hypothesis.

3.3 Studies Testing All Three VISC Hypotheses

Nearey and Assmann (1986) tested the three basic VISC hypotheses using pattern recognition models trained on different parameterizations of English nominal monophthongs and phonetic diphthongs. Parameters were initial F1 and F2 values plus:

- Offset: final F1 and F2 values ($\mathbf{\Delta F}$).
- Slope: change in F1 and F2 values over the duration of the glide ($\mathbf{\Delta F}/\Delta t$)
- Direction: change in F1 and F2 values each over the magnitude of the total change ($\mathbf{\Delta F}/||\mathbf{\Delta F}||$)

All formant values were transformed to log-hertz prior to making any other calculations. Correlations with listeners' responses were slightly higher for the offset and the direction models than for the slope model, but in general all three parameterizations provided adequate characterizations of listeners' response patterns.

Gottfried et al. (1993) compared the three hypotheses using pattern recognition models trained on different parameterizations of English phonetic and true diphthongs. They used two sets of parameterizations: one was similar to that of Nearey

and Assmann (1986) in that it used log-F1 and log-F2 measurements, but differed in that only the F2 slope was included,[5] and that the direction was specified as an angle in degrees (∠**ΔF**), with adjustments made to avoid discontinuities at 0°/360°. The second set of parameters transformed F1, F2, and F3 values into Miller's *auditory-perceptual space* (APS: Miller 1989). Across speaking conditions (slow stressed, slow unstressed, fast stressed, and fast unstressed) the log-formant parameterizations had slightly higher correct-classification rates for the offset and the slope models than for the direction model, and in the APS parameterizations the offset model had higher correct-classification rates than the slope and the direction models. However, no one hypothesis-based model clearly outperformed the others in all contexts.

Morrison and Nearey (2007) tested the three basic hypotheses using a synthetic English /e/–/ɪ/–/ɛ/ continuum which had a fixed onset but systematic variation in duration, offset values, and slope (see Figs. 3 and 4). For each set of final formant values and vowel duration, two slope values were created by synthesizing one stimulus with a straight line trajectory in the log-hertz F1–F2 space, and synthesizing another with an elbow (piecewise linear, see Figs. 2d and 4). The elbowed stimuli had no formant movement during the first quarter of the vowel and straight transition from onset to offset values during the last three quarters. Pairs of elbowed and straight stimuli therefore had the same offset values, but the elbowed member of each pair had a slope which was a third steeper than the straight member. Direction was fixed to a single diagonal in the F1–F2 space, but stimuli had several magnitudes of positive and negative movement along this diagonal, thus multiple stimuli had the same direction but different offset values (see Fig. 3). Formant movement along the diagonal was either in the rising-F1–falling-F2 (/ɪ/-like and /ɛ/–like) direction, the opposite falling-F1–rising-F2 (/e/-like) direction, or had zero formant movement (/i/-like direction).

Listeners identified the stimuli, and general additive models were fitted to the perception results. Adding offset parameters to models already containing slope or direction parameters resulted in a significant improvement of fit to listeners' responses, but adding slope or direction parameters to a model already containing offset parameters did not. Compared to the direction and slope hypotheses, the offset hypothesis therefore gave a better account of human listeners' VISC perception.

[5] Assmann and Katz (2000) tested the perception of stimuli in which the F1 trajectory was flattened and F2 unchanged, and stimuli in which the F2 trajectory was flattened and F1 unchanged. Listeners' correct identification rates for English nominal monophthongs and phonetic diphthongs significantly decreased when either formant was flattened. Although some vowels were affected more by F1 flattening, some were affected more by F2 flattening. The results indicate that a VISC theory applicable across vowel categories should refer to formant movement in both F1 and F2.

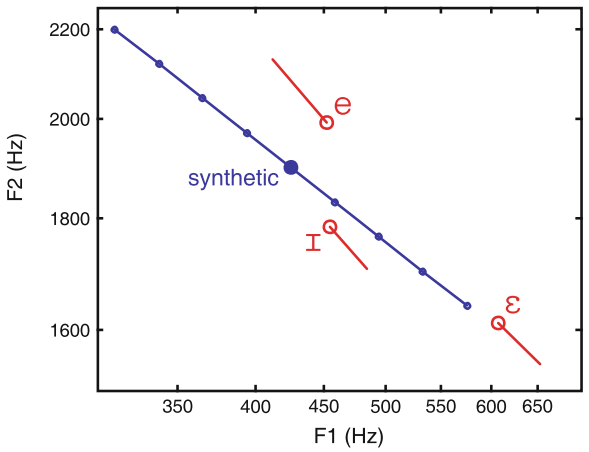


Fig. 3 First and second formant properties of natural and synthetic vowels from Morrison and Nearey (2007). The comets labeled /e/, /ɪ/, and /ɛ/ represent the mean onsets and offsets of these vowels produced in isolated-word /bVpə/ context by seven male speakers of Alberta English (ten replications per speaker). Comet heads and tails represent the formant values at 25 and 75 % of the duration of the vowels respectively. For the synthetic stimuli, the large dot represents the initial formant values and the small dots represent the nine sets of final formant values. From Morrison and Nearey (2007)

4 Curve-Fitting Parameterizations

4.1 Problems with Multi-Point Parameterizations

Parameterizations of VISC based on formant measurements at two or three points in the vowel have been criticized as being crude measures incapable of capturing all the relevant details of inherently complex time-varying patterns (Clermont 1993;

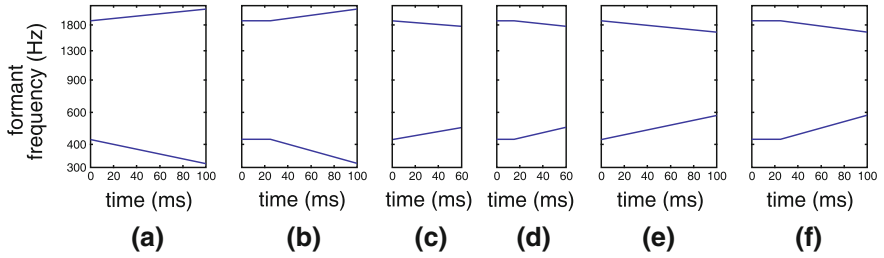


Fig. 4 Examples of formant trajectories of synthetic stimuli in Morrison and Nearey (2007). The examples are straight (**a**, **c**, **e**) and elbowed (**b**, **d**, **f**) versions of the stimuli which, in the perception experiment in Morrison and Nearey (2007), received the greatest number of /e/ (**a**, **b**), /ɪ/ (**c**, **d**), and /ɛ/ (**e**, **f**) responses. The elbowed stimuli had steeper slopes than their corresponding straight stimuli. From Morrison and Nearey (2007)

Jenkins et al. 1994, see also the discussion of Neel 2004 in Sect. 3.1). A concrete problem with a two-point model is the choice of time points at which to measure formant values. Different studies have selected different points at which to measure the vowel onset and offset values, e.g., at the earliest and latest measurable values with amplitudes not less than 15 dB below the vowel's maximum amplitude (Nearey and Assmann 1986), 40 ms after the initial consonant release and 40 ms before the final consonant closure (Andruski and Nearey 1992), at 20 and 80 % of the duration of the vowel (Hillenbrand and Nearey 1999), and at 20 and 70 % of the duration of the vowel (Hillenbrand et al. 2001). Gottfried et al. (1993) measured at points immediately following and preceding the consonant transitions, which they determined on the basis of an algorithm which made use of the rate of change of formant movement over time. The choice of time points will clearly have an influence on an onset + offset parameterization (see Assmann et al. 2013 Chaps. 1 and 9). It may also affect a slope parameterization: if there is any steady state portion between the measurement points then the true slope will be underestimated (most studies using data based on acoustic measurements of productions have not attempted to divide vowels into steady state and glide portions). The direction parameterization is least likely to be affected. These philosophical and practical problems may be overcome by using curve-fitting parameterizations, examples of which are discussed in the remainder of Sect. 4.

4.2 Discrete Cosine Transforms

The most popular parametric-curve in VISC research is the *discrete cosine transform* (DCT). The zeroth DCT coefficient gives the mean value of the signal (e.g., the mean of the time-varying formant values measured from the beginning to the end of the vowel), and the first coefficient gives the sign and magnitude of a half period of a cosine (a backward "S" shape) fitted to the signal after subtraction of the mean. Continuing the series, the second DCT coefficient is associated with a whole period of a cosine (a "U" shape), the third with one-and-a-half periods (a backward "N" shape) etc. (see Fig. 5). There are several variants of the DCT formula, and the zeroth coefficient could be a mean value or an initial intercept value. Also, depending on the variant, the coefficient values may or may not include a scaling factor related to the length of the signal. Unless otherwise specified, further discussion in the present chapter will assume that the zeroth coefficient is the mean and that coefficient values are not scaled relative to the length of the signal. Sometimes, as was the case in Zahorian and Jagharghi (1991, 1993) and Watson and Harrington (1999), what is referred to here as the zeroth coefficient is referred to as the first, the first coefficient is referred to as the second, etc. Figure 6 shows an example of a first-order DCT fitted to an F2 trajectory of an /aɪ/ token, the zeroth coefficient value is 1,490 Hz (the mean of F2 over time) and the first coefficient value is −446 Hz (the peak value of the deviation of a half-cosine from the mean value, negative because F2 is rising over time).

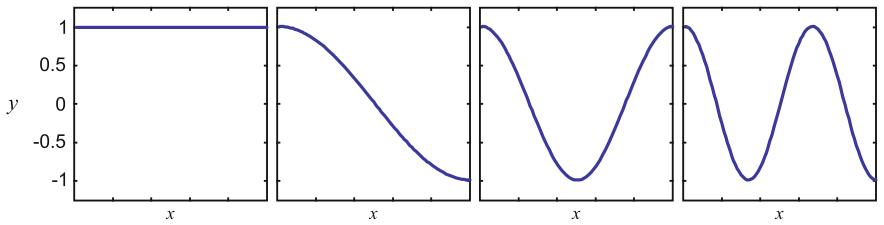


Fig. 5 *Left* to *right* Zeroth through third DCT basis functions

When a vowel-duration scaling factor is not included or vowel durations are equalized prior to DCT calculation, the value of the zeroth DCT coefficient is the mean value of that formant, and the value of the first DCT coefficient is a symmetrically constrained measure of the direction and distance of the onset and offset of a formant relative to the mean value of that formant. This parameterization is therefore an onset + offset parameterization, but based on a curve fitted to the whole trajectory rather than only two points. If a vowel-duration scaling factor is included and there is no time normalization, then the first-order DCT becomes an onset + slope parameterization. DCTs above first-order allow for more complex shapes which do not directly map to any of the three basic hypotheses.

4.3 Experimental Results

Zahorian and Jagharghi (1991, 1993) fitted second-order DCTs to the time-varying spectral properties of English duration-equalized nominal monophthongs (although /e/ was excluded, /o/ was included).[6] The spectrum at each time frame was parameterized as formant values and as cepstral coefficients, and then DCTs were fitted to the formant values or cepstral coefficient values. For both spectral parameterizations, the higher correct-classification rates and higher correlations with listeners' responses were obtained for these models, which included dynamic spectral information, than for static spectral models (cepstral models also outperformed formant models).

As mentioned in Sect. 4.2, a second-order DCT is more complex than any of the three basic hypotheses. Zahorian and Jagharghi (1991, 1993) implied that in pilot tests second-order DCT models outperformed first-order DCT models, hence this more complex model outperformed a parameterization of the onset + offset model (although, since the details of the pilot tests were not reported, it is not clear by how much the former may have outperformed the latter).

[6] Zahorian and Jagharghi (1991, 1993) reserved the term *discrete cosine transform* (DCT) for a curve fitted to a spectrum at a single time frame (what is normally referred to as a *cepstrum*), and used the term *discrete cosine series* (DCS) for curves fitted to a time-ordered series of formant or cepstral coefficient values.

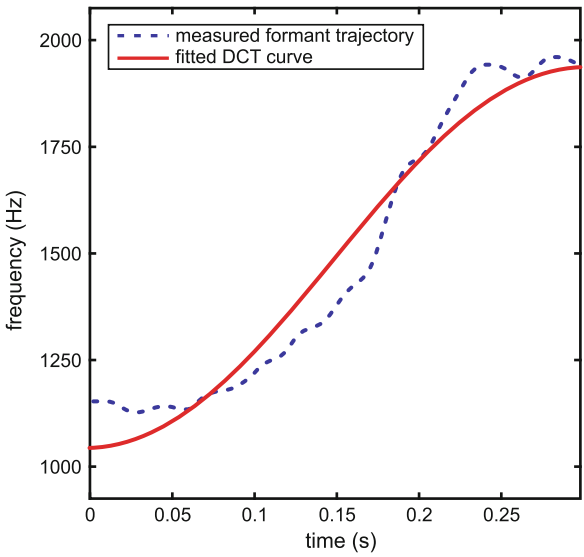


Fig. 6 First-order DCT *curve* fitted to the F2 trajectory of an /aɪ/ token

Zahorian and Jagharghi (1993) reported also testing Legendre polynomial basis functions, and least-squares polynomial fitting, but obtained slightly higher correct-classification rates using DCTs. They speculated that the relative success of DCTs may be related to greater stability due to their edge constraints.

Watson and Harrington (1999) fitted DCTs to time-normalized formant trajectories from Australian English vowels. They obtained higher correct-classification rates for models using the zeroth and first DCT coefficients compared to models using only the zeroth, but no significant additional improvement when the second DCT coefficient was added, i.e., the DCT model representing a parameterization of the onset + offset hypothesis outperformed the static model, but, in contrast to Zahorian and Jagharghi (1993), a more complex model did not outperform the onset + offset model. It may be that the difference in the results between these two studies is related to the difference in the English dialect studied.

Hillenbrand et al. (2001) reported fitting polynomials and DCTs to formant trajectories from English nominal monophthongs and phonetic diphthongs, and comparing the results to two-point models. They concluded that in terms of correct-classification rates the curve-fitting parameterizations were not superior to the simpler two-point parameterization. Thus, as yet, there is no evidence to suggest that curve-fitting models are superior to the basic two-point onset + offset model with respect to the substantive issues of correct-classification and correlation with listeners' responses.

5 Conclusion

The weight of evidence in the literature reviewed here indicates that the onset + offset hypothesis provides a better account of the perceptually relevant aspects of VISC with respect to vowel identity than either the onset + slope hypothesis or the onset + direction hypotheses. In terms of correct-classification rates and correlation with listeners' responses, more sophisticated curve-fitting parameterizations of VISC have not, as yet, been found to outperform the simple two-point parameterizations of the onset + offset hypothesis. The first-order DCT model is interpretable as a parameterization of the onset + offset hypothesis, and should probably be preferred to the two-point parameterization given that it exploits all the available formant-track data without having to make arbitrary decisions as to which points to measure, and is therefore likely to have a smaller variance.

Although the weight of evidence favors the onset + offset hypothesis, one should not conclude that onset + offset models of VISC will necessarily capture all perceptually relevant information: Zahorian and Jagharghi (1993) indicate that a more complex parametric-curve model may outperform the first-order DCT parameterization of the onset + offset hypothesis. Perception of vowel-plus-consonant diphones may required more detailed descriptions of VISC (Jacewicz et al. 2003; Moreton 2004; Nearey 2013 Chap. 4), and more complex VISC models are more effective for capturing indexical information such as speaker identity (see Morrison 2013 Chap. 3).

Acknowledgments The writing of this chapter began when the author was a PhD student at the Department of Linguistics, University of Alberta, and was supported by a Social Sciences and Humanities Research Council of Canada Doctoral Fellowship. Thanks to Terrance M. Nearey, Peter F. Assmann, James M. Hillenbrand, and Christian E. Stilp for comments on earlier versions of this chapter.

References

Adank, P., van Hout, R., Smits, R.: An acoustic description of the vowels of Northern and Southern standard Dutch. J. Acoust. Soc. Am. **116**, 1729–1738 (2004). doi:10.1121/1.1779271

Adank, P., van Hout, R., van de Velde, H.: An acoustic description of the vowels of Northern and Southern standard Dutch II: regional varieties. J. Acoust. Soc. Am. **121**, 1130–1141 (2007). doi:10.1121/1.2409492

Andruski, J.E., Nearey, T.M.: On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables. J. Acoust. Soc. Am. **91**, 390–410 (1992). doi:10.1121/1.402781

Assmann, P.F., Katz, W.F.: Time-varying spectral change in the vowels of children and adults. J. Acoust. Soc. Am. **108**, 1856–1866 (2000). doi:10.1121/1.1289363

Assmann, P.F., Katz, W.F.: Synthesis fidelity and time-varying spectral change in vowels. J. Acoust. Soc. Am. **117**, 886–895 (2005). doi:10.1121/1.1852549

Assmann, P.F., Nearey, T.M., Bharadwaj, S.V.: Developmental patterns in children's speech: patterns of spectral change in vowels. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 9). Springer, Heidelberg (2013)

Assmann, P.F., Nearey, T.M., Hogan, J.T.: Vowel identification: orthographic, perceptual, and acoustic aspects. J. Acoust. Soc. Am. **71**, 975–989 (1982). doi:10.1121/1.387579

Bladon, A.: Diphthongs: a case study of dynamic articulatory processing. Speech Commun. **4**, 145–154 (1985). doi:10.1016/0167-6393(84)90040-2

Bond, Z.S.: The effects of varying glide duration on diphthong identification. Lang. Speech **21**, 253–278 (1978)

Bond, Z.S.: Experiments with synthetic diphthongs. J. Phonetics **10**, 259–264 (1982)

Borzone de Manrique, A.M.: Acoustic analysis of Spanish diphthongs. Phonetica **36**, 194–206 (1979)

Clermont, F.: Spectro-temporal description of diphthongs in F1–F2–F3 space. Speech Commun. **13**, 377–390 (1993). doi:10.1016/0167-6393(93)90036-K

Divenyi, P.: Perception of complete and incomplete formant transitions in vowels. J. Acoust. Soc. Am. **126**, 1427–1439 (2009). doi:10.1121/1.3167482

Hillenbrand, J.M.: Static and dynamic approaches to understanding vowel perception. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel inherent spectral change (Chap. 2). Heidelberg, Springer (2013)

Fox, R.: Perceptual structure of monophthongs and diphthongs in English. Lang. Speech **26**, 21–49 (1983)

Fox, R.: Dynamic information in identification and discrimination of vowels. Phonetica **46**, 97–116 (1989)

Fox, R.A., McGory, J.T.: Second language acquisition of a regional dialect of American English by native Japanese speakers. In: Bohn, O.-S., Munro, M.J. (eds.) Language experience in second language speech learning: in honor of James Emil Flege, pp. 117–134. John Benjamins, Amsterdam (2007)

Gay, T.: Effects of speaking rate on diphthong formant movements. J. Acoust. Soc. Am. **44**, 1570–1573 (1968). doi:10.1121/1.1911298

Gay, T.: A perceptual study of American English diphthongs. Lang. Speech **13**, 65–88 (1970)

Gottfried, M., Miller, J.D., Meyer, D.J.: Three approaches to the classification of American English diphthongs. Journal of Phonetics **21**, 205–229 (1993)

Hargus Ferguson, S., Kewley-Port, D.: Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. **112**, 259–271 (2002). doi:10.1121/1.1482078

Harrington, J., Cassidy, S.: Dynamic and target theories of vowel classification: evidence from monophthongs and diphthongs in Australian English. Lang. Speech **37**, 357–373 (1994). doi:10.1177/002383099403700402

Hillenbrand, J.M., Nearey, T.M.: Identification of resynthesized /hVd/ syllables: Effects of formant contour. J. Acoust. Soc. Am. **105**, 3509–3523 (1999). doi:10.1121/1.424676

Hillenbrand, J.M., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**, 3099–3111 (1995). doi:10.1121/1.411872

Hillenbrand, J.M., Clark, M.J., Nearey, T.M.: Effect of consonant environment on vowel formant patterns. J. Acoust. Soc. Am. **109**, 748–763 (2001). doi:10.1121/1.1337959

Holbrook, A., Fairbanks, G.: Diphthong formants and their movements. J. Speech Hear. Res. **5**, 38–58 (1962)

Huang, C.B.: Modelling human vowel identification using aspects of format trajectory and context. In: Tohkura, Y., Vatikiotis-Bateson, E., Sagisaka, Y. (eds.) Speech Perception, Production and Linguistic Structure, pp. 43–61. IOS, Tokyo, Ohmsha/Amsterdam (1992)

Jacewicz, E., Fox, R.A.: Cross-dialectal differences in dynamic formant patterns in American English vowels. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 8). Springer, Heidelberg (2013)

Jacewicz, E., Fujimura, O., Fox, R.A.: Dynamics in diphthong perception. In: Sole, J., Recasens, D., Romero, J. (eds.) Proceedings of the 15th international congress of phonetic sciences, Barcelona. pp. 993–996. Causal Productions, Australia (2003)

Jenkins, J.J., Strange, W.: Perception of dynamic information for vowels in syllable onsets and offsets. Percept Psychophysics **61**, 1200–1210 (1999). doi:10.3758/BF03207623

Jenkins, J.J., Strange, W., Miranda, S.: Vowel identification in mixed-speaker silent-center syllables. J. Acoust. Soc. Am. **95**, 1030–1043 (1994). doi:10.1121/1.410014

Kewley-Port, D., Goodman, S.G.: Thresholds for second formant transitions in front vowels. J. Acoust. Soc. Am. **118**, 3252–3560 (2005). doi:10.1121/1.2074667

Lehiste, I., Peterson, G.E.: Transitions, glides, and diphthongs. J. Acoust. Soc. Am. **33**, 268–277 (1961). doi:10.1121/1.1908681

Miller, J.D.: Auditory-perceptual interpretation of the vowel. J. Acoust. Soc. Am. **85**, 2114–2134 (1989). doi:10.1121/1.397862

Moreton, E.: Realization of the English postvocalic [voice] contrast in F1 and F2. J. Phonetics **32**, 1–33 (2004). doi:10.1016/S0095-4470(03)00004-4

Morrison, G.S., Nearey, T.M.: Testing theories of vowel inherent spectral change. J. Acoust. Soc. Am. **122**, EL15–EL22 (2007) doi:10.1121/1.2739111

Morrison, G.S.: Vowel inherent spectral change in forensic voice comparison. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 11). Springer, Heidelberg (2013)

Nábělek, A.K., Czyzewski, Z., Crowley, H.: Vowel boundaries for steady-state and linear formant trajectories. J. Acoust. Soc. Am. **94**, 675–687 (1993). doi:10.1121/1.406885

Nábělek, A.K., Czyzewski, Z., Crowley, H.: Cues for perception of the diphthong /aɪ/ in either noise or reverberation. Part I. duration of the transition. J. Acoust. Soc. Am. **95**, 2681–2693 (1994). doi:10.1121/1.409837

Nearey, T.M., Assmann, P.F.: Modeling the role of vowel inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986). doi:10.1121/1.394433

Nearey, T.M.: Evidence for the perceptual relevance of vowel-inherent spectral change for front vowels in Canadian English. In: Elenius, K., Branderud, P. (eds.) Proceedings of the 13th congress of phonetic sciences, Stockholm, (pp. 678–681). KTH, Sweden (1995)

Nearey, T.M.: Vowel inherent spectral change in the vowels of North American English. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 4). Springer, Heidelberg (2013)

Neel, A.T.: Formant detail needed for vowel identification. Acoust. Res. Lett. Online **5**, 125–131 (2004). doi:10.1121/1.1764452

Pols, L.C.W.: Spectral analysis and identification of Dutch vowels in monosyllabic words. PhD dissertation, University of Amsterdam. Amsterdam, Academishe pers B.V (1977)

Strange, W., Jenkins, J.J., Johnson, T.L.: Dynamic specification of coarticulated vowels. J. Acoust. Soc. Am. **74**, 695–705 (1983). doi:10.1121/1.389855

Strange, W., Jenkins, J.J.: Dynamic specification of coarticulated vowels: Research chronology, theory, and hypotheses. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 5). Springer, Heidelberg (2013) doi:10.1007/978-3-642-14209-3_5

Watson, C., Harrington, J.: Acoustic evidence of dynamic formant trajectories in Australian English vowels. J. Acoust. Soc. Am. **106**, 458–468 (1999). doi:10.1121/1.427069

Wise, C.M.: Acoustic structure of English diphthongs and semivowels vis-a-vis their phonetic symbolization. In: Zwirner, E., Bethge, W. (eds.) Proceedings of the 5th international congress on phonetic sciences, Münster pp. 589–593. Switzerland: S. Kager (1964)

Zahorian, S.A., Jagharghi, A.J.: Speaker normalization of static and dynamic vowel spectral features. J. Acoust. Soc. Am. **90**, 67–75 (1991). doi:10.1121/1.402350

Zahorian, S.A., Jagharghi, A.J.: Spectral-shape features versus formants as acoustic correlates for vowels. J. Acoust. Soc. Am. **94**, 1966–1982 (1993). doi:10.1121/1.407520

Vowel Inherent Spectral Change in the Vowels of North American English

Terrance M. Nearey

Abstract Nearey and Assmann (1986) coined the term 'vowel inherent spectral change' (VISC) to refer to change in spectral properties inherent to the phonetic specification of vowels. Although such change includes the relatively large formant changes associated with acknowledged diphthongs, the term was explicitly intended to include reliable (but possibly more subtle) spectral change associated with vowel categories of North American English typically regarded as monophthongs. This chapter reviews statistical and graphical evidence of dynamic formant patterns in vowels of several CV and CVC syllable types in three regional dialects of English: Dallas, Texas (Assmann and Katz, 2000), Western Michigan (Hillenbrand et al., 1995) and Northern Alberta (Thomson 2007). Evidence is reviewed for the importance of VISC in vowel perception. While certain apparent VISC patterns show up across dialects, both dialect differences and differences in context make it clear that more sophisticated methods will be required to fully separate several factors affecting formant change in vowels. Promising preliminary results are presented using a new non-linear regression method that extends compositional models of Broad and Clermont (1987, 2002, 2010) to include dual vowel targets.

Abbreviations

CVC	Consonant–vowel–consonant
f0	Fundamental frequency
F1	First formant
F2	Second formant
F3	Third formant

T. M. Nearey (✉)
Department of Linguistics, University of Alberta, Edmonton, Canada
e-mail: t.nearey@ualberta.ca

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_4,

IPA International phonetic alphabet
RMS Root mean squared
VISC Vowel inherent spectral change
VOT Voice onset time

1 Introduction

Research into VISC in the phonetics laboratories at the University of Alberta was spurred by the efforts of my colleagues and I to explore the results of Jenkins et al. (1983) concerning the perception of 'silent-center' CVC stimuli. In such signals, brief sections gated from the beginnings and ends of the syllables were preserved while most of the vocalic portion was deleted, rendering the syllables (in their description) 'vowelless'. Jenkins et al. showed that vowel categories were well identified in such syllables even when the duration of the silent middle section was varied substantially. The authors suggested that the robustness of vowel perception to this extreme manipulation was likely due to co-specification of vowels and consonants by the consonantal transition information that remained in the syllable edges. Jenkins et al. noted similar good performance of complementary "centers only" condition (wherein the vocalic centers preserved, but CV and VC transitions removed).

We conducted a series of informal listening experiments in our laboratory using silent-center versions of both CVCs and isolated vowels. Expanding on earlier statistical analysis of isolated vowels (Assmann et al. 1982), we found excellent classification performance using discriminant analysis of formant and f0 measurements from two time points, one early and one late in the syllable. These preliminary observations led Nearey and Assmann (1986) to propose that the efficacy of silent-center stimuli might be largely due to residual information about vowel-inherent properties preserved near the interior edges of formant transition regions. To test this hypothesis, we conducted an analogous experiment with silent-center syllables extracted from isolated vowels and found that such 'silent-center isolated vowels' were also very well perceived.

In Nearey and Assmann (1986), we compared listeners' performance to several alternative pattern recognition models. Our analyses led us to favor a 'dual-target' specification for the majority of vowels of Western Canadian English.[1] This specification was based on measurements of formant frequencies from relatively

[1] Following Nearey and Assmann (1986), I will use the term Western Canadian English to refer to a combination of what Boberg (2008) refers to as the Prairie and the British Columbia dialects. For all of the Canadian samples discussed in this chapter, the majority were likely Prairie dialect speakers, primarily from Alberta. For brevity, the labels Alberta and Western Canadian will be used interchangeably in much of the text. However, a small minority of speakers who lived at least part of their lives in British Columbia were also included in at least some of the samples

early *nucleus* targets and relatively late *offglide* temporal portions of the vocalic signal.[2] Formant frequency changes between such nuclei and offglides were expected for the vowels /e/ (or more narrowly, [eɪ]) and /o/ ([oʊ]). These vowels are traditionally described as diphthongal, as are sometimes /i/ and /u/ (as e.g., [ij] and [uw], respectively; see Sect. 3). The latter two showed relatively little formant movement in our data. However, the front vowels /ɪ/, /ɛ/ and /æ/, usually described as monophthongs, also showed significant formant movement in the isolated vowels in the frequencies of the first and second formants (F1 and F2) or both. Nearey and Assmann proposed the term "vowel inherent spectral change" to cover both generally acknowledged cases of diphthongization and other, perhaps subtler, cases of formant movement. This awkward term was abbreviated to VISC by Strange (1989) and will be referred to as such below.

In this chapter, I review some of the evidence from production and perception supporting the importance of VISC in the vowel system of several dialects of North American English, with special emphasis on VISC patterns in lax vowels. I also present results from a new modeling technique that shows promise of successfully decomposing vowel- and consonant-directed formant movement.

2 Structure of the Chapter

Section 3 reviews traditional accounts of monophthongs and diphthongs in North American English. Section 4 outlines the earliest published evidence we have found for lax vowel VISC in production data in the lax vowels of North American English. Section 5 provides a descriptive typology of possible VISC patterns and a breakdown of the expected relations between VISC patterns and specific vowel

(Footnote 1 continued)
discussed here. For the contexts discussed here, Boberg's analysis suggests that it is likely that only the F1 and F2 values of /u/ and /æ/ are to be affected by a finer-grained dialect distinction.

[2] A reviewer has suggested that the terms "nucleus" and "offglide", which have varying interpretations in the literature, may be the source of some confusion in theoretical discussions among some of authors represented in Chap. 4. The usage below follows the relatively atheoretical senses of the terms as used by Nearey and Assmann 1986: "The terms nucleus and offglide are used for mnemonic purposes only. In particular, no attempt was made (either in the course of stimulus preparation or in measurements described below) to segment the stimuli into 'steady-state' versus 'transition' sections" (p. 82). Our goal was to allow for modest expansion of the commonly-used single target "steady-state" accounts of putative monophthongs to include some measure of spectral change. In screening the measurements of stimuli, Assmann and I experimented with re-synthesizing the stimuli based on simplified trajectories. We found that straight line interpolation of formant tracks from an early to a late target value for each formant often produced perceptual results that were remarkably similar in impressionistic vowel quality to more fully specified formant tracks. Since this simple representation was convenient for our early modeling methods, we decided to run with this idea. In the interim we have, until recently, found no compelling reason to move to more complex representations of vowel formant parameters for the problems we have studied. See Sects. 7 and 8 for additional discussion.

categories for some dialects of English. Section 6 reviews published evidence from the literature for the relevance of VISC in production and perception. Production data is analyzed for three distinct dialect regions of North America: Western Canadian English collected in Edmonton, Alberta; Inland Northern English, collected in Kalamazoo, Michigan; and Southern English collected in Dallas, Texas. (Dialect labels adopted from Boberg 2008, 2010 and Labov et al. 2006). Results from modified natural and synthetic speech supporting the relevance of VISC in perception are also summarized.

Section 7 examines in more detail statistical patterns of formant movement from two temporal slices of vowels, comparing patterns from Western Canadian English in several consonantal contexts and explores vowel-by-vowel patterns in Inland Northern and Southern English in /hVd/. Although some general patterns observed in early studies of Western Canadian English survive, some differences between dialects and contexts indicate possible limitations of simple two-target representations or of the robustness of lax-vowel VISC patterns. Section 8 explores more detailed trajectory patterns (beyond two-slice representations) for /CVC/ syllables in the Michigan data of Hillenbrand et al. (2001), with some preliminary discussion of the potential separability of vowel- and consonant-related aspects of trajectory patterns. Section 9 presents results on how the pioneering work of Broad and Clermont (1987) modeling formant trajectories in CVC syllables might be extended to elucidate the interplay of vowel- and consonant-dominated formant patterns in syllables with VISC. Finally, Sect. 10 outlines preliminary conclusions and remaining research questions.

3 Traditional Phonetic Descriptions of Monophthongs and Diphthongs

The question of which vowels show substantial formant movement is rarely discussed directly in acoustic studies of English vowels. For the most part, there is reliance on descriptions from informal or descriptive studies in linguistics or traditional phonetics texts. In describing Canadian English, the introductory textbook in Linguistics by O'Grady et al. (1997) and the introduction to descriptive phonetics by Rogers (2000) follow similar texts focusing on American English. The vowels /i/ and /e/ are also sometimes transcribed as /ij/ and /ej/, or in a non-IPA notation widely adopted in American structuralist traditions (e.g., Trager and Smith 1957) /iy/ and /ey/. Such vowels are typically said to show movement toward more extreme [i] or [j]-like values. Similarly, the vowels /o/ and /u/ (sometimes /ow/, /uw/) are described as showing change toward extreme [u] or [w]. See Ladefoged (1999) for a brief summary of alternate notations of the phonetic details. This change in vowel quality is typically described as a low-level, allophonic property. Other, more robust, diphthongs, sometimes designated "phonemic diphthongs", include /aɪ/, /aʊ/, /ɔɪ/ and perhaps /ju/. Formant movement in the latter group in most dialects is beyond

dispute, it will receive little attention below,[3] since the focus of this study is on the more subtle cases of VISC.

Regarding other vowels, there is sometimes (e.g., Hockett 1958) passing mention of centering offglides of lax vowels in some (usually Southern) dialects of North American English. The term "lax vowel" in the discussion below is taken to refer to the phonological property of (near) total absence as word-final stressed vowels in English without any other phonetic or phonological presuppositions. In the case of "General American" and "General Canadian", however, the consensus is that the lax vowels /ɪ ɛ ʊ ʌ/ are essentially monophthongal. The vowel /æ/ is usually also classed as a monophthong, though in some dialects (in some or all contexts), it is what Labov has dubbed a "tense æ". This is often described as having a higher, more fronted nucleus and centralizing offglide that could be rendered [eə] (Labov et al. 2006).[4] In other cases (e.g., Lehiste and Peterson 1961; Strange 1989; Di Benedetto 1989) any formant movement of lax vowels is attributed to coarticulation with, or a "style of movement" influenced by, consonantal context.

Broad and colleagues (Broad and Fertig 1970; Broad and Clermont 1987, 2002, 2010) appeal to related concepts in their modeling of vowel trajectories in (C)V(C). In all these works, they focus on models with single vowel targets and exclude vowels like /e/ and /o/. They also treat as exceptional or avoid altogether vowels in #(C)V# contexts. Broad and Fertig (1970) in a study of the vowel /ɪ/ in various contexts, including word final, note: "The termination of syllables with final silence, however, display configurations that are indicative of more centralized vowel articulations, and transitions away from the syllable nucleus configurations are apparently in progress by the end of such syllables" (p. 2577). Broad and Clermont (1987) avoid CV# contexts altogether because of what they call "a tendency to glide into a /CVə/ for some lax vowels in citation-form syllables".

4 Early Indications of VISC in Lax Vowels in North American English

Klatt (1980) provides a notable, though understated, exception to this general assumption of the monophthongal nature of North American English lax vowels. The alternate interpretation appears only in the caption to Table 2 (p. 986). This table displays parameters for the synthesis of selected American English vowels.

[3] Of course a full account of VISC, and perhaps especially the issue of how VISC combines with consonantal context effects, must ultimately include the phonemic diphthongs.

[4] Or even sometimes as [eæ]. See Language Samples Project (n.d.). Labov et al. (2006) also note a phenomenon they refer to as "Northern breaking" (where a former monophthong has become a bi-moraic vowel) for /æ/ in Northern Cities dialects (Labov et al. 2006, pp. 303–305). The term "breaking" is fairly widely used in largely impressionistic discussions of diphthongization in some Southern American dialects (See, e.g., Thomas 2003).

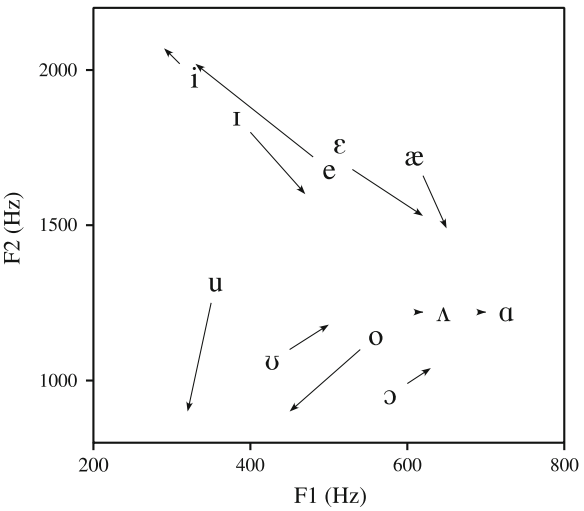


Fig. 1 Graphical representation of formant movement in Table 2 of Klatt (1980)

In that caption Klatt notes: "If two vowels are given, the vowel is diphthongized or has a schwa-like offglide in the speech of the author." A graphic representation of F1 and F2 values in this table is shown in Fig. 1. The "schwa-like" offglide is reminiscent of Broad and colleagues' characterization. What is new here is that Klatt includes the offglide as part of the definition of key acoustic properties of the vowels themselves.

The patterns in Fig. 1 bear a strong a resemblance to those of Fig. 2, adapted from Nearey and Assmann (1986), who measured the formant frequencies of 10 speakers of Western Canadian English in isolated vowels. Measurements were taken at 24 and 65 % of total vowel duration.

5 A Descriptive Typology of VISC

In what follows, it will be useful to refer to a catalog of some possible VISC patterns for North American English. Although both Broad and colleagues and Klatt refer to "schwa offglides" in certain vowels, a glance at Figs. 1 and 2 suggests another possible interpretation. This is true especially for /ɛ/, but also arguably for /ɪ/, /æ/ and even perhaps /ʊ/ and /ɔ/. In these cases it would be reasonable to posit a movement toward the third point, [a], of the classic [i–a–u] vowel triangle, where [a] here is meant to denote the lowest possible central vowel, not the standard IPA lowest front vowel. (See Barry and Trouvain 2009 for a summary of the controversy about symbol choice for a low, or in official IPA terminology, "open", central vowel.)

The following terminology will be relied on in the discussion below. (Examples in parentheses refer to Figs. 1 and 2):

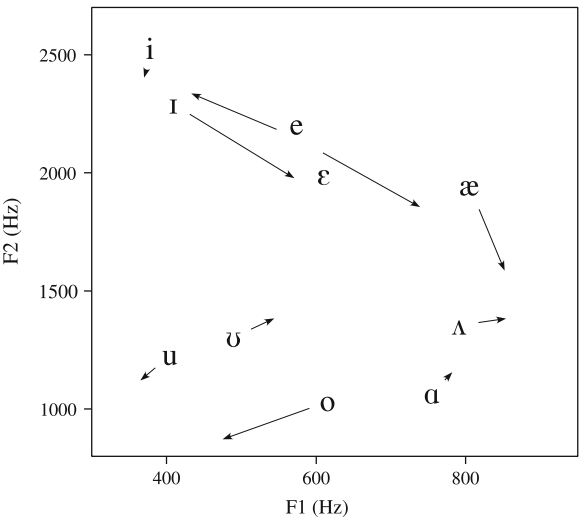


Fig. 2 A reconstruction of Fig. 2 of Nearey and Assmann (1986) showing formant movement in isolated vowels of 10 speakers of Western Canadian English. *Beginning of the shaft of the arrows* represents formant measurements at 24 % and *tips of arrowheads* at 65 % of total vowel duration

1. i-VISC or iota-VISC: movement toward [i] or [j] (/e/)
2. u-VISC or upsilon-VISC: movement toward [u] (/o/ and perhaps weakly /u/)
3. a-VISC or alpha-VISC: movement toward [a] (/ɛ/ and perhaps /æ/ and /ɔ/, also possibly /ɪ/ and /ʊ/)
4. ə-VISC or schwa-VISC: movement toward [ə], that is toward relatively neutral values of F1 and F2 (perhaps /ɪ/ and /ʊ/)

Given the overall pattern of raising F1 and centering F2, it is not clear that schwa-VISC can be distinguished from alpha-VISC. More compelling evidence of schwa-VISC would require a movement in F2 only for a vowel with an F1 starting close to the neutral position, or a movement from a high F1 and extreme F2 value at onset toward a more central F1 and F2 position.

5.1 Expected Patterns of Movement

In the discussion below, it is useful to refer to four subgroups of vowels. The divisions are based on a combination of traditional phonetic classification discussed in the introduction and on the patterns of lax vowel movement observed in Figs. 1 and 2. To organize the discussion, I will consider the movement patterns (or lack thereof) to be "expected", based on the findings of Nearey and Assmann (1986) or of Andruski and Nearey (1992), summarized in Sect. 6.

I. The mid tense vowels /e/ and /o/ with expected iota-and upsilon-VISC respectively.
II. The high tense vowels /i/ and /u/ which are said to have similar VISC patterns to /e/ and /o/ respectively in at least some dialects and or contexts.

III. The lax vowels /ɪ/, /ɛ/, /æ/ and /ʊ/, which showed at least some tendency toward alpha-VISC patterns in Nearey and Assmann (1986) or Andruski and Nearey (1992) and in Klatt's (1980) synthesis tables.
IV. The vowels /ɑ/ and /V/, which have not been explicitly identified as exhibiting any intrinsic formant movement either in the general phonetic literature or in any acoustic studies to date.

6 Existing Evidence for VISC in Production and Perception

6.1 Production Evidence in Western Canadian English

Using the terminology described above, I will review the significant formant movement patterns in the production of isolated vowels in Western Canadian English found by Nearey and Assmann (1986). See Fig. 2. The tense mid vowels /e/ and /o/ behaved as expected from traditional descriptions: /e/ showed significant iota-VISC, while /o/ showed significant upsilon-VISC in both F1 and F2. The front vowels /ɪ/ and /ɛ/ showed significant alpha-VISC in both formants. The vowel /æ/ showed a trend toward alpha-VISC with significant downward movement in F2 and a non-significant upward movement of F1. The back lax vowel /ʊ/ did not show significant movement, though there were trends in F1 and F2 suggesting alpha-VISC. The high tense vowels, /i/ and /u/ did not show significant movement, although there were trends consistent with upsilon VISC in /u/ in both formants.

It is of course possible that the alpha-VISC patterns of the isolated lax vowels are the result of some extraneous offgliding in the phonotactically unusual word-final position. Nearey and Assmann argued that this seemed unlikely in view of examples of similar movements observed in small sample /bVb/ syllables from a single speaker. This was later confirmed by Andruski and Nearey (1992) who showed similar statistically significant formant movement for vowels in /bVb/ syllables in a multi-speaker database. All the significant movement patterns shown in Nearey and Assmann (1986) for isolated vowels were confirmed by Andruski and Nearey (1992) in /bVb/ context. In addition the vowel /ʊ/ showed clear trend toward alpha-VISC with significant upward movement in F2 and a trend for upward movement in F1. Note that this pattern is in rough agreement with Klatt's (1980) target values for this vowel.

6.2 Evidence for VISC in Production Data from Two American Dialects

Hillenbrand et al. (1995) and Assmann and Katz (2000) both report substantial improvement of classification of vowels when formants from two vowel sections are used in discriminant analysis in /hVd/ syllables of Western Michigan and

North Central Texas speech respectively. Hillenbrand et al. also report significant movement from the 20 to 80 % movement for the formants of most vowels.

Assmann and Katz report significant interactions: of formant frequency change patterns with vowel category for slices taken at 33 and 66 % of vocoid duration. Although these studies did not focus on movement patterns in lax vowels, in fact the vowels corresponding to Western Canadian English /ɪ/, /ɛ/, /æ/ and /ʊ/ sometimes show trends in the same direction, as discussed further in Sect. 7.

6.3 Evidence for the Relevance of VISC in Perception

In several articles, Strange, Jenkins, and their colleagues (see Strange and Jenkins 2013 Chap. 5 for a review) present perceptual results on what they termed "silent-center syllables". These are CVC syllables in which the central, most nearly steady-state vocalic portions were excised and replaced by silence, leaving only the relatively rapidly changing portions of near the consonantal margins. These can be denoted [CV…VC]. Identification experiments showed that the silent-center syllables were recognized almost as well as the original CVCs. Furthermore, the complementary signals, the excised vowel centers […V…], were themselves recognized at about the same rate as the silent-center CVCs.

6.3.1 Windowed Natural Speech

Nearey and Assmann (1986) speculated that the similarity of the performance of silent-center syllables and the complementary vowel centers might indicate that the information about vowel identity was essentially the same. Specifically, the formant frequency values at the internal edges of the silent centers (i.e., the right edge CV part and the left edge of the VC part) should be very similar to the information at the outer edges of the excised vowel centers […V…]. If listeners' judgments were strongly influenced by the formant patterns near these junction points, then it would not be surprising that performance of the two physically very different stimuli was about the same. If this were so, then similar manipulations of edges and centers of isolated vowels should yield similar results.

To explore this idea, Nearey and Assmann adopted a working hypothesis of dual-target specification of English vowels, consisting of a nucleus target followed by a transition to an offglide target. VISC could thus be characterized by nucleus or 'head' values early in the vowel and offglide or 'tail' values late in the vowel. "Early" and "late" were defined rather arbitrarily following the rough desideratum that they be far enough from edges to avoid unstable measurements and breathy offsets in isolated vowels and to avoid rapid formant CV and VC transitions in CVCs. (See Assmann et al. 2013 Chap. 1 for evidence regarding the sensitivity of the distinctness of vowel patterns to precise time selection points).

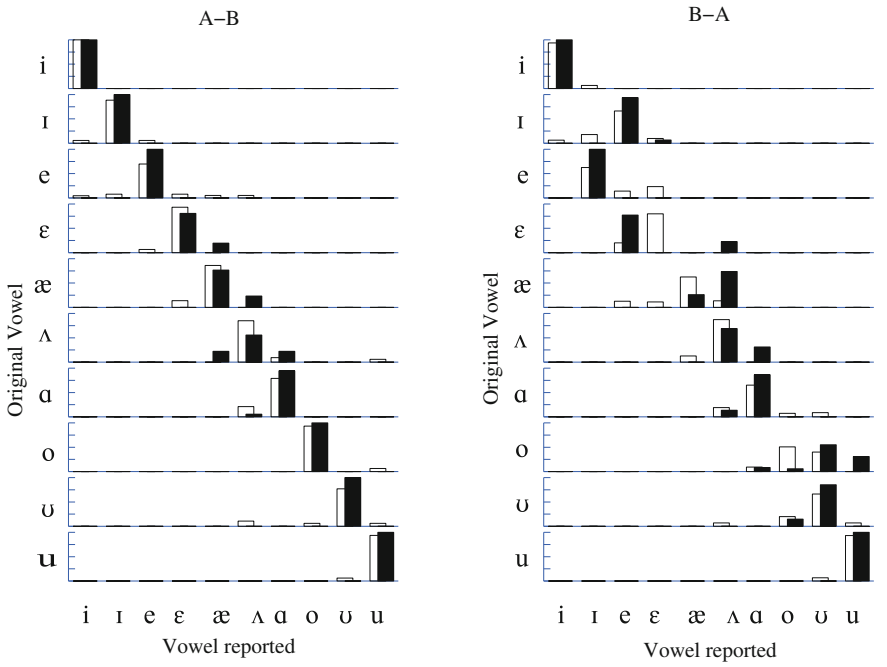


Fig. 3 Identification of original order (A–B) and reversed (B–A) gated isolated vowel snippets. After Nearey and Assmann (1986)

To characterize the dual targets in isolated vowels, Nearey and Assmann chose two brief windowed sections. The nucleus section, labeled A, was centered at 24 % of vowel duration. The offglide section, labeled B, was centered at 64 % of the duration. (I will return to the possibility of more complex characterization below.) When these windowed sections were played in their natural (A–B) order (with a short 10 ms silence between them), they were identified at almost as well (85.6 % correct) as the unmodified full-duration vowels (87.5 % correct).

However, reversing time order of head and tails produced important perceptual changes (see Fig. 3). When the order of the sections was reversed (B–A) performance fell substantially (to 62.5 % correct). The most dramatic change in the B–A condition is the reversal of dominant responses to original /ɪ/ and /e/; specifically, most original /e/ were identified as /ɪ/ and vice versa.[5]

A simple pattern recognition algorithm (linear discriminant analysis) was trained on F1 and F2 measurements from nucleus and offglide sections (together with average f0) from the unmodified syllables. This simple model was able to predict the general effects of the time reversal with fairly high accuracy.

[5] The general flavor of this reversal can be demonstrated by using a waveform editor to time reverse syllables containing /I/ or /e/ in many tokens of /hVd/ syllables of the publically available Hillenbrand et al. (1995) database.

6.3.2 Resynthesized Speech

Andruski and Nearey (1992) provide evidence of a reasonably good fit of a similar pattern recognition model to listeners' perception of /bVb/ syllables, including resynthesized 'hybrid' synthesis, where head formant patterns from one speaker were combined with tail formant patterns from different speakers in a manner inspired by the natural vowel cross-splicing experiments of Verbrugge and Rakerd (1986).

There is additional indirect evidence for the perceptual relevance of spectral change in vowels of two other dialects of North American English: Hillenbrand and Nearey (1999) for Western Michigan, Assmann and Katz (2000) for North Texas. These experiments included "flattened" formant trajectories based on vowels in naturally produced /hVd/ stimuli. Results show that resynthesized vowels are more intelligible when formant movement modeled after measured natural speech contours is preserved, compared to resynthesis where the vocalic portion is replaced with steady-state trajectories. Such resynthesis may contain information other than the kind of simple dual-target VISC patterns considered here. However, pattern recognition studies by Hillenbrand and Nearey (1999) and Hillenbrand, Clark and Nearey (2001) show significant correlations of listeners' response patterns for natural speech stimuli with explicit dual-target characterizations similar to those of Nearey and Assmann (1986) for Western Canadian English.

Continuum studies. The case for the perceptual relevance of VISC (and of alpha-VISC in particular) for front vowel contrasts of Western Canadian English is confirmed in two studies of phonetic continua. Nearey (1995) showed that alpha-VISC in F1 and F2 produced substantially different classification patterns in front vowels compared to vowels with no formant movement or other movement patterns. Furthermore, listeners' naturalness judgments for front lax vowels were higher for alpha-VISC patterns than for steady-states.

Morrison (2006, 2008) provides a dramatic illustration of the effects of VISC patterns in front vowels of Western Canadian English. (See also Morrison and Nearey 2007). Morrison produced a three-dimensional continuum spanning the front vowels /i ɪ e ɛ/ in a /ˈbVpə/ context. The first dimension consisted of a correlated F1 and F2 onset-target continuum F1 ranged from 283 to 580 Hz in 10 steps. F2 changed in a negatively correlated fashion as appropriate for front vowels of increasing height. The second dimension, vowel duration was set at three different levels ranging from 80 to 110 ms. The third dimension involved three vowel offset conditions, a steady-state condition and two VISC conditions, one corresponding to iota-VISC and the other to alpha-VISC. A summary of the pooled results is shown in Fig. 4.

The effects of VISC are dramatic and in the direction expected from Nearey and Assmann (1986) patterns. With iota-VISC, only tense /i/ and /e/ are heard. With no VISC, /i/ and /e/ dominate, except at highest F1 and at shorter durations, where /ɛ/ emerges. Only one stimulus in the no-VISC condition shows dominant /ɪ/ responses. Only in the alpha-VISC condition do lax /ɪ/ and /ɛ/ responses emerge as dominant.

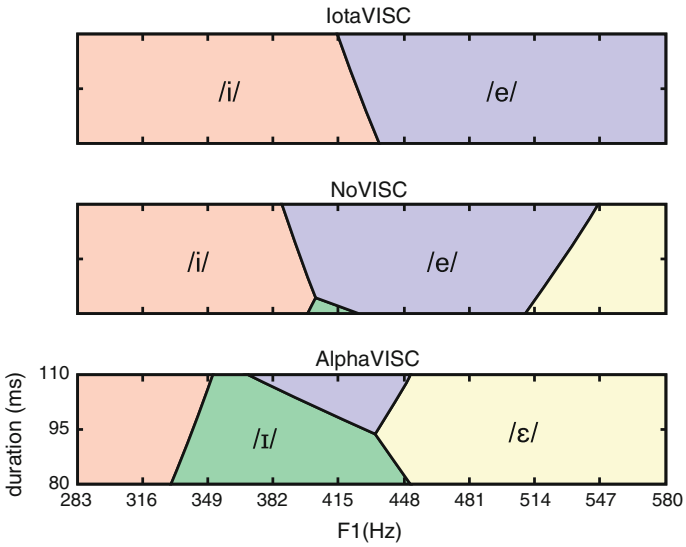


Fig. 4 Identification of an F1 continuum at three different durations and VISC patterns

7 Further Analysis of VISC in Production Data for Varying Contexts and Dialects

The evidence reviewed above clearly suggests that dual-target VISC patterns have perceptual relevance in at least some circumstances in at least some dialects of North American English. However, it remains to be demonstrated whether such patterns, and in particular the alpha-VISC patterns observed for lax vowels, represent stable production patterns across a variety of environments, even in careful speech, and which, if any, lax vowel VISC patterns are stable across several dialects for which dual targets models have been proposed. The remainder of this paper aims to fill in some of the gaps in our knowledge.

7.1 VISC in Several Contexts in Western Canadian English

In this section, I consider additional data from Western Canadian English (Edmonton, Alberta) for which direct evidence of the perceptual relevance of VISC in isolated vowels in /bVb/ and /bVpə/ contexts is available. Four additional contexts, /pV/, /bV/, /pVt/, and /bVt/, are considered. While richer representations will be discussed below, for now the focus is on a simple two-slice representation

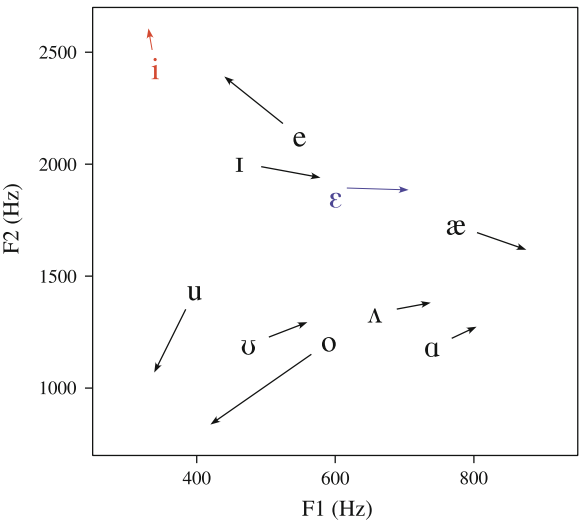


Fig. 5 Vowels from Alberta English in /bV/ contexts. *Beginnings of shafts of arrows* indicate 20 % of vowel duration and *points of arrowheads* indicate 70 % of duration. *Black* significant F1 and F2 movement; *blue* significant movement in F1 only; *red* significant F2 movement only

taken 20 and 70 % of vocoid duration.[6] It will become clear that context matters. Furthermore, the variability of apparent VISC patterns raises serious questions about the adequacy and stability of the simple dual target characterization. The focus here is on graphic presentation of data together with an indication of which formant movement patterns are significant. Details of the statistical analysis are presented in the Appendix.

For the Alberta /bV/ syllables shown in Fig. 5 the mid vowels /e/ and /o/ show significant movement in the expected direction. The high tense vowels /i/ and /u/ show at least small movement in the direction expected by transcriptions like [ij] and [uw], although F1 movements are very small and not significant for /i/. The Group III lax vowels /ɪ ɛ æ ʊ/ all show movement in directions consistent with alpha-VISC and all changes are significant except for F1 of /ɛ/. Thus far, the patterns are generally compatible with the other studies of Western Canadian English.

However, the Group IV vowels /V/ and /ɑ/ show unexpected movement in the direction of increasing F1 and decreasing F2 across the vowels. Perhaps they also involve alpha-VISC. However, it is also possible that this pattern is due to effects of the surrounding context, as will be discussed below.

Figure 6 shows a pattern of formant movements in /pV/ syllables similar to the /bV/ case. However, the magnitude of alpha-VISC like movement is generally diminished. Although /V/, /æ/ and /ɑ/ all show some significant movement, it is generally quite small in magnitude. Could the additional apparent movement of the formants in the /bV/ context of Fig. 5 be simply due to consonantal effects?

There are several kinds of effects that deserve consideration. First, F1 is expected to be lower in the early portion of a vocoid following a voiced stop (see

[6] These time values were found to be the best of those explored in pattern-recognition studies of stop+vowel+stop syllables in Michigan speakers by Hillenbrand et al. (2001).

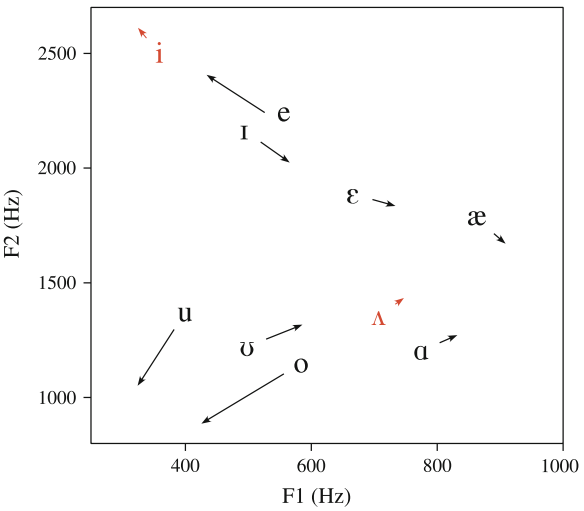


Fig. 6 Vowels from Alberta English in /pV/ contexts. *Beginnings of shafts of arrows* indicate 20 % of vowel duration and *points of arrowheads* indicate 70 % of duration. *Black* significant F1 and F2 movement; *red* significant F2 movement only

Kingston and Diehl 1994 "low frequency property" for extensive references). Second, F2 might also be expected to be low due to transition from a low F2 locus (Sussman et al. 1991) for /b/. These two effects together could account for the generally larger movement in F1 and smaller movement in F2 the /bV/ and /pV/ for lax front vowels. This might also account for the bulk of the movement in lax /V/ and tense /ɑ/ in /bV/ syllables.

A third factor to consider involves possible side effects of positive VOT in /pV/ syllables. The 20 and 70 % points for measurement are defined relative to duration of the voiced signal. Measurement of the 20 % point following voiceless aspirated /pV/ is relatively later in the entire syllable compared to /bV/ contexts. This could diminish apparent CV transition effects of the preceding paragraph. It also might mean that the 20 % measurement points represent a relatively later point in a vowel-intrinsic movement pattern.

Finally, the previous two examples involved open syllables, where the possibility of context-specific "schwa-like" offglides has been raised by Broad and colleagues (Broad and Fertig 1970; Broad and Clermont 1987). While similar alpha-VISC patterns have been observed in /bVb/ syllables for Edmonton English by Andruski and Nearey (1992), the question has not been studied in detail in other CVC contexts. We must therefore consider other contexts.

Figure 7 presents a summary of measurements from Thomson (2006) for /bVt/ syllables. Here, we see general agreement in the movement patterns of the usual alpha-VISC group III suspects /ɪ ɛ æ ʊ/; however, large F1 movements are also found for /ɑ/ and /ʌ/, similar to those seen in /bV/ context. This suggests again that much of the upward movement of F1 is due to carryover from initial voiced stop. That is, qualitatively, the /bVt/ environment could favor upward movement in F1 through the vowel. This is so because the initial /b/ could be expected to lower F1 near the beginning of the syllable, while a voiceless final stop would not show

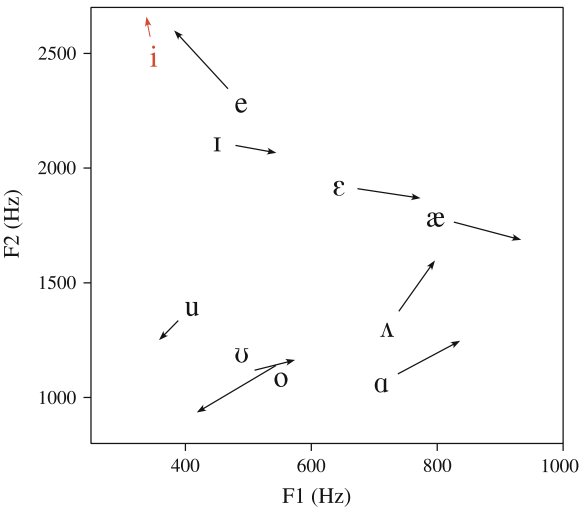


Fig. 7 Vowels from Alberta English in /bVt/ context. *Beginnings of shafts of arrows* indicate 20 % of vowel duration and *points of arrowheads* indicate 70 % of duration. *Black* significant F1 and F2 movement; *blue* significant movement in F1 only; *red* significant F2 movement only

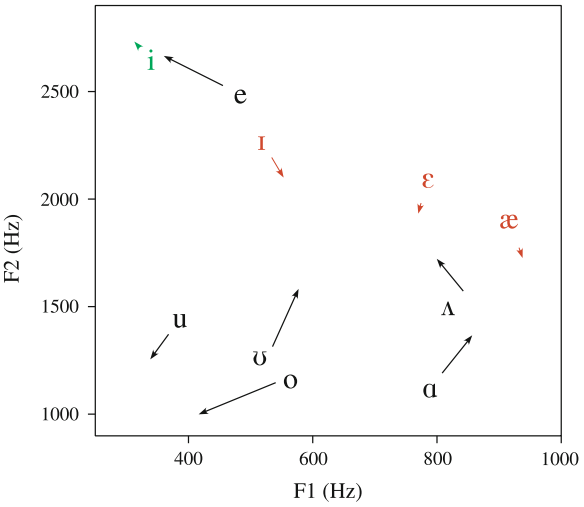


Fig. 8 Vowels from Alberta English in /pVt/ context. *Beginnings of shafts of arrows* indicate 20 % of vowel duration and *points of arrowheads* indicate 70 % of duration. *Black* significant F1 and F2 movement; *blue* significant movement in F1 only; *red* significant F2 movement only. *Green* no significant movement of F1 or F2

such a lowering. Similarly, movement toward a middle F2 value might suggest simple movement toward an F2 locus for the final alveolar stop (Sussman et al. 1997).

The analysis presented above for F1 is buttressed somewhat by considering Fig. 8 involving Alberta vowels in /pVt/ frames. Here there is no significant upward movement of F1 for any of the front lax vowels. It is perhaps noteworthy that the mean duration of the voiced vocalic region of these syllables is the shortest of the four examined here, because of both the voiceless and murmured portions after the initial consonant release and the foreshortening of the vowel associated with final voiceless stops.

The question raised by the variation of VISC patterns in the four contexts shown above is: How much of what we have been analyzing as alpha-VISC is really just consonantal context effects? The presence of movement in vowels such as /V/ and /ɑ/, which showed no movement in other contexts, suggests clearly that some apparent VISC is a side effect of consonantal context. However it is unlikely that all observed VISC patterns so far can be dismissed in this way. First, many of the lax vowel alpha VISC patterns observed above were also found in /#V#/ and /bVb/ in the studies reviewed in Sect. 6.1. Second, several perceptual experiments reviewed in Sect. 6.2 show clear evidence for the relevance of alpha VISC at least for the front vowels /ɪ/ and /ɛ/. Finally, there is evidence from other dialects that significant rising F1 VISC persists at least for the high lax vowel /ɪ/ in contexts that should be antagonistic to such changes.

7.2 *Evidence for VISC in /hVd/ Contexts in Michigan and Texas English*

Figure 9 shows results from the dataset of the 93 speakers of Hillenbrand et al. (1995) for vowels in /hVd/ environment, collected in Western Michigan. It is to be expected that the initial /h/ has relatively little effect on the formant trajectories of the early part of the vowel.[7] The final consonant, /d/ should be expected to lower F1 at the end of the syllable (Summers 1987). The fact that the vowels /ʌ/, /ɑ/ and /ɝ/ show significant downward movement in F1 seems likely to be due to this factor. Given this negative-trending influence of the /hvd/ context, any upward movement F1 is almost certainly due to vowel-inherent characteristics. Both /ɪ/ and /ʊ/ show highly significant upward movement of F1 in these conditions. /ɛ/ shows very little movement, though it is significant (evidence, perhaps, of the statistical power of a large sample size). The "low vowel" /æ/ shows a very robust upward formant movement. This is almost certainly due largely to presence of the Northern Cities dialect raised or tense /æ/ in these speakers. This vowel is widely acknowledged to have a strong offglide pattern. The vowel /ɔ/ is often also characterized in impressionistic phonetic accounts as showing salient diphthongization in many dialects (See Labov et al. 2006).

The interpretation of F2 movement pattern is not entirely straightforward because of the likely contamination of movement toward the final /d/ locus near the middle of the F2 space. Only the tense high and mid Group I and II tense vowels /i u/ and /e o/ show movement unambiguously in a contrary direction for F2. On the other hand, the F2 movement patterns of /ɪ/, /ʊ/ and /æ/ are consistent

[7] There might be some concern about effects of murmured voice for the first couple of glottal pulses following /h/. While this may indeed affect F1 measurements very early on in the syllable, any such effects seem unlikely to persist in measurements taken as late as 20 % of vowel duration.

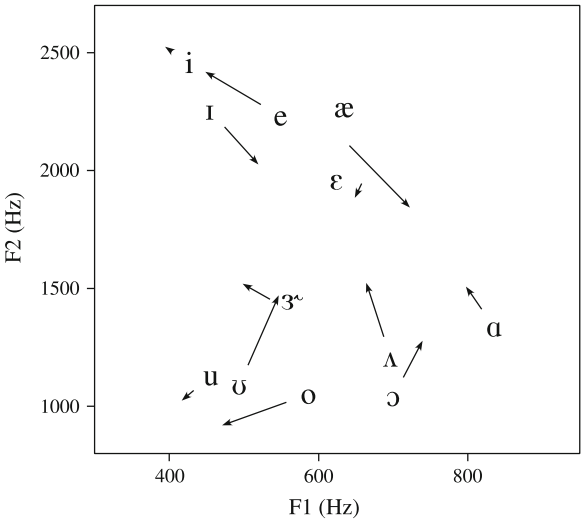


Fig. 9 Vowels from Western Michigan in /hVd/ context. *Beginnings of shafts of arrows* indicate 20 % of vowel duration and *points of arrowheads* indicate 70 % of duration. (All F1 and F2 movements are significant)

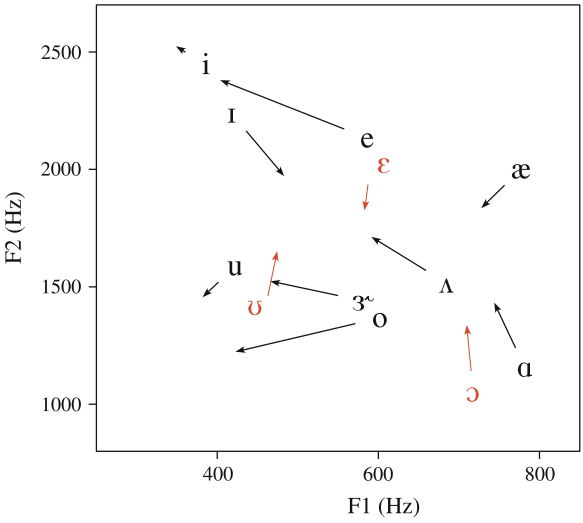


Fig. 10 Vowels from North Texas in /hVd/ context. *Beginnings of shafts of arrows* indicate 20 % of vowel duration and *points of arrowheads* indicate 70 % of duration. *Black* significant F1 and F2 movement; *red* significant F2 movement only

with those observed in Alberta English in vowel final contexts, and for that matter, with Klatt's (1980) formant synthesis tables.

Figure 10 shows data from 20 speakers in /hVd/ syllables collected by Assmann and Katz (2000) in North Texas. The only lax vowel showing significant upward F1 movement is /ɪ/. The fact that the vowels /ɝ/, /V/, /æ/, /ɑ/ and /ɔ/ show significant downward F1 movement is again suggestive of the influence of the final /d/. From this perspective, it is perhaps noteworthy that lax /ʊ/ shows at least a non-significant rising F1 trend that bucks this general tendency. Once again, for F2, the movements of the lax vowels and low tense vowels might be due to F2 locus effects of the final alveolar. Again, the Group I and II tense vowels /i e o u/

show significant (though for /i/ very small) movements in a direction contrary to the expected 1800–1900 Hz /d/ locus.

8 A More Detailed Look at Some CVC Trajectories in Michigan English

From the above discussion it seems clear that much observed formant movement in the 20 to 70 % temporal region of the vocalic portion in CV(C) contexts must surely be strongly influenced by consonantal context effects. The question must be raised again: Is there any alpha-VISC left in lax vowels or can it all be attributed to consonantal effects? While some qualitative evidence has been given above that at least some residual lax-vowel VISC patterns are not consistent with expected consonantal effects, more direct evidence needs to be brought to bear if the concept is to have any force. In the following, I present a sketch of two lines of research my colleagues and I are pursuing. The first is a more complete graphic depiction of F1–F2 trajectories. The inspiration of these plots comes from those developed by Assmann et al. (2013 Chap. 1).[8]

Summers (1987) demonstrated that initial and final consonantal voicing can have effects that influence virtually the entire formant pattern of the vowel. As discussed in several cases above, such effects might either enhance or mask apparent VISC pattern. It seems reasonable therefore to consider first syllables in symmetrical voicing contexts, specifically in [voiced stop + vowel + voiced stop] and [voiceless stop + vowel + voiceless stop] contexts where any effects of voicing on F1 would be relatively symmetrical, rather than differentially affecting early versus late parts of the vowel formant trajectories.

Figure 11 shows data from the study of Hillenbrand et al. (2001) for 12 speakers recorded in Western Michigan for vowels in [voiced stop + vowel + voiced stop] frames. This study was a replication and extension of the study of Stevens and House (1963) and includes only vowel categories corresponding to those considered by the latter authors as monophthongal. With the exception of the vowels /i/ and /u/, all vowels show a general looping upward movement of F1 at the beginning followed by downward movement toward the end. A general tendency of F2 to end near a central value is also evident.[9] Although most vowels share a generally similar overall trajectory pattern, the timing of the movements is quite different. Lax /ʌ/ and tense /ɑ/, which have shown little or no formant movement in several of the cases

[8] A reviewer pointed out that Neel (2004) presented similar diagrams for one male and one female speaker (both from Central Indiana) for vowels in a /dVd/ frame.

[9] Interpretation of this F2 movement as a final consonant locus effect is somewhat more nuanced than in the case of /hVd/. At least some of the labial and velar contexts would lead to more extreme, rather than centralized F2 values. The net effects averaged for all three places of articulation may, however, be similar. Rather than pursuing this issue graphically in the several contexts, the question is dealt with implicitly in the trajectory modeling section below.

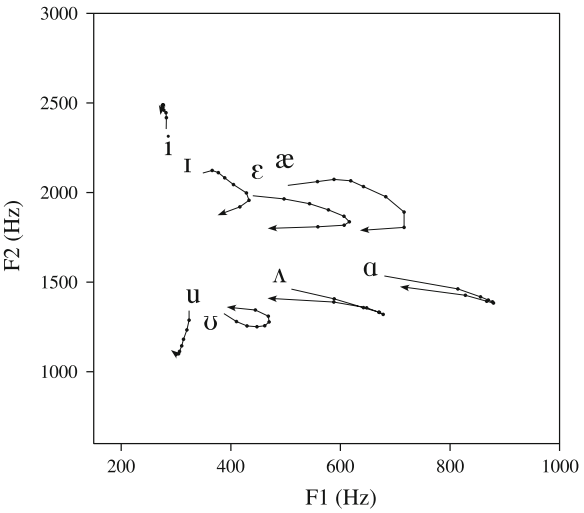


Fig. 11 Average trajectories of [voiced stop + V + voiced stop] syllables measured at intervals of 10 % of vowel duration. *Start of line near symbol* indicates measurements at 10 % of vowel duration. *Subsequent dot symbols moving toward arrowhead* indicate at 20, 30 and so on to 80 % and *tip of arrowhead* indicates measurement at 90 % of duration

discussed above, here show rapid movement from 10 to about 30 % then little movement until about 70 %, followed by rapid downward movement to the 90 % point. The pattern seems consistent with an analysis of a single vowel target modulated by voicing effects of the surrounding context.

The vowel /æ/, which is a raised or tensed version for the Michigan speakers, shows a very different F1 movement pattern: A rather slow and steady upward movement from 10 to 70 %, followed by a very rapid decrease. Although the magnitude of the movement is less extreme, the lax vowels /ɪ/, /ɛ/ and /ʊ/ show F1 movement patterns relatively similar to /æ/, with rather steady movement to a late maximum, followed by rapid movement to lower F1 values. The "late F1 maximum" aspect of the movement patterns in F1 seems clearly related to phenomena discussed by Di Benedetto (1989). In that work, however, the late maximum was illustrated only for the vowel /ɪ/ and contrasted with /ɛ/, which showed a flatter trajectory. In the Michigan data, in these contexts at least, the pattern is quite general.

Generally similar patterns of movement are observed for these same vowels in [voiceless stop + vowel + voiceless stop] environments as shown in Fig. 12. As expected from (e.g.) Summers (1987), the achieved maximum F1 frequencies are somewhat higher in the voiceless stop environments. However, the general shapes and relative timing of the movement patterns are otherwise remarkably similar to those in the voiced consonantal environments.

While other interpretations may be possible, the patterns observed in the last two figures seem consistent with dual-targets for the lax vowels /ɪ/, /ɛ/, /ʊ/ and the raised /æ/, one relatively early in the syllable and one relatively late. However, any such vowel-inherent patterns must be viewed as combining with consonant-directed formant movement. Though the consonant-directed aspects are most prominent at the syllable margins, they likely persist to some degree throughout the duration of the vocoid.

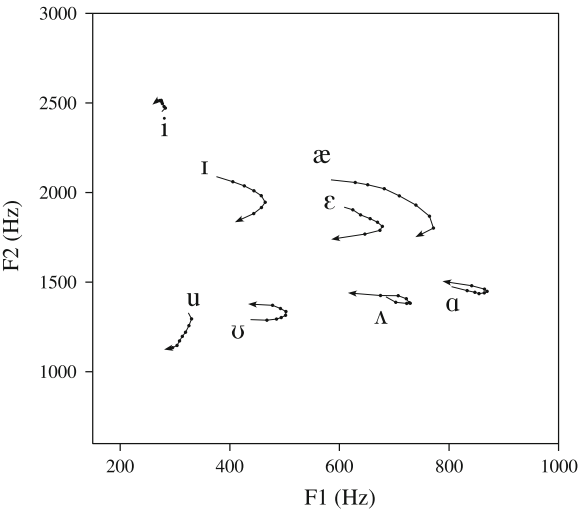


Fig. 12 Average trajectories of [voiceless stop + V + voiceless stop] syllables measured at intervals of 10 % of vowel duration. *Start of line near symbol* indicates measurements at 10 % of vowel duration. *Subsequent dot symbols moving toward arrowhead* indicate at 20, 30 to 80 %, and *tip of arrowhead* indicates measurement at 90 % of duration

9 Towards a Formal Model of VISC and Consonantal Context

The graphical methods above are useful for gaining intuitive insight into possible sources of variation in formant trajectories in a relatively non-parametric way. However, some formal modeling is necessary to adequately test competing hypotheses. In this section, I describe some preliminary results from attempts to do just that, extending methods developed by Broad and Clermont (1987). This analysis was first presented in Nearey (2010).

9.1 Description of the Model

The data used are a superset of those described by Hillenbrand et al. (2001) but included initial /h/ as well as voiced and voiceless stops. There were 12 speakers. The syllables consisted of all possible sequences of the [C1, V, C2] patterns[10] where C1 consisted of the consonants /h p t k b d g/, V consisted of the vowels /i ɪ ɛ æ ʌ ɑ ʊ u/, and C2 comprised the stops /p t k b d g/.

Broad and Clermont (1987) describe several related models[11] to account for the formant trajectories of vowels they consider monophthongs. The authors achieve

[10] There were a few missing values, but the non-linear regression techniques used here do not require fully balanced data.

[11] The model described below is very close to Broad and Clermont's Model IVb, except that the current model uses normalized time (as does their model IVa), dual vowel targets and slope-intercept specification for vowel. See also Broad and Clermont (2002, 2010) for related models.

what appear to be quite good fits from a model that consists of three components sketched below:

1. A single (steady state) target per formant per vowel.
2. An exponential decay (in time from onset) toward the vowel target from an initial consonant onset value related to a consonantal locus.
3. An exponential decay (in time from offset) toward the vowel target from a final consonant offset value related to a consonantal locus.

Subsequent development of closely related modeling frameworks by Broad and Clermont (2002, 2010) have shown very promising results with super positional models of this general kind, where vowel-directed effects can be parsed out statistically from consonantal aspects of the trajectories.

Encouraged by the success of Broad and Clermont, I sought to develop a related model that could be used to evaluate formant movement related to VISC. The extended model allows for VISC by implementing the following modification:

1. a compound vowel target including *nucleus* and *offglide* portions per formant per vowel.
   The extend model also generalizes the locus equation work of Broad and Clermont (2010) as follows:
2. and 3. Both C1 onset and C2 offset frequencies are calculated based on a locus-equation approach. (Sussman et al. 1991, 1997). The locus equations for C1s are based on nucleus target frequencies, and those for C2s are based on offglide target frequencies.

More specifically for F1 and for F2 separately, the vowel-inherent component trajectory is specified by:

a. a vowel nucleus target (first 20 % of vocoid), specified as $T1_v$
b. a linear transition (middle 60 % of vocoid) from nucleus to offglide
c. a vowel offglide target frequency-state (last 20 % of vocoid), specified as $T2_v$

The full trajectory equation the entire duration of the vocoid for a single formant will be denoted by $F_{c1,v,c2}(t)$, where t is normalized time ranging from 0.0 to 1.0; and where the three indices *c1*, *v* and *c2* index the initial consonant, vowel and final consonant respectively. In a manner similar to Broad and Clermont (1987) the trajectory is decomposed into three components

$$F_{c1,v,c2}(t) = C_{c1}(t) + V_v(t) + D_{c2}(t) \quad (1)$$

where $V_v(t)$ is the context-independent vowel trajectory component, and $C_{c1}(t)$ and $D_{c2}(t)$ are time dependent deviations from $V_v(t)$ trajectory. Some insight into

(Footnote 11 continued)
All their argumentation is exquisitely set out and carefully evaluated on well-designed single-speaker datasets. Clermont and Millar (1986) do, however, explore the Broad and Clermont (1987) approach on /CVd/ data from three Australian male speakers.

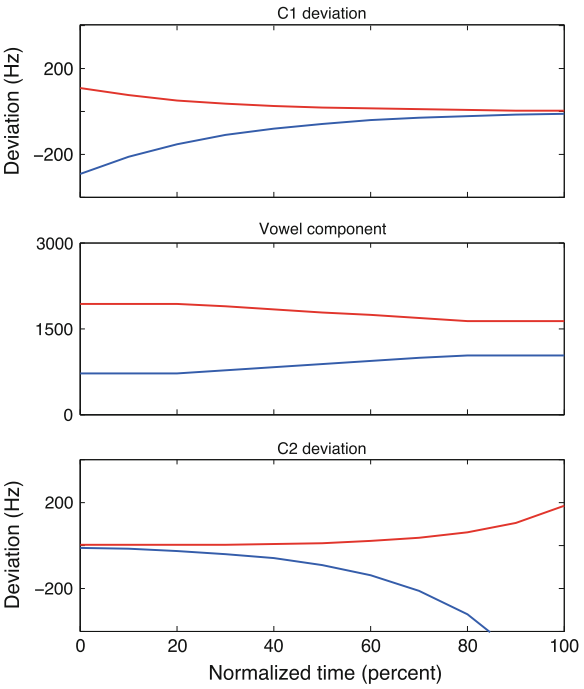


Fig. 13 Decomposition of the fitted model of Fig. 14 for the syllable /dæg/. *Top panel* predicted contribution of C1 deviation function $C_{c1,v}(t)$ as deviation from vowel component. *Middle panel* predicted contribution of vowel-inherent trajectory $V_v(t)$. *Bottom panel* predicted contribution of C2 deviation $D_{c2,v}(t)$ (*blue lines* F1, *red* F2)

the key properties of these components can be had with reference to Fig. 13, which illustrates both F1 and F2 patterns for a single syllable.

The vowel-specific component, $V_v(t)$, is specified by a piecewise linear function of time:

$$V_v(t) = \begin{array}{ll} T1_v & \text{if } t \leq 0.2 \\ T1_v + (t - 0.2)(T2_v - T2_v)/0.6 & \text{if } 0.2 < t < 0.8\,. \\ T2_v & \text{if } t \geq 0.8 \end{array} \tag{2}$$

This specifies an initial nucleus steady state target at frequency $T1_v$ for the first 20 % of the duration of the vocoid, a second steady state offglide target at frequency $T2_v$ for the final 20 % and a linear transition between the two. The middle panel of Fig. 13 illustrates an example pattern. The contribution of the initial consonant $C_{c1,v}(t)$ is specified as

$$C_{c1,v}(t) = \left(L_{c1,v} - T1_v\right) exp(-P_{c1}t) \tag{3}$$

where P_{c1} is a consonant-specific exponential decay parameter. Thus $C_{c1,v}(t)$ represents a deviation from the nucleus vowel target $T1_v$ that reaches its maximum magnitude at $t = 0$, where it equals the full difference between the locus $L_{c1,v}$ (defined below in Eq. 4) and the vowel target. During the time course of the trajectory, the magnitude of the deviation decreases toward zero because, with relative time t ranging from 0 to 1.0, the expression $exp(-P_{c1}t)$ has its maximum

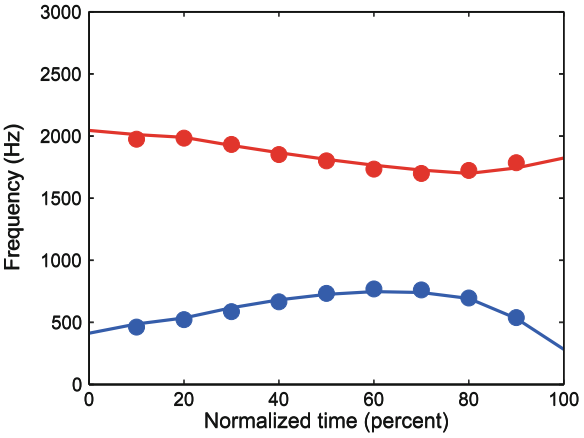


Fig. 14 Fitted trajectory (*lines*) and observed trajectory (*filled circles*) for a single syllable using the extended non-linear model. *Blue* indicates F1; *red* F2

of 1.0 at $t = 0$, and it approaches 0 asymptotically as t increases, reaching a value of $exp(-P_{c1})$ at the end of the vocoid trajectory at $t = 1.0$. The contribution of $C_{c1,v}(t)$ for a sample syllable is shown in the top panel of Fig. 14.

The initial consonant locus is specified by a Sussman et al. style locus equation

$$L_{c1,v} = a_{c1}T1_v + b_{c1} \tag{4}$$

where a_{c1} and b_{c1} are a initial consonant-specific slope and intercept coefficients that specify a predicted frequency at syllable onset, given the nucleus target for the vowel $T1_v$.

The contribution of the final consonant $D_{c2,v}(t)$ is specified as

$$D_{c2,v}(t) = \left(M_{c2,v} - T2_v\right)exp[-Q_{c2}(1-t)] \tag{5}$$

where Q_{c2} is a consonant specific exponential decay parameter. $D_{c2,v}(t)$ is a deviation from the offglide vowel target $T2_v$ that reaches its maximum magnitude at $t = 1$, where it equals the full difference between locus $M_{c2,v}$ (defined in Eq. 6) and second vowel target $T2_v$. It decreases in magnitude as t approaches 0.0 (representing 0 % of the trajectory duration), where it reaches a minimum value (within the syllable) of $exp(-Q_{c2})$. The contribution of $D_{c2,v}(t)$ for a sample syllable is shown in the top panel of Fig. 14.

The final consonant locus is specified by another Sussman et al. style locus equation

$$M_{c2,v} = a'_{c2}T2_v + b'_{c2} \tag{6}$$

where a'_{c2} and b'_{c2} are a final consonant-specific slope and intercept coefficients that specify a predicted frequency at syllable offset, given the nucleus target for the vowel $T2_v$.

It is important to note that, as in the Broad and Clermont (1987) models, the core parameters, those on the right sides of Eqs. 2–6, are indexed only by a single

phoneme-level index, $\{c1, v, c2\}$. The apparent exceptions to this claim $L_{c1,v}$ and $M_{c2,v}$ are typographic conveniences that abbreviate expressions that themselves meet these conditions. Importantly for present purposes, as in Broad and Clermont's models, the vowel-specific target parameters can be separated from consonant-directed deviation patterns.

9.2 Model Fits

The parameters of this model are fitted for individual talkers using a non-linear least squares approach, using the MATLAB program *nlshybrid* in the *immoptibox* package of Nielsen (2009).

Figure 14 shows the composite trajectory for one of the better fitting /dæg/ syllables for one of the speakers (Figure 13 shows the initial consonant, vowel and final consonant components for this same syllable are added to yield the composite track shown by the solid lines).

Table 1 shows results of *t*-tests of coefficients representing vowel inherent spectral change (offglide $T2_v$ minus nucleus $T1_v$ values) for F1 of each vowel. The tests are based on the methods of Lorch and Myers (1990), where regression models are first fitted to individual subject's data, and then *t*-tests are performed on coefficients (or contrasts) of interest across subjects.[12] If Broad and Clermont's single target account of vowels were adequate, we would expect the offglide components to be estimated at approximately the same value as the nuclei, and consequently for the spectral change measures reported below to fluctuate around zero across talkers. Instead, there are numerous significant trends across talkers. The vowels /ɪ/, /ɛ/, /æ/ and /ʊ/ show significant upward movement of F1 as expected by earlier hypotheses. The lax /V/ and tense /ɑ/ do not show such movement in F1. There is, however, one unexpected result. Namely F1 of /i/ shows an unexpected significant increase in F1. However, the lax vowel alpha-VISC suspects show positive movements that more than twice as large as that observed for /i/.

Table 2 shows results of *t*-tests of coefficients representing vowel inherent spectral change (offglide minus nucleus values) for F2 of each vowel. The vowels /ɪ/, /ɛ/, and /æ/ show significant downward movement of F2 as expected by earlier hypotheses. The vowel /ʊ/ shows significant upward movement as expected, but vowels /ʌ/ and /ɑ/ show significant unexpected upward movement as well (though substantially smaller in magnitude than that of /ʊ/). The tense vowel /u/ shows a significant downward movement of F2, but this is not unexpected at least for diphthongal accounts of this vowel.

[12] Fitting to individual subjects has two main advantages (1) it avoids the need for normalization for the tests of VISC of interest here and (2) for intuitive and relatively robust (see Gumpertz and Pantulla 1989) statistical tests of generalization of coefficient patterns across speakers.

Table 1 Estimated deviations of offglide $T2_v$ minus nucleus $T1_v$ for F1

F1 (Hz)	i	ɪ	ɛ	æ	ʌ	ɑ	ʊ	u
Mean	43	132	135	263	23	72	93	10
t value	4.091	5.194	3.328	4.5	0.665	1.842	3.883	0.812
p value	0.00179	0.0003	0.00673	0.0009	0.51955	0.09255	0.00255	0.43411

Table 2 Estimated deviations of offglide $T2_v$ minus nucleus $T1_v$ for F2

F2 (Hz)	i	ɪ	ɛ	æ	ʌ	ɑ	ɪ	u
Mean	31	−247	−171	−303	99	76	187	−74
t value	1.065	−5.188	−3.599	−4.479	5.709	6.102	7.265	−5.073
p value	0.30968	0.0003	0.00418	0.00093	0.00014	0.0000	0.0000	0.00036

In sum, even after accounting for consonantal context in the best way yet known, there is ample evidence that there are consistent deviations of second targets that are roughly in accord with the general patterns described in the graphical analyses above. Many of the alpha-VISC patterns observed for the class III lax vowels [i ɪ ɛ æ] observed in Alberta English by Nearey and Assmann (1986) and Andruski and Nearey (1992) for isolated vowels and /bVb/ respectively are evident, on average, in the Michigan CVC data after controlling statistically for consonantal context.

9.3 Comparisons with Other Models and Directions for Future Refinements

While this model generally supports the graphic analysis of this data in the previous section, some caution is in order. There is no guarantee that average effects actually due to consonantal context are not sometimes incorrectly "parsed" by the model as part of the vowel environment, rather than being accommodated by the consonant-locus equations. This is potentially a very profound problem. There are a number of ways it might be partially addressed. One that has been lightly explored involves forcing the vowels /ʌ/ and /ɑ/ to be purely monophthongal on a priori grounds. This would encourage to some degree any average consonantal effects to stay with the consonants, since they cannot attach to at least some of the vowels. A second and more compelling approach would be to use additional contexts including #VC# which plays an important role in the models of Broad and Clermont (1987, 2002, 2010). Although Broad and Clermont have avoided #CV# and #V# environments because of possible centralization artifacts, in the present modeling framework, the assumption is that the formant movement at the end of such syllables is not artifactual, but a genuine reflection of vowel-inherent properties.

One characteristic of the model is not particularly attractive: namely, the stylized piecewise linear virtual vowel pattern, with 20 % initial fixed target, 60 % medial transitional part and 20 % final fixed target, is not a very satisfying representation, as

it leads inevitably to glitches in the formant tracks (discontinuities in the first derivative) at the vowel junction points. This approach is viewed as a first approximate step beyond a static target model. The selection of the 20 and 80 % time points is at least weakly motivated by previous exploration dual-target models (e.g., Hillenbrand and Nearey 1999; Assmann et al. 2013 Chap. 1). There is some indication in Morrison and Nearey (2007) that listeners may not be extremely sensitive to the precise shape of formant movement patterns in the more central vocalic regions, but we need to model them better in production in any event. Alternative vowel trajectory patterns clearly need to be investigated thoroughly.

One alternative that was explored involved a single linear trajectory to replace the three-piece pattern of Eq. 2 for each formant of each vowel by:

$$V_v(t) = T1_v + (1 - t)(T2_v - T2_v) \tag{7}$$

This linear vowel trajectory model led to a very similar overall fit to trajectories, but with only very slightly larger RMS errors overall. Although the simple linear vowel formant trajectories are technically feasible, they lead to a conceptual problem in using Sussman-style locus equations. Those equations assume for example, that initial formant measures of a CV syllable are linear functions of a target at a later position in the syllable, near the vowel nucleus. In the approach with simple linear vowel formant patterns, the only target values available were associated with the onset and offset of the vowels, at the terminals of the trajectories to be estimated.

More importantly, for straight-line vowel trajectories of (7) in place of the piecewise linear functions of (2), an analysis of component plots like those of Fig. 13 showed unrealistic vowel targets for some vowels of some speakers. Extremely high F2 values were estimated in some cases and in others F1 and F2 targets crossed (F2 < F1). Adjustments to consonant locus equation and exponential deflection parameters corrected this to yield plausible complete trajectories very similar to those using (2) to specify the vowel trajectory. The piecewise linear trajectories anchored at 20 and 80 %, though not altogether satisfying in their own right, avoid these problems.[13]

Other more general approaches to specifying the trajectory of the vowel component are on the drawing board. These include smooth vowel trajectory functions (e.g. scaled logistic or hyperbolic tangent functions) and changes in the timing of nucleus to offset transition functions. These appear to involve both conceptual and technical difficulties of their own, including ones of parameter identification or overparameterization. As yet no results are available with more elaborate models. Similarly, more flexibility might be contemplated in specifying shapes of consonant

[13] The sensitivity of the solution to slight changes in the model is a well-known pitfall in nonlinear modeling. Models involving sums of exponentials (which includes the models discussed above) are prone to a number of problems (Seber and Wild 1988). Even a superficial reading of the literature makes it clear that much care must be taken in constructing models and the design space on to which a model is fitted. A thorough scouring of that literature for similarly structured problems is warranted in future modeling efforts in this domain.

transitions. For example, varying aspiration durations might best be modeled by adjusting a virtual transition onset time in advance of first voiced onset frame as adopted here. All such complications must be handled carefully in non-linear modeling (see note 13) and would best be done in conjunction with supplementary data collection focused specifically on the questions at issue.

9.4 Comparison with Aspects of Other Models

The model sketched above uses an approach to specifying consonant onset parameters that is somewhat different than either the locus equations of Sussman et al. (1991) and from the (vowel) target locus scaling of Broad and Clermont (1987, 2010). The current model uses two-parameter locus equations involving one slope and one intercept per consonant per formant and per position (initial versus final). However, while identical in form, the locus equations used in this model are conceptually distinct in at least one important way from the locus equations of Sussman et al. (1991) or of Nearey and Shammass (1987). Those studies involved what might be called empirical locus equations, because measured (empirical) formant onsets (offsets for finals) are predicted from measured properties of individual tokens. The fitted equations, as Sussman notes, are abstractions. In the present model, the locus equations are even more abstract, and might reasonably be called virtual locus equations. The locus parameters operate on virtual (ideal, roughly "average") vowel targets vowel targets at the 20 and 80 % points to specify virtual formant onsets and offsets at the 0 and 100 % time points.

Broad and Clermont have focused since their 1987 paper explicitly on a rather different type of virtual locus model, similar to that of Delattre et al. (1955), which relates to an apparent common origin "just beyond the vowel boundary" (Broad and Clermont 2010, p. 180) though Broad and Clermont's method of estimation is different. Broad and Clermont (2010, pp. 15–19) discuss the relation between the two in some detail. These are vowel-independent. The model presented above does not rely on a fixed consonantal locus,[14] rather the locus equations predict onsets of formant transitions at time 0 % and offsets at 100 % from formant frequencies of the vowel targets (Thus the formant patterns in Figs. 13 and 14 extrapolate out to 0 and 100 % of the voiced formant trajectory duration of the vowel for initial and final consonants respectively, even though only 10 to 90 % are used in the fitting process).

Broad and Clermont's final model IVb, unlike their earlier models, was based on absolute rather than relative durations. As they note, an attractive feature of this approach is that it then bears strong relations to models of undershoot related to the

[14] However, a locus can usually be calculated following Sussman et al. (1991) and solving the locus equation $y = b1\ x + b0$ for $y = x$, namely at $b0/(b1-1)$, is possible (except where $b = 1$). Here b1 is the slope and b0 is the intercept and (for initial C) y is the formant onset value and x is a vowel target. This form reveals a technical difficulty with the explicit locus model, that is not present in the slope-intercept form, as the locus is undefined if the slope equals 1.0.

work of Lindblom (1963). They also note that there is at best ambiguous support in their own data for a real-time model over their relative-time model IVa. It seems likely that a wider variety of prosodic contexts, inducing within-vowel duration differences will be required to differentiate between relative and normalized time models.

10 Summary and Discussion

Although many refinements of full trajectory modeling remain to be explored, these initial results constitute an encouraging proof of concept. They show that using an explicit model incorporating effects of consonantal contexts, consistent improvements of fit of trajectories are obtained when the model includes an explicitly dual-target representation of many nominal monophthongs. While the consonantal trajectory modeling may be far from perfect, it does follow the general lines of the only explicitly decompositional account of formant trajectories yet proposed, that of Broad and Clermont (1987, 2002, 2010). Until a clear alternative[15] is available, it seems well worth pursuing variations of this general line of inquiry.

Although the situation is not nearly as straightforward as anticipated from some of our earlier work (especially Nearey and Assmann 1986; Andruski and Nearey 1992), it does appear that many of the findings from that work have a more general validity in multiple contexts and in several dialects. It is also clear, as Strange and colleagues have long rightly exhorted, that consonantal effects are substantial and may, perhaps especially in rapid or reduced speech, dominate vowel trajectory patterns. The general trajectory modeling framework pioneered by Broad and Clermont—and extended here to allow dual targets for vowels—offers a promising approach to a detailed study of the many issues involved in a systematic, quantitative way.

Acknowledgments I would like to thank Peter Assmann, Michael Kiefte, Geoffrey Stewart Morrison and Santiago Barreda for comments and suggestions. Special thanks go to David Broad and Frantz Clermont for very insightful comments on an earlier draft of this paper. Faults remaining are my own. Portions of this work were presented as paper 4ASCa2 'Spectral change in the Front Vowels of North American English' to the 157th meeting of the Acoustical Society of America, Portland Or 21 May 2009 and as poster 5aSc18 'A new non linear regression model for formant trajectories in English monosyllables incorporating dual targets for vowels' presented April 23 to the 159th Meeting of the Acoustical Society of America and NOISE–CON 2010, Baltimore, Maryland, 19–23 April 2010. Work supported in part by Social Sciences and Humanities Research Council of Canada Grant number 3010325 to T.M. Nearey. Thanks are due to the following for sharing summary data: Ron Thomson for Edmonton AB English; Peter Assmann for Dallas TX English; Jim Hillenbrand for N. Michigan English. Thanks also to Geoffrey Stewart Morrison for permission to use Fig. 4 adapted from his PhD thesis.

[15] Broad and Clermont (1987, 2002, 2010) do provide alternative, less parametric, approaches to the shapes of consonantal trajectories that could be investigated in the context of VISC. However, the limited degrees of freedom of the exponential models are attractive in evaluating VISC statistically and might be used in such a role even if more accurate, but more complex, accounts of consonantal trajectories are established; provided, of course, the exponential approximation does not lead to obvious artifacts that impact the interpretation of vowel targets.

Appendix A: Statistical Tables for Formant Movement Patterns

The following tables form the basis for Figs. 5, 6, 7, 8, 9, and 10. The formant measures used were those taken at 20 and 70 % of vocoid duration in all cases. Measures were log-transformed before processing and the *t*-tests and significance levels are based on the log values. The mean values in the 20 and 70 % columns are geometric means of the formants across speakers; that is, means were first calculated in the log-formant space and converted by exponentiation back to Hz. The column labeled ΔF is the difference in the geometric means. The *t*-statistics are based on the change in log space, that is, $ln(F_{70\%}) - ln(F_{20\%})$ within individual speakers (Table A.1).

Table A.1 Formant movement for Alberta /bV/ vowels of Fig. 5

Vowel	Fk	20 %	70 %	Δ*F*	*t*	*df*	*p*	Notes
I. Mid tense with expected iota- and upsilon-VISC								
e	F1	524.6	439.8	−84.7	−9.565	19	0.00000	*
e	F2	2181.6	2392.0	210.4	11.313	19	0.00000	*
o	F1	565.4	419.9	−145.5	−7.226	19	0.00000	*
o	F2	1151.0	837.7	−313.3	−10.217	19	0.00000	*
II. High tense with possible iota- and upsilon-VISC								
i	F1	335.7	330.4	−5.3	−0.969	19	0.34490	.r
i	F2	2510.25	2606.80	96.6	6.682	19	0.00000	*
u	F1	384.20	338.80	−45.5	−5.340	19	0.00004	*
u	F2	1352.63	1069.6	−283.0	−12.562	19	0.00000	*
III. Lax with suspected alpha-VISC								
ɪ	F1	493.1	578.5	85.4	6.612	19	0.00000	*
ɪ	F2	1989.5	1940.3	−49.3	−2.229	19	0.03804	*
ɛ	F1	616.5	705.7	89.3	8.834	19	0.00000	*
ɛ	F2	1893.7	1885.9	−7.7	−0.412	19	0.68512	.
æ	F1	804.4	875.5	71.1	3.768	19	0.00130	*
æ	F2	1694.9	1617.7	−77.2	−4.618	19	0.00019	*R
ʊ	F1	504.3	559.8	55.5	5.772	19	0.00001	*R
ʊ	F2	1228.6	1295.0	66.5	6.551	19	0.00000	*
IV. No expected movement								
ʌ	F1	688.7	737.9	49.1	7.979	18	0.00000	!
ʌ	F2	1353.1	1381.5	28.4	2.909	18	0.00936	!
ɑ	F1	768.1	803.9	35.8	3.018	19	0.00708	!
ɑ	F2	1223.0	1274.6	51.7	3.939	19	0.00088	!

Notes (*see text*)
* Significant movement in expected direction
. Non-significant movement in expected direction
! Significant movement when none was expected
r Non-significant retrograde movement from 70 to 90 %
R significant retrograde movement

Table A.2 Formant movement for Alberta /pV/ vowels of Fig. 6

Vowel	Fk	20 %	70 %	ΔF	t	df	p	Notes
I. Mid tense with expected iota- and upsilon-VISC								
e	F1	525.9	433.5	−92.4	−8.839	19	0.00000	*
e	F2	2242.41	2406.7	164.3	13.229	19	0.00000	*
o	F1	556.2	425.0	−131.2	−9.169	19	0.00000	*
o	F2	1103.14	885.8	−217.3	−11.261	19	0.00000	*
II. High tense with possible iota- and upsilon-VISC								
i	F1	338.0	324.6	−13.4	−1.846	19	0.08050	.r
i	F2	2567.17	2611.3	44.1	3.118	19	0.00567	*R
u	F1	381.9	323.9	−58.0	−5.592	19	0.00002	*
u	F2	1296.31	1051.1	−245.2	−10.682	19	0.00000	*
III. Lax with suspected alpha-VISC								
ɪ	F1	519.3	565.6	46.3	4.620	19	0.00019	*
ɪ	F2	2113.56	2023.8	−89.8	−4.848	19	0.00011	*
ɛ	F1	696.8	733.3	36.5	4.632	19	0.00018	*
ɛ	F2	1863.19	1835.0	−28.2	−2.294	19	0.03338	*
æ	F1	889.8	908.4	18.6	2.133	19	0.04618	*
æ	F2	1714.50	1670.8	−43.7	−2.653	19	0.01569	*
ʊ	F1	527.9	585.7	57.8	3.399	19	0.00301	*
ʊ	F2	1254.01	1317.3	63.3	3.624	19	0.00181	*
IV. No expected movement								
ʌ	F1	733.2	747.1	13.9	2.052	19	0.05426	=
ʌ	F2	1406.83	1434.0	27.2	2.590	19	0.01795	!
ɑ	F1	803.6	831.8	28.2	2.340	19	0.03036	!
ɑ	F2	1238.81	1271.7	32.9	3.301	19	0.00376	!

Notes (*see text*)
* Significant movement in expected direction
. Non-significant movement in expected direction
= Non-significant movement when none is expected
! Significant movement when none was expected
r Non-significant retrograde movement from 70 to 90 %
R significant retrograde movement

An additional measurement was taken at 90 % of vowel duration. Although this measurement is not reported in the tables, its value was used to check for possible 'retrograde' or 'switchback' motion of formant frequencies. If the movement patterns from 20 to 70 to 90 % are all in the same direction, then the movement is not retrograde. But if a formant moves up from 20 to 70 % then down again from 70 to 90 %, the pattern is classed as retrograde. (This pattern occurs, for example, in a number of cases for the F1 of /ɪ/ in the data below). Note that this is a sufficient, but not necessary, condition for the detection of VISC. Thus, for the syllable /hed/, VISC in F1 is not controversial, even though we would expect a non-retrograde pattern where F1 moves lower from nucleus to offglide, then lower still to the low F1 offset associated with a voiced stop. The vowels are organized into groups I-IV of Sect. 5.1. Patterns with the designation *R in the "Notes" columns of Tables A.2,

Table A.3 Formant movement for Alberta /bVt/ vowels of Fig. 7

Vowel	Fk	20 %	70 %	ΔF	t	df	p	Notes
I. Mid tense with expected iota- and upsilon-VISC								
e	F1	461.5	367.6	−93.8	−13.492	32	0.00000	*
e	F2	2395.33	2644.2	248.9	20.033	32	0.00000	*R
o	F1	554.0	416.5	−137.6	−11.824	32	0.00000	*
o	F2	1213.95	1012.9	−201.1	−14.076	32	0.00000	*R
II. High tense with possible iota- and upsilon-VISC								
i	F1	325.9	319.4	−6.5	−1.303	32	0.20182	.
i	F2	2616.86	2702.5	85.7	8.035	32	0.00000	*R
u	F1	384.3	350.6	−33.7	−5.422	32	0.00001	*
u	F2	1405.47	1321.9	−83.5	−2.939	32	0.00607	*R
III. Lax expected alpha VISC								
ɪ	F1	474.4	545.7	71.3	9.627	32	0.00000	*R
ɪ	F2	2153.18	2121.2	−32.0	−2.358	32	0.02464	*
ɛ	F1	672.0	781.9	110.0	11.003	32	0.00000	*R
ɛ	F2	1968.44	1928.0	−40.5	−2.756	32	0.00957	*
æ	F1	839.6	957.0	117.4	12.894	31	0.00000	*R
æ	F2	1825.71	1749.9	−75.8	−4.750	31	0.00004	*R
ʊ	F1	517.0	587.4	70.4	7.387	31	0.00000	*R
ʊ	F2	1193.44	1237.1	43.7	4.090	31	0.00028	*
IV. No expected movement								
ʌ	F1	738.8	801.6	62.8	6.377	31	0.00000	!R
ʌ	F2	1445.09	1662.0	216.9	11.041	31	0.00000	!
ɑ	F1	737.1	845.5	108.4	10.980	32	0.00000	!R
ɑ	F2	1177.22	1319.6	142.3	8.597	32	0.00000	!

Notes (*see text*)
* Significant movement in expected direction
. Non-significant movement in expected direction
! Significant movement when none was expected
R significant retrograde movement

A.3, A.4, A.5 might be interpreted as showing particularly strong evidence for VISC patterns in the expected direction (see the classification of Sect. 5.1), as they show significant movement from 20 to 70 % in the expected direction and also a significant movement in the opposite direction from 70 to 90 %.

There are two cases of retrograde movement in Table A.1 for Alberta /bV/. F2 of /æ/ decreases from 20 to 70 %, but then increases a little again (not shown) at 90 %. F1 of /ʊ/ shows a rising–falling pattern. These cases are not very compelling because we have no good reason to expect a final silence to induce a particular context effect, beyond some speculation in the literature of a return toward a neutral position at the end of a word final vowel.

In Table A.2, for Alberta /pV/ contexts, there is only one case of significant retrograde motion, involving F2 of /i/, where there is also non-significant retrograde motion for F1. Again this is not very compelling, because there is no

Table A.4 Formant movement for Alberta /pVt/ vowels of Fig. 8

Vowel	Fk	20 %	70 %	ΔF	t	df	p	Notes
I. Mid tense with expected iota- and upsilon-VISC								
e	F1	455.3	360.1	−95.2	−9.431	32	0.00000	*
e	F2	2528.30	2666.1	137.8	9.752	32	0.00000	*R
o	F1	540.3	416.1	−124.2	−10.464	32	0.00000	*
o	F2	1146.79	999.9	−146.9	−9.369	32	0.00000	*R
II. High tense with possible iota- and upsilon-VISC								
i	F1	319.2	312.8	−6.4	−1.258	32	0.21759	.
i	F2	2711.62	2732.4	20.8	1.305	32	0.20131	.r
u	F1	367.7	338.2	−29.5	−3.234	32	0.00284	*
u	F2	1371.53	1254.7	−116.8	−5.092	32	0.00002	*R
III. Lax expected alpha VISC								
ɪ	F1	533.7	553.2	19.5	1.530	30	0.13649	.r
ɪ	F2	2193.52	2100.1	−93.4	−6.801	30	0.00000	*
ɛ	F1	774.1	769.7	−4.4	−0.419	30	0.67828	?
ɛ	F2	1982.32	1933.2	−49.1	−4.458	30	0.00011	*
æ	F1	931.9	937.5	5.7	0.575	31	0.56919	.r
æ	F2	1774.79	1726.4	−48.4	−3.482	31	0.00151	*R
ʊ	F1	535.2	577.0	41.8	3.207	31	0.00311	*R
ʊ	F2	1314.62	1582.5	267.8	12.689	31	0.00000	*
IV. No expected movement								
ʌ	F1	842.1	799.5	−42.6	−4.323	29	0.00017	!
ʌ	F2	1572.85	1722.7	149.9	10.914	29	0.00000	!
ɑ	F1	808.3	856.7	48.4	5.291	31	0.00001	!R
ɑ	F2	1191.29	1366.8	175.5	8.498	31	0.00000	!

Notes (*see text*)
* Significant movement in expected direction
. Non-significant movement in expected direction
? Non-significant movement in the opposite of expected direction
! Significant movement when none was expected
r Non-significant retrograde movement from 70 to 90 %
R significant retrograde movement

unambiguous contextual motivation for F1 and F2 to change direction so late in the syllable.

No further comment will be added in the case of significant movement between 20 and 70 %, as these have been discussed in the text. However a few comments are added here regarding retrograde movement of formants, specifically for syllables with significant VISC movements from 20 to 70 % in the expected direction but then show a further reliable turning in the opposite direction.

For the Alberta /bVt/ in Table A.3 there are 9 instances of significant retrograde motion in the directions expected based on predictions of Sect. 5.1. For the F2s, these are largely consistent with movements from nucleus (near 20 %) through offglide (near 70 %) to a final alveolar F2 locus (1800–1900 Hz). The F1 patterns of the group III lax vowels are consistent with a pattern of moving from a lower

Table A.5 Michigan /hVd/

Vowel	Fk	20 %	70 %	ΔF	t	df	p	Notes
I. Mid tense with expected iota- and upsilon-VISC								
e	F1	522.8	448.3	−74.5	−23.031	92	0.00000	*
e	F2	2280.35	2419.1	138.8	17.191	92	0.00000	*
o	F1	556.8	470.8	−86.1	−26.596	92	0.00000	*
o	F2	1016.58	919.3	−97.3	−15.161	92	0.00000	*R
II. High tense with possible iota- and upsilon-VISC								
i	F1	401.6	394.4	−7.2	−3.397	92	0.00101	*R
i	F2	2509.22	2525.0	15.8	3.413	92	0.00096	*R
u	F1	432.6	416.4	−16.2	−6.274	92	0.00000	*R
u	F2	1066.92	1024.0	−42.9	−4.698	92	0.00001	*R
III. Lax expected alpha VISC								
ɪ	F1	474.0	519.4	45.4	13.795	92	0.00000	*R
ɪ	F2	2185.11	2025.6	−159.6	−17.910	92	0.00000	*
ɛ	F1	657.7	648.4	−9.3	−2.215	92	0.02922	!!
ɛ	F2	1945.01	1884.3	−60.7	−5.903	92	0.00000	*R
æ	F1	641.3	722.5	81.2	13.180	92	0.00000	*R
æ	F2	2106.13	1843.1	−263.1	−16.134	92	0.00000	*
ʊ	F1	505.4	546.6	41.2	13.095	92	0.00000	*R
ʊ	F2	1174.50	1471.0	296.5	22.406	92	0.00000	*
IV. No expected movement								
ʌ	F1	687.5	663.9	−23.6	−7.293	92	0.00000	!
ʌ	F2	1294.51	1523.1	228.6	24.464	92	0.00000	!
ɑ	F1	818.3	797.4	−21.0	−5.426	92	0.00000	!
ɑ	F2	1412.08	1507.8	95.8	11.718	92	0.00000	!
ɔ	F1	713.5	739.2	25.7	7.100	92	0.00000	!R
ɔ	F2	1122.45	1278.4	156.0	17.217	92	0.00000	!
ɝ	F1	534.9	498.2	−36.7	−12.821	92	0.00000	!
ɝ	F2	1453.63	1520.1	66.5	9.405	92	0.00000	!

Notes (*see text*)
* Significant movement in expected direction
! Significant movement when none was expected
!! Significant movement opposite of expected direction
R significant retrograde movement

nucleus value through a higher offglide then back toward a lower F1 locus for the final stop. While voiced stops have lower F1 offsets than voiced, there is on average some measurable lowering of F1 even with voiceless stops (Flege et al. 1992; Summers 1987).

For the Alberta /pVt/ contexts, there are far fewer significant retrograde movements in either F1 or F2. The significant retrograde F2 movements are largely consistent with again with a late movement toward an alveolar locus. For the group III vowels, only F1 of /ʊ/ shows reliable upward movement in F1 followed by downward movement from 20 to 90 %.

For the /hVd/ of both Michigan and Texas, again the significant retrograde F2 movements are for class I, II and III vowels are interpretable as expected nucleus

Table A.6 Texas /hVd/

Vowel	Fk	20 %	70 %	ΔF	t	df	p	Notes
I. Mid tense with expected iota- and upsilon-VISC								
e	F1	557.5	403.4	−154.2	−14.390	18	0.00000	*
e	F2	2171.97	2380.4	208.4	14.112	18	0.00000	*R
o	F1	571.3	422.6	−148.7	−11.683	19	0.00000	*
o	F2	1343.29	1223.2	−120.1	−7.235	19	0.00000	*R
II. High tense with possible iota- and upsilon-VISC								
i	F1	360.4	348.5	−11.9	−2.910	17	0.00974	*R
i	F2	2499.29	2524.3	25.1	2.144	17	0.04678	*R
u	F1	401.4	381.7	−19.7	−3.948	19	0.00086	*
u	F2	1520.14	1455.7	−64.4	−6.038	19	0.00001	*R
III. Lax expected alpha-VISC								
ɪ	F1	435.9	482.9	47.1	4.570	18	0.00024	*R
ɪ	F2	2164.11	1971.5	−192.6	−10.350	18	0.00000	*
ɛ	F1	587.1	582.7	−4.4	−0.537	18	0.59800	?
ɛ	F2	1935.88	1824.4	−111.5	−8.812	18	0.00000	*
æ	F1	756.6	727.5	−29.2	−2.541	18	0.02046	!!
æ	F2	1932.41	1834.8	−97.6	−7.724	18	0.00000	*
ʊ	F1	462.6	474.2	11.6	1.345	19	0.19460	.r
ʊ	F2	1459.92	1651.8	191.9	7.088	19	0.00000	*
IV. No expected movement								
ʌ	F1	659.0	591.2	−67.8	−8.847	18	0.00000	!
ʌ	F2	1568.88	1712.5	143.6	9.152	18	0.00000	!
ɑ	F1	770.1	743.7	−26.4	−3.393	19	0.00305	!
ɑ	F2	1240.28	1432.9	192.6	10.335	19	0.00000	!
ɔ	F1	714.9	709.7	−5.3	−0.729	18	0.47563	=
ɔ	F2	1141.09	1338.0	196.9	8.327	18	0.00000	!
ɝ	F1	550.0	465.8	−84.2	−8.767	19	0.00000	!
ɝ	F2	1462.45	1523.9	61.5	8.512	19	0.00000	!

Notes (*see text*)
* Significant movement in expected direction
. Non-significant movement in expected direction
? Non-significant movement in the opposite of expected direction
= Non-significant movement when none is expected
! Significant movement when none was expected
!! Significant movement opposite of expected direction
r Non-significant retrograde movement from 70 to 90 %
R significant retrograde movement

to offglide movement followed by return to intermediate alveolar locus values. For F1, Michigan shows a pattern consistent with alpha-VISC followed by movement to a lower F1 locus for the final /d/ only for the lax high vowels /ɪ/ and /ʊ/, and for the impressionistically strongly diphthongal "tense" /æ/. Texas shows significant retrograde F1 movement only for /ɪ/ (Table A.6).

References

Andruski, J., Nearey, T.M.: On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. J. Acoust. Soc. Am. **91**(1), 390–410 (1992)

Assmann, P.F., Katz, W.F.: J. Acoust. Soc. Am. **108**, 1856–1866 (2000)

Assmann, P.F., Nearey, T.M., Hogan, J.: Vowel identification: orthographic, perceptual and acoustic aspects. J. Acoust. Soc. Am. **71**, 975–989 (1982)

Assmann, P.F., Nearey, T.M., Bharadwaj, S.V.: Developmental patterns in children's speech: patterns of spectral change in vowels. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (ch. 9). Springer, Heidelberg (2013)

Barry, W.J., Trouvain, J.: Do we need a symbol for a central open vowel? The discussion so far and a reply to Daniel Recasens and Martin Ball. J. Int. Phonetic Assoc. **39**(03), 365–366 (2009). doi:10.1017/S0025100309990235

Boberg, C.: Regional phonetic differentiation in standard Canadian English. J. Engl. Linguist. **36**(2), 129 (2008)

Boberg, C.: The English Language in Canada: Status, History and Comparative Analysis. Cambridge University Press, Cambridge (2010)

Broad, D., Clermont, F.: A methodology for modeling vowel formant contours in CVC context. J. Acoust. Soc. Am. **81**, 155–165 (1987). doi:10.1177/0075424208316648

Broad, D.J., Clermont, F.: Linear scaling of vowel-formant ensembles (VFEs) in consonantal contexts. Speech Commun. **37**(3–4), 175–195 (2002). doi:10.1016/S0167-6393(01)00010-3

Broad, D.J., Clermont, F.: Target-locus scaling methods for modeling families of formant transitions. J. Phonetics **38**(3), 337–359 (2010). doi:10.1016/j.wocn.2010.02.004

Broad, D.J., Fertig, R.H.: Formant-freqency trajectories in selected CVC-syllable nuclei. J. Acoust. Soc. Am. **47**(6), 1572–1582 (1970)

Clermont, F., Millar, B.: Multi-speaker validation of coarticulation models of syllabic nuclei. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86, vol. 11, pp. 2671–2674). Presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86 (1986). doi: 10.1109/ICASSP.1986.1168577

Delattre, P.C., Liberman, A.M., Cooper, F.S.: Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. **27**(4), 769–773 (1955)

Di Benedetto, M.G.: Frequency and time variations of the first formant: properties relevant to the perception of vowel height. J. Acoust. Soc. Am. **86**(1), 67–77 (1989). doi:10.1121/1.398221

Flege, J.E., Munro, M.J., Skelton, L.: Production of the word-final English /t/–/d/ contrast by native speakers of English, Mandarin, and Spanish. J. Acoust. Soc. Am. **92**(1), 128–143 (1992). doi:10.1121/1.2029176

Gumpertz, M., Pantula, S.G.: A simple approach to inference in random coefficient models. Am. Sta. **43**(4), 203–210 (1989)

Hillenbrand, J.M., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**(5, pt. 1), 3099–3111 (1995)

Hillenbrand, J.M., Nearey, T. M.: Identification of resynthesized/hVd/syllables: Effects of formant contour. J. Acoust. Soc. Am. **105**, 3509–3523 (1999)

Hillenbrand, J.M., Clark, M.J., Nearey, T.M.: Effects of consonant environment on vowel formant patterns. J. Acoust. Soc. Am. **109**(2), 748–763 (2001). doi:10.1121/1.1337959

Hockett, C.: A Course in Linguistics. MacMillan, New York (1958)

Jenkins, J.J., Strange, W., Edman, T.R.: Identification of vowels in 'vowelless' syllables. Percept. Psychophys. **34**, 441–450 (1983). doi:10.3758/BF03203059

Kingston, J., Diehl, R.: Phonetic knowledge. Language **70**, 419–454 (1994). doi:10.2307/416481

Klatt, D.: Software for a cascade/parallel synthesizer. J. Acoust. Soc. Am. **67**, 971–995 (1980)

Labov, W., Ash, S., Boberg, C.: The Atlas of North American English: Phonetics, Phonology, and Sound Change: A Multimedia Reference Tool. Walter De Gruyter, The Hague (2006). doi:10.1215/00031283-2007-014

Ladefoged, P.: American English. In: Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet, pp. 41–44. Cambridge University Press, Cambridge (1999)

Lehiste, I., Peterson, G.: Transitions, glides and diphthongs. J. Acoust. Soc. Am. **33**(3), 268–277 (1961). doi:10.1121/1.1908638

Lindblom, B.: Spectrographic study of vowel reduction. J. Acoust. Soc. Am. **35**, 1773–1781 (1963)

Lorch, R.F., Myers, J.L.: Regression analyses of repeated measures data in cognitive research. J. Exp. Psychol. Learn. Mem. Cogn. **16**, 149–157 (1990). doi:10.1037/0278-7393.16.1.149

Morrison, G.S.: Perception of synthetic vowels by monolingual Canadian-English, Mexican-Spanish, and Peninsular-Spanish listeners. Can. Acoust. **36**(4), 17–23 (2008)

Morrison, G.S., Nearey, T.M.: Testing theories of vowel inherent spectral change. J. Acoustic. Soc. Am. **122**, EL15–EL22 (2007). doi: 10.1121/1.2739111

Morrison, G.S.: L1 and L2 production and perception of English and Spanish vowels: a statistical modelling approach. PhD dissertation, University of Alberta (2006)

Nearey, T.M.: Evidence for the perceptual relevance of vowel-inherent spectral change for front vowels in Canadian English. Proc. XIII Int. Cong. of Phonetic Sciences. Stockholm **2**, 678–681 (1995)

Nearey, T.M., Assmann, P.F.: Modeling the role of inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986)

Nearey, T.M.: A new non-linear regression model for formant trajectories in English monosyllables incorporating dual targets for vowels. J. Acoustic. Soc. Am. **127**, 2020 (abstract) (2010). doi: 10.1121/1.3385273

Nearey, T.M., Shammass, S.: Formant transitions as partly distinctive invariant properties in the identification of voiced stops. Can. Acoust. **15**, 17–24 (1987)

Neel, A.T.: Formant detail needed for vowel identification. Acoust. Res. Lett. Online **5**(4), 125–131 (2004). doi:10.1121/1.1764452

Nielsen, H.B.: Immoptibox: a Matlab toolbox for optimization and data fitting. Retrieved 13 Aug 2009, from http://www2.imm.dtu.dk/~hbn/immoptibox/ (2009)

O'Grady, W., Dobrovolsky, M., Katamba, F.: Contemporary Linguistics: An Introduction. Addison-Wesley, Boston (1997)

Rogers, H.: The Sounds of Language: An Introduction to Phonetics. Longman, Harlow (2000)

Seber, G.A.F., Wild, C.J.: Nonlinear Regression. Wiley, New York (1988)

Stevens, K.N., House, A.S.: Perturbation of vowel articulations by consonantal context: an acoustical study. J. Speech Hear. Res. **6**, 111–128 (1963)

Strange, W.: Dynamic specification of coarticulated vowels spoken in sentence context. J. Acoust. Soc. Am. **85**(5), 2135–2153 (1989). doi:10.1121/1.397863

Strange, W., Jenkins, JJ: Dynamic specification of coarticulated vowels: research chronology, theory, and hypotheses. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (ch. 5). Springer, Heidelberg (2013)

Summers, W.V.: Effects of stress and final-consonant voicing on vowel production. J. Acoust. Soc. Am. **82**, 847–863 (1987). doi:10.1121/1.395284

Sussman, H.M., Bessell, N., Dalston, E., Majors, T.: An investigation of stop place of articulation as a function of syllable position: a locus equation perspective. J. Acoust. Soc. Am. **101** (5 Pt 1), 2826–2838 (1997). doi:10.1121/1.418567

Sussman, H.M., McCaffrey, H.A., Matthews, S.A.: An investigation of locus equations as a source of relational invariance for stop place categorization. J. Acoust. Soc. Am. **90**, 1256–1268 (1991)

Thomas, E.R.: Secrets revealed by Southern vowel shifting. Am. Speech **78**(2), 150–170 (2003). doi:10.1215/00031283-78-2-150

Thomson, R.I.: Modeling L1/L2 interactions in the perception and production of English vowels by Mandarin L1 speakers: a training study. PhD dissertation, University of Alberta (2007)

Trager, G.L., Smith, H.L.: An Outline of English Structure. American Council of Learned Societies, New York (1957)
Verbrugge, R., Rakerd, B.: Evidence of talker-independent information for vowels. Lang. Speech **29**(1), 39–57 (1986). doi:10.1177/002383098602900105

Dynamic Specification of Coarticulated Vowels

Research Chronology, Theory, and Hypotheses

Winifred Strange and James J. Jenkins

Abstract This chapter summarizes research conducted over a 35 year period on the dynamic specification of vowels. A series of experiments comparing vowels in consonant context with vowels produced in isolation failed to support talker normalization theories that predicted higher accuracy through prior exposure to a talker's "point vowels." Instead, these studies showed that vowels in consonant context were more accurately identified than isolated vowels, supporting a dynamic specification of vowels theory over static target theories, leading to the proposal that important information is contained in the formant transitions. Consonant–vowel coarticulation is not a source of "noise", rather it gives rise to an acoustic array in which the consonants and vowels are cospecified in the time-varying spectral configuration which we call dynamic specification. Subsequent experiments showed high identification accuracy for "silent center vowels" in which the central portion of the CVC syllable was removed by gating. Identification accuracy was not disrupted when the onset and offset portions were produced by different speakers. Vowel identification improved with increasing duration of the onsets or offset portions. Onsets were identified more accurately than offsets but neither was as well identified as the silent center syllables. Collectively these and other experiments summarized herein support the view that the most important source of information for speaker-invariant vowel identity is carried in dynamic specification of vowel onset and offset spectral patterns, with vowel duration also playing a role. Subsequent experiments with North German vowels, which do not exhibit the degree of vowel diphthongization reported in

W. Strange (✉) · J. J. Jenkins
PhD Program in Speech, Language, Hearing Sciences Graduate School and University Center City University of New York, New York, USA
e-mail: strangepin@aol.com

J. J. Jenkins
e-mail: Jcube01@gmail.com

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_5,

American English dialects, showed that listeners rely on dynamic spectro-temporal information specified by syllable onsets and offsets, in addition to cues provided by inherent vowel duration. Cross-language comparisons are presented from the perspective of adaptive dispersion theory. These comparisons support the view that dynamic properties are perceptually more important in differentiating vowels in languages with large vowel inventories.

Abbreviations

CVC	Consonant vowel consonant
CV	Consonant vowel
VC	Vowel consonant
F1	First formant
F2	Second formant
F3	Third formant
AE	American English
NG	North German
PF	Parisian French
VISC	Vowel inherent spectral change

1 Introduction

Because the research reported here has taken place over a 35 year period, it is important to recognize that positions and issues have changed over time and the motivation for particular studies has changed accordingly. When we began our studies in the 1970s, it was widely accepted that the "steady-state target" portions of vowels provided the primary (and in most models, the only) information for their perception (see Hillenbrand 2013, Chap. 2). The primary theoretical problem in this view was that static acoustic targets (positions in an F1/F2 acoustic vowel space) were not invariant across speakers of different ages and genders; this was termed the *speaker normalization problem*. However, when we started investigating American English vowels coarticulated with consonants in different consonant–vowel–consonant (CVC) syllables, we also found considerable variability of static acoustic targets of vowels produced in different consonantal contexts by the same speaker. This is often referred to as the *target undershoot problem* (Lindblom 1963, 1983) because the midsyllable formant values of coarticulated vowels often do not reach the static steady-state values of vowels produced in isolation. In general, target undershoot leads to more overlap in the midsyllable formant values of different vowels in a language such as English which has a large vowel inventory. It is at this point that we, along with Donald Shankweiler and several of our students began our research program in the 1970s.

The first phase of our research resulted in changing the view that North American English vowels are completely specified by static target formant values.

Almost no investigator holds that position today (c.f. many chapters in this volume). Later, as dynamic spectral factors received more attention from many researchers, there was an ongoing debate concerning which aspects of spectral change were most important e.g., "nucleus plus offglide" versus "onset and offset transitions" (c.f. Andruski and Nearey 1992; Nearey 2013, Chap. 4; Sect. 3). Now it seems to be widely agreed that the dynamic information for vowel perception is highly redundant and that rich information of several kinds is available throughout the course of the syllable. Finally, at the present time, research is being extended to questions of dialect, to diachronic change, to languages other than English, and to perception of vowels by young native learners and by non-native and second-language learners.

2 History of the Specification of Vowels

In the history of research on the acoustics of spoken vowels, there was a strong predisposition to concentrate on the steady-state resonances of the vocal tract. Given the state of instrumentation prior to the 1940s, there was little option but to study sustained, relatively steady-state utterances. Helmholtz (1885/1954), studying vowels with his tuning forks and resonators, concluded that front vowels were differentiated by two resonances but back vowels by a single resonance tone. Bell (1911) agreed in general but argued for two resonances as the critical features for all vowels, a front cavity resonance and a back cavity resonance. With the invention of the speech spectrograph in the mid-1940s (Potter et al. 1947), more detailed analysis became possible. Taking advantage of the new instrument, Joos (1948) published a classic monograph on the spectral analysis of speech sounds suggesting that two or three resonances were sufficient to characterize vowels. Finally, with the invention of speech synthesis techniques, especially the Pattern Playback at Haskins Laboratories (Cooper 1950), the demonstration seemed to be complete: two steady-state resonances were sufficient to produce identifiable vowels in English and the consonants could be synthesized as needed by adding linear formant transitions at the ends of the vowel resonances (see Fig. 1).

Thus, as Hillenbrand (2013, Chap. 2), points out, it seemed entirely appropriate for Peterson and Barney (1952) to make spectrographic measurements on one slice near the center of the syllable of /hVd/ words spoken by 76 speakers of different ages, genders and dialects of English, and to graphically present their data as in Hillenbrand's Fig. 2.1. It is testimony to the power of graphic representations that it was the remarkable overlap in formant values of different vowels that remained with the readers, rather than the high identification accuracy of native listeners when they were presented the /hVd/ words, with speakers randomly varying from token to token. The caution of the authors in their discussion that other aspects of the acoustics needed to be measured was largely neglected. As a result of this understandable focus on steady-state formant targets, many theorists and investigators regarded the vowel identification problem almost completely as a speaker normalization problem.

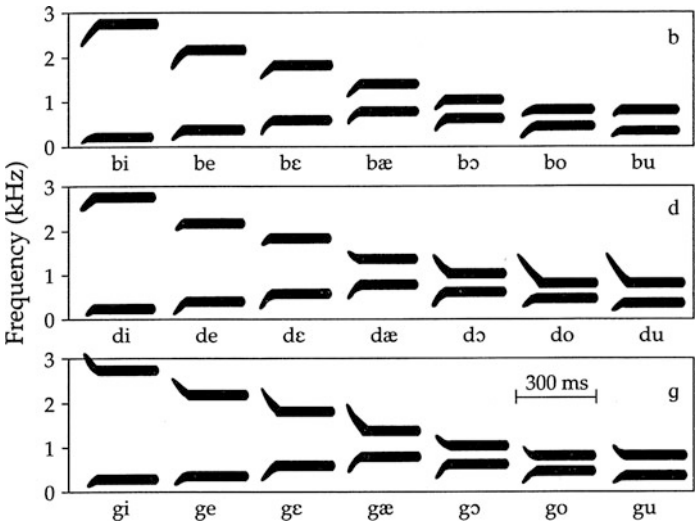


Fig. 1 Spectral patterns for formants on the Pattern Playback to produce the English vowels in three consonant contexts. From Johnson (1997 p. 135) (originally published in Haskins Laboratory reports)

The notion that vowels were perceived relative to some other aspect of the speaker's speech production patterns was given support by Ladefoged and Broadbent's (1957) demonstration that the identification of an ambiguous vowel in a synthesized syllable could be changed (e.g., from /ɪ/ to /ɛ/) by shifting the synthesis parameters of the sentence in which the syllable was embedded. The implication was that the listener was adjusting to some aspect of the speaker's vocal tract. One suggestion was that the listener used the incoming speech to build a model of the speaker's "vowel space" and interpreted the speaker's vowels in that context. Lieberman et al. (1972) suggested, as Joos (1948) had earlier, that the "point vowels" /i/, /ɑ/, /u / could serve this purpose and provide a framework for identifying the remaining vowels. This hypothesis seemed reasonable but had not been empirically tested at the time our research began.

3 Vowels in Consonantal Context Versus Isolated Vowels

3.1 The Original Findings

In the speech perception laboratory at the University of Minnesota, Shankweiler, Strange, Verbrugge, and Edman set out to test the point-vowel hypothesis about vowel normalization and to examine other possibilities. Verbrugge et al. (1976) reported three experiments, each with multiple conditions utilizing 15 to 30 native speakers of American English (AE) (men, women, and children), 9 to 15 AE

vowels, /hVd/ and /pVp/ syllables, citation form syllables (i.e. read in a list), as well as stressed syllables excised from sentences, destressed syllables excised from sentences, and those same destressed syllables imbedded in their original sentences. The authors reached the conclusion that talker-dependent variation was not a major perceptual problem within a dialect group. Contrary to the point-vowel hypothesis, exposure to a talker's point vowels before each test syllable did not decrease, and sometimes increased, errors of identification. Surprisingly, listeners who heard only a single talker made only 7 % fewer errors than those listening to many talkers. Finally, the writers conjectured that a naturally-produced carrier sentence aided vowel identification more by establishing the proper tempo for the syllable than by specifying a spectral vowel space.

Strange et al. (1976) directly challenged the generality of prior experiments that utilized only isolated vowel tokens. In two studies with several experimental conditions in each, they found that vowels produced in both single consonant or randomly varying consonant–vowel-consonant (CVC) contexts were more accurately identified than isolated vowels produced by the same speakers under the same test conditions. Overall, vowels in consonant context with speakers varying randomly were correctly identified about 80 % of the time whereas isolated vowels were identified less than 60 % of the time. Surprisingly, this difference between the two contextual conditions was observed in tests where the listeners heard only one speaker at a time. (90 % correct in a fixed consonant context, 77 % correct in varying consonantal context and 69 % as isolated vowels). The writers concluded by suggesting that important information for vowel identity is contained in formant transitions, as Lindblom and Studdert-Kennedy (1967) had suggested earlier. From the perceptual standpoint they argued "the definition of a vowel ought to include a specification of how the relevant acoustic parameters change over time" (Strange et al. 1976, p. 221). The significance of these and other experiments and further discussion of questions of "normalization" versus dynamic specification of vowels can be found in Shankweiler et al. (1977).

3.2 Failures to Replicate the Original Findings

As might be expected, these results which seemed to fly in the face of "accepted wisdom" aroused considerable controversy. At the meeting of the Acoustical Society of America in 1979 an entire symposium was devoted to "failures to replicate" Strange et al. (1976). However, most of the papers consisted of demonstrations of the greater accuracy of identification of synthesized steady-state vowels compared to the identification of synthesized CVC syllables with artificial formant transitions tacked on the ends of steady states. Such studies, of course, conflate the experimental question with the adequacy of the synthesis. The symposium paper of most relevance was subsequently published by Macchi (1980). Her study replicated Strange et al. (1976) using naturally produced syllables and young, phonetically naïve listeners, listening over high quality headphones rather

than loud speaker. The speakers and listeners were carefully matched in dialect and the listeners were thoroughly trained in the use of the response forms. In Macchi's study no difference was found in the accuracy of identification of isolated vowels as compared to vowels in CVC syllables. However, the accuracy rates were so high (about 95 %) that perception of both kinds of tokens was essentially at ceiling. The study, however, replicated the small effect on perception for multiple speakers versus a single speaker (about 6 %), suggesting that speaker variability made only a small, though consistent contribution to problems in vowel identification.

Diehl et al. (1981) also conducted a partial replication of Strange et al. (1976). The researchers failed to find the consonant advantage when they employed synthetic materials in their first two experiments. However, they rejected these findings (and similar findings of other writers) because of high error rates in all conditions (between 20 and 30 %) attributed to the unnaturalness of the synthetic stimuli. When they replicated the study with naturally-produced stimuli, they found the expected advantage of CVC syllables over isolated vowels, with about twice as many errors on the isolated vowels. While this validated the finding of Strange et al. the writers suggested that the difference between conditions could be attributed to differences in the compatibility of response forms. (It should be noted, however, that in their experiments with natural stimuli, the isolated vowels were better identified on the CVC answer sheets than on the supposedly compatible isolated vowel answer sheets.) They also suggested that some of the difference might lie in the memory requirements of the tasks rather than differences in perception *per se*.

4 Methodological Studies of Vowel Identification

Macchi's and Diehl's studies as well as Assmann et al. (1982) called attention to the importance of appropriate response forms and the training of listeners in their use. Our research group then turned to studies of response forms and task variables in an effort to minimize the role of these factors in vowel studies. A good deal of this work was reported in posters and papers at the Acoustical Society of America; (see, for example, Verbrugge and Shankweiler 1977; Edman and Soskin 1977), but little of it was published in detail. Studies by Strange and Gottfried (1980), Gottfried and Strange (1980), Strange et al. (1979), and Carney et al. (1983), however, illustrate some of the extensive research that was being carried on. A few general findings are mentioned here:

1. In no study were naturally spoken, isolated AE vowels perceived more accurately than vowels produced and presented in CVC, VC, or even CV context. In most experiments, isolated vowels were much more poorly perceived. Experiments with the six oral stops as context for the vowel found that consonant environment significantly aided in vowel identification for all stops except /g/ (Gottfried and Strange 1980).

2. Naïve monolingual American English listeners required practice with feedback in the use of response forms. Without such training there were intrusive errors, most arising from the capricious English orthography. Many potential listeners could not overcome this problem in a reasonable time and had to be excluded from the experiments. All response tasks were not equivalent. However, CVC keyword response forms and rhyming word response forms, such as the one used by Macchi (1980) gave closely equivalent results and did not interact with the type of stimuli employed (Strange and Gottfried 1980). See also Assmann et al. (1982) for another examination of these response problems.
3. Vowel studies encounter phonotactic constraints that affect both CV syllables and isolated vowel utterances; namely, that lax vowels are not permissible in open syllables (word-finally) in English. Thus, lax vowels in both CV and isolated vowel syllables are disadvantaged in identification tasks. Even accounting for this factor, however, which contributes to identification differences between vowels in different conditions, there is still an advantage for vowels coarticulated with consonants. Additionally, it was found that relative duration information sometimes plays a role in disambiguating spectrally similar vowels (Strange et al. 1979).
4. Listeners had a slight advantage (about 4 %) in identifying vowels when they listened to recordings of their own voice as opposed to listening to recordings of other speakers of their dialect of English. However, they showed the consonantal context advantage even with their own productions (Carney et al. 1983).
5. When all concern with response forms was removed by asking listeners to perform a categorial ABX discrimination task with the stimuli being either vowels in CVCs or isolated vowels, significantly higher accuracy was observed for vowels in consonantal context (Gottfried et al. 1985).

5 The Silent Center Paradigm

5.1 Studies with CVC Citation-Form Syllables

In spite of several experiments showing the superiority of the identification of vowels in consonantal environment, many of our colleagues were reluctant to accept our interpretation that onset and offset transitional portions of syllables provided important information about vowel identity. At this point we conceived of a radical test of the hypothesis. Suppose we eliminated the quasi-steady-state portion of syllables and just presented the onsets and offsets alone at the proper temporal interval. If listeners could identify the vowels in this condition, it would conclusively demonstrate the importance of the transitional information. Of course, we were not sure what the listeners would hear. The syllables might be

incomprehensible, might fall apart, or simply be heard as two different signals. Would listeners even hear such stimuli as speech? Our friends at Haskins Laboratories allowed us to prepare such materials with their computer facilities amid great skepticism about the outcome. To everyone's surprise the vowels were directly perceived without difficulty even when as much as 65 % of the center of the syllable had been removed. We might also mention that the silences were also perceived; that is, the syllables were heard as containing a very large glottal stop in the middle of the identifiable vowel.

In the experiments that followed (Jenkins et al. 1983) we presented /bVb/ full syllables as *Controls*, *Silent Center* syllables where the center portion had been attenuated to silence as shown in Fig. 2, the *Centers* alone without the transitions, *Hiss-Center* syllables where the center portion was filled in with noise (from a /ʃ/ spoken in isolation), *Abutted* syllables where the beginning and end of each syllable were juxtaposed (silent interval removed), and the *Initial* portions and the *Final* portions of the syllables, presented separately. Figure 2 shows a schematic representation of the stimuli. Separate groups of listeners heard the stimuli in the various conditions. Both the *Control* syllables and the *Silent Center* syllables were identified with approximately 93 % accuracy. The *Centers* alone were somewhat poorer (87 % correct) but not significantly so, whereas *Hiss Centers* (81 % correct) were significantly poorer than the *Silent Center* syllables. *Abutted* syllables and the *Initials* and *Finals* alone were identified with far less accuracy: 70, 54 and 53 %, respectively. The results provided clear evidence that onset and offset transitions *taken together* and in their original temporal relationship furnished valid and important information concerning vowel identity.

A repetition of the experiment added a new condition, *Fixed Centers*. In this condition, the center portions of the syllables were trimmed to 60 ms so they were all of the same length, thus depriving the listeners of relative duration information and significant spectral change, in the longer vowels. The results of the study were almost identical to the first study. Interestingly, the accuracy of the *Fixed Center* condition was only 70 % as compared to 80 % for *Variable Centers*, thus confirming our earlier finding regarding the (small) contribution of relative duration information in identifying spectrally adjacent AE vowels.

We concluded with a view of vowels as intrinsically dynamic in nature. "According to this view, coarticulation of consonants and vowels is not to be considered as the introduction of 'noise' in the acoustic signal. On the contrary, the act of coarticulating phonemes in syllables gives rise to an acoustic array in which the consonants and vowels are cospecified in the time-varying spectral configuration" (Jenkins et al. 1983, p. 449).

At the same time, Strange et al. (1983) published a more analytic approach to examining the dynamic parameters of the stimuli. In the first experiment the original stimuli (*Controls*) were /bVb/ syllables spoken by a single male speaker with multiple tokens spoken at somewhat different rates of speech. The digitally altered stimuli consisted of *Silent Center* stimuli, *Variable Center* stimuli, *Fixed Center* stimuli trimmed to the length of the shortest original Center stimulus, *Shortened Silent Centers* where the silent interval was set to the length of the

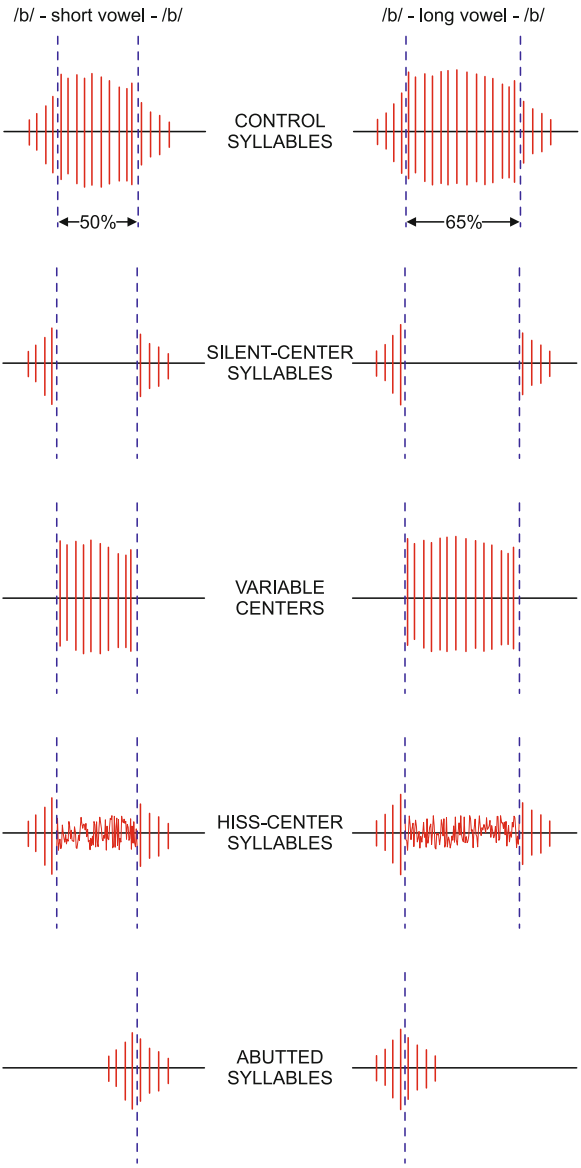


Fig. 2 Schematic representation of the stimuli in the first Silent Center vowel experiment. Adapted from Jenkins et al. (1983)

shortest interval of the *Silent Centers* (57 ms), *Lengthened Silent Centers* where the silent interval was set to the longest interval of the *Silent Centers* (163 ms), and the *Initia*l and *Final* stimuli as before.

Accuracy in the identification of vowels was above 90 % for the *Control*s, *Silent Centers*, *Shortened Silent Centers* and *Variable Centers* indicating that important and independent sources of information for vowel identity were carried in onset-offset dynamic spectral patterns, relative duration, and vowel inherent

spectral change (VISC) (especially for intrinsically long vowels). *Lengthening Silent Centers* resulted in 87 % correct identifications with most of the errors being the short vowels identified as their longer spectral counterparts. *Fixed Centers* were correctly identified only 79 % of the time with most of the errors being long vowels identified as their short vowel counterparts. Again it was apparent that both dynamic spectral patterns and temporal relationships were important in specifying the intended vowels in AE (see Hillenbrand 2013, Chap. 2, for a discussion of the role of relative duration information in the perception of AE vowels).

The study was repeated in a second experiment with stimuli from four speakers, two men and two women, with very much the same results, although one of the female speakers had a dialect very different from the Midwest dialect of the listeners and generated a higher error rate for all of the *Silent Center* conditions. Overall, it was impressive that dynamic spectral information was effective even when temporal relationships specifying intrinsic vowel duration differences were eliminated. Acoustic analysis found consistent differences between tense and lax vowels (including lax, but long /æ/) not only in vocalic duration but also in the length and manner of transition into and out of the "vowel target." This reflected the findings of Lehiste and Peterson (1961) that lax vowel production involved not only a short "target" but also a slower release into the final consonant while tense vowels involved both a longer hold in the target position as well as a more rapid closing gesture, as indicated by more rapid offset transitions.

5.2 Studies with Sentence Materials

At this point we decided to use sentence materials in our experiments as a first step in developing a more ecologically appropriate study of the perception of vowels as they are produced in continuous speech. Strange (1989b) presents detailed analyses of the results of three such experiments. In these experiments a new editing procedure was employed for the *Silent Centers.* Instead of designating proportions of the syllables as *Initials, Centers* and *Finals*, as we had done in the earlier studies, the *Initial* portion was defined as the first three pitch periods of the syllable and the *Final* portion was defined as the last four pitch periods of the syllable. This procedure provided no temporal cue to the duration of the syllable in the initial and final segments taken separately; long vowels simply had greater silences in them than shorter vowels. A second change was that all stimuli were presented in their original sentence context in the test situation. The experimental conditions were then designed to examine the relative contribution of information available in the vocalic nuclei, the intrinsic duration information specified by syllable length and the dynamic information available in the syllable onsets and offsets, taken together.

The first two experiments with three different consonant contexts (/bVb/, /dVd/, /dVt/) demonstrated first, that *Silent Centers* were perceived more accurately than other modified conditions (84 % correct, with most of the errors on short, lax vowels); second, that equalizing duration in the *Silent Center* syllables decreased

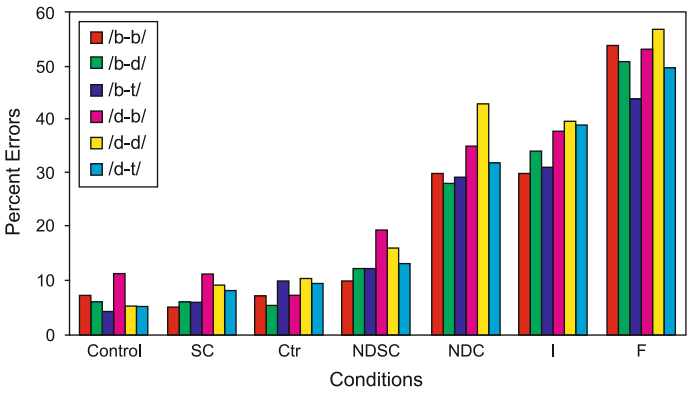


Fig. 3 Percent identification errors by consonantal context for each stimulus condition collapsed over all ten vowels. Control: full syllables. *SC* silent centers, *Ctr* centers, *NDSC* neutral duration silent centers, *NDC* neutral duration centers, *I* initial portions only, *F* Final portions only. Adapted from Strange (1989b)

accuracy somewhat (77 % correct), and, third, that presenting only one consonant environment at a time, (i.e. blocked presentation) vs. randomized presentation of the consonant types did not affect identification accuracy.

The third experiment with a new talker employed six consonantal contexts, /bVb/, /bVd/, /bVt/, /dVb/, /dVd/ and /dVt/) produced in the sentence frame, "I say the word CVC some more". Seven stimulus conditions were prepared: *Controls* with the full syllables, *Silent Centers* with the center of the syllables attenuated to silence, *Neutral Duration Silent Centers* with the silent interval set equal to the average length of the original silent portions, *Centers* where the Initial and Final portions of the syllables were attenuated to silence, *Neutral Duration Centers* where only the 4th through 7th pitch period of the original syllables were retained, the *Initial* portions alone, and the *Final* portions alone. The results of the study are shown in Fig. 3.

It is apparent that the *Controls*, *Silent Centers* and (variable duration) *Center*s were highly informative as to the identity of the vowel, with accuracy ranging from 89 to 95 % across contexts. *Neutral Duration Silent Centers* were slightly poorer, especially in the initial /d/ contexts, with accuracy from 81 to 90 %. *Neutral Duration Centers* fared even more poorly in all contexts with accuracy ranging from 57 to 72 %. Finally, the *Initials* (65 % correct overall) and especially the *Finals* (48 % correct overall) were very poorly identified. Detailed analyses showed that neutralizing duration in the silent centers led to increased errors mostly on the lax vowels, including long /æ/. Neutralizing duration in the centers, however, resulted in increased errors on all 10 vowels.

Acoustic analysis of the formant trajectories examined the symmetry of patterning of the first formant (F1) onsets and offsets. Again, as Lehiste and Peterson (1961) and our earlier study had reported, rapid onsets and offsets were observed in tense vowels, whereas lax vowels (including /æ/) showed rapid onsets but

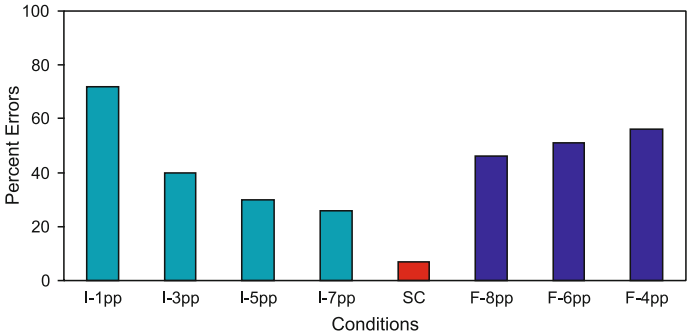


Fig. 4 Accuracy of vowel identification in the parametric study. *I* initial, *pp* pitch periods, *SC* silent center, *F* final

asymmetrically slower offsets. The discussion again emphasized the importance of dynamic spectral change and of time-varying characteristics in specifying the identity of the vowel.

In an effort to explore the effect of the amount and location of syllable portions on vowel identification, a parametric study was conducted (Jenkins and Strange 1999). Ten vowels spoken in a short sentence in a /dVd/ syllable were recorded. These stimuli were edited so listeners heard the sentence frame with only a short segment at the beginning of the syllable or at the end of the syllable or in the usual Silent Center preparation. Testing the *Initial* part of the syllable, there were four conditions: the listener heard 1, 3, 5, or 7 pitch periods with the rest of the syllable attenuated to silence. Testing the *Final* part of the syllable the listener heard 8, 6, or 4 pitch periods with the beginning of the syllable attenuated to silence. The *Silent Center* stimuli consisted of the Initial 3 pitch period portion plus the Final 4 pitch period portion with a silent interval that was the average of the silent period of all the vowels. Thus, this was a *Neutral Duration Silent Center* condition. The results of the study are shown in Fig. 4. Vowels in the *Neutral Duration Silent Center* condition (93 % correct) were identified significantly more accurately than vowels in any of the *Initials* (<75 % correct) or *Finals* (<55 % correct) conditions.

Overall, initial segments were more informative of vowel identification than were final segments. In initial segments, identification of all vowels except /o/ improved rapidly with increasing numbers of pitch periods. No such clear trend was seen for final segments; identification of /æ, ɑ, e, o/ never rose above 20 % correct even with eight pitch periods. Examining identification accuracy with respect to vowel height and position revealed an interesting pattern. Identification as to front-back position was fairly accurate (87 %) with four final pitch periods and highly accurate (95–98 %) with three to seven initial pitch periods. Correct decisions as to vowel height, however, were poor for all the final segments (less than 25 % correct) and required five to seven initial pitch periods to reach 80 % accuracy. Importantly, the combination of two ambiguous segments (three pitch periods initially and four pitch periods finally) resulted in very accurate vowel

identification even in the absence of durational differences in the syllables. This finding provides strong evidence that the spectro-temporal *relationships* between opening and closing transitional portions (e.g., F1 transition symmetry vs. asymmetry) play an important role in specifying AE vowels. It argues against the hypothesis that nucleus + offglide direction information provides the critical dynamic spectral information for AE vowels in continuous speech contexts (Andruski and Nearey 1992; Nearey 2013, Chap. 4; Morrison 2013, Chap. 11). The longest *Finals* (F-8 pp) contained target plus offglide information, and yet, were very poorly perceived.

In 1987, a special symposium on vowel perception was organized at a meeting of the Acoustical Society of America. Three papers from that meeting and the overview by the Chair were published in the society's journal in 1989 (see Miller 1989; Nearey 1989; Strange 1989a, b). These papers provide valuable summaries of somewhat different approaches to the problem of vowel perception but all of them mark the end of the belief that steady states or targets are the basis of vowel perception. The papers are also rich in suggestions as to further experimental and theoretical work.

5.3 Dynamic Invariance Over Speakers

In an effort to further test the limits of dynamic invariance, we performed a new version of an experiment reported by Verbrugge and Rakerd (1986) and Andruski and Nearey (1992) in which they switched speaker identity in the midst of the test syllable. Such a study examines the robustness of dynamic spectro-temporal specification, independent of speaker identity. It also speaks to the counter-hypothesis that syllable onsets and offsets merely "point to" the targets, which provide the most important information for vowel identity. Unfortunately, both previous studies used citation-form syllables and both had relatively low rates of accuracy in vowel identification (75 % for /bVb/ syllables in Verbrugge and Rakerd study, 65 % for /bVb/ syllables and 75 % for isolated vowels in Andruski and Nearey). Because these error rates are two to three times as great as in our Silent Center studies, and because we wanted to test the finding in a more natural context, we replicated the study with CVC syllables spoken in sentential context (Jenkins et al. 1994).

To induce strong coarticulation of the vowel we chose /dVd/ syllables spoken briskly in the sentence "I say the word dVd some more." We recorded a male speaker first. The female speaker then listened to each of the male's sentences until she felt she could match its tempo and then recorded the same sentence. *Silent Center* stimuli were prepared for the male using the beginning of the sentence plus the first three pitch periods of each test syllable as the *Initial* portion and the last four pitch periods of the test syllable and "some more" as the *Final* portion. For the female speaker the same procedure was followed except that the portions of the test syllable were matched in duration to the male portions rather than using the number of pitch periods (average of 6 pitch periods for the initial portion, 8 pitch periods for the final portion).

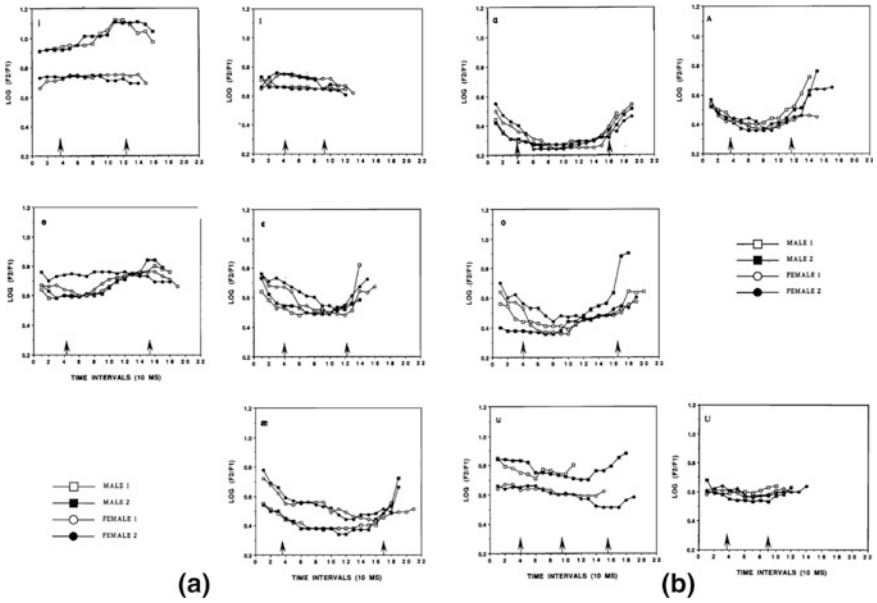


Fig. 5 The ratios of F1/F2 in the hybrid vowel study. From Jenkins et al. (1994 p. 1038)

Five conditions were generated from these materials: *Controls*, *Initials,* and *Finals,* were constructed as usual. *Silent Centers* were constructed by cross-splicing two different tokens of sentences using the same vowel by the same speaker so that the Initial portion was taken from one sentence and the Final portion was taken from another. The *Hybrid Silent Centers* were made by appending the Initial portion of one speaker's sentence and the Final portion of the other speaker's matching sentence. The results showed that the identification accuracy in the *Controls* was 97 %; both the *Silent Centers* and the *Hybrid Silent Centers* were 87 %, while the *Initials* and *Finals* were 52 and 36 % respectively. The dramatic finding, of course, was the equality of the results for the S*ilent Centers* and the *Hybrid Silent Centers*; it was apparent that the switch from one talker's voice to the other's had no effect on the accuracy of identification of the vowels. Acoustic analysis showed that the results could not be accounted for by nucleus plus offglide remaining invariant across both *Silent Center* conditions, as Andruski and Nearey (1992) had suggested. The direction and extent of change from the end of onsets to the beginning of offsets for the *Hybrid Silent Center* stimuli differed markedly from that of the *Silent Centers* in many cases. The outcome is much more in accord with the view of vowels as having distinctive *styles of articulatory gestures* that necessarily produce a variety of dynamic acoustic consequences in their execution.

In an effort to capture the changes for these two speakers, following the suggestion of Miller (1989), we plotted the ratios of the first two formants (the logarithm of F2/F1) during the syllable for each speaker for each of their two tokens of each of the

10 vowels (Fig. 5). It is apparent in the graphs that there is global similarity in the *temporal course of formant ratio change* for most of the vowels for these two speakers, even though the direction and extent of change from "target" to "offglide" varies within and across speakers. No one index, of course, captures all of the dynamics conveyed in the acoustics, but the graphs suggest that single points of comparison must vastly oversimplify the information that is available.

5.4 Dynamic Invariance Over Consonantal Context

A second severe test of the importance of transitional information was to use the Silent Center technique to splice together syllables of a single vowel but change the consonant environment in mid syllable. For example, from a /bVb/ syllable and a /dVd/ syllable one could create the usual *Silent Center* versions. Then, by rearranging the initial and final portions of those syllables, one could create a bVd *Silent Center* syllable and a /dVb/ *Silent Center* syllable. Jenkins et al. (1999) performed that experiment with ten vowels and six different consonant environments. Cross splicing of the test syllables resulted in new syllables that varied place of articulation, voicing, and both place of articulation and voicing from the original syllables. The experimental conditions compared *Silent Center* syllables, *Mixed Consonant Silent Center* syllables, and versions of these two conditions with the silent durations equated across the ten vowels (*Neutral duration Silent Centers* and *Neutral duration Mixed Consonant Silent Centers*). Results indicated that in the *Silent Center* and *Mixed Consonant Silent Center* conditions vowels were identified very accurately (96 % overall in both conditions). When the duration differences across vowels were held constant (*Neutral duration* conditions), vowels in both *Silent Center* and *Mixed Consonant Silent Center* conditions were identified with slightly lower accuracy (91 % overall). The increased errors, as usual, were largely on the lax vowels, including long, lax /æ/. As in previous studies, vowel *Initial* and *Final* portions presented separately were very poorly identified (66 and 47 % correct, respectively). Acoustic analyses again confirmed that direction and extent of formant change from the end of the initial portions to the beginning of the final portions of the mixed syllables differed from those of the original stimuli. However, the temporal style of F1 change of tense and lax vowels appeared to be invariant over changes in both place of articulation and voicing of the surrounding stop consonants, suggesting that the symmetry vs. asymmetry of F1 onsets and offsets provide critical information for differentiating adjacent tense and lax vowels in an F1/F2 target vowel space.

6 Dynamic Specification Over Speaking Rate and Speaking Style

Two studies explored the effects of rate of speech on vowel identification. In the first study (Johnson and Strange 1982) the speaker produced 11 vowels in the context of the sentence, "Was it the tVt sound that you heard?" The sentences were produced at the speaker's normal rate and then at a rate 25 % faster than normal. Test syllables were excised from their original sentences. Three listening tests were constructed: the test syllables in isolation (*Isolated Syllables*), the rapid and normal test syllables embedded in the sentence produced at the normal rate (*Normal Sentences*) and both kinds of syllables embedded in the sentence produced at the rapid rate (*Rapid Sentences*).

Test syllables spoken at the normal rate were almost perfectly perceived (about 98 % correct) in all three experimental conditions, suggesting that the normal rate syllables had robust sources of information as to the identity of the vowels even when they were placed in a misleading temporal context. Test syllables spoken at the rapid rate, however, were much more affected by context. In the *Rapid Sentences,* they were perceived accurately 94 % of the time; embedded in *Normal Sentences* they were perceived accurately 89 % of the time, and presented as *Isolated Syllables* they were correctly identified only 81 % of the time. Analysis of the data disclosed that increased errors were almost entirely due to the misperception of the rapid long vowels (/e/ → /ɪ/, /æ/ → /ɛ/, /ɑ, ɔ/ → /ʌ/); identification of long vowels in the *Isolated Syllables* condition was only 68 %. This replicates the findings of Verbrugge and Shankweiler (1977).

A further experiment (Johnson and Strange 1982 Exp. III) found that identification of the vowels produced in rapid test syllables was improved to 95 % correct when the test included just the words preceding and following the test syllable ("the tVt sound") and to 92 % when only the following word was included. Finally, in two other conditions (Exp II) where the listeners were given only excised rapid syllables or were given both normal and rapid excised syllables, but were told that they had been taken from both normal and rapid rate speech, identification accuracy was significantly improved for the long vowels (82 and 80 %, respectively, compared to 68 % in Exp I). The writers interpreted the results as evidence for the importance of contextual temporal information in disambiguating long and short vowels in rapid speech where vowel targets are more centralized due to articulatory undershoot (see Fig. 1 in Johnson and Strange 1982). However, informing listeners that they would hear syllables spoken at both rapid and normal rates allowed them to switch to other sources of information to disambiguate long and short vowels with spectrally similar vowel targets.

In a more recent study (Stack et al. 2006), a female speaker produced 11 vowels in six different stop-vowel-stop contexts in the following sentence, "I hear the sound of /həCVC/ some more." The sentences were recorded at both the speaker's normal rate of speech and at a fast rate. In addition, a "neutral consonantal context" corpus was recorded with the disyllables /əhVt/ and /əhVd/ produced in

sentences at both rates and in citation form (lists) to establish the traditional canonical formant targets for this speaker. Separate groups of listeners were asked to identify the rate of speech of the various types of tokens. Listeners were 98 % correct in labeling speaking rate when presented the entire sentences but only 75 % successful in labeling rate of test disyllables excised from the sentences. In spite of this difference, listeners were highly successful in identifying the vowels in sentences (98 %) and almost as accurate in identifying vowels in the excised disyllables when they were blocked by rate (96 %) and even when the rapid and normal rate excised syllables were randomly intermixed (93 %). Although these results are for a single speaker, the findings are the same as those of von Son and Pols (1990, 1992) with a speaker of Dutch. The patterns of identification of the vowels were very little changed as the rate of speech increased.

Acoustical analysis disclosed that the formant values showed almost equal amounts of undershoot of mid-syllable formant frequencies in both fast and normal sentences, relative to "canonical values" established from citation-form /əhVd/ and /əhVt/ utterances (c.f. Hillenbrand et al. 2001; Stevens and House 1963). In addition, the patterns of formant movement from 25 through 50 to 75 % temporal points in each syllable were for the most part highly similar in direction and extent under both rates of speech, within the same consonantal context. However, both direction and extent of formant movement differed across different consonantal contexts; especially for the short vowels and the back rounded long vowels (see Stack et al. 2006). Thus, in sentence materials, formant movements within syllables appear not to be invariant for AE vowels across different consonantal contexts, but rather reflect both vowel inherent change (VISC) and coarticulatory patterning. This suggests that any studies of vowel inherent change should be tested on vowels produced in multiple consonantal contexts in sentence-length materials to see the extent to which such information remains available to perceivers in continuous speech utterances. The situation is further complicated by findings like those of Jacewicz and Fox (2013, Chap. 8) showing variations in spectral change across dialects and age cohorts of speakers of AE even within the same consonantal context.

7 Cross-Language Studies of Coarticulated Vowels

7.1 Dynamic Specification of German Vowels

The studies reviewed so far were all conducted with AE vowels and native AE listeners. Indeed, most other chapters in this volume report studies of perception of AE or Canadian English vowels by native listeners. (See Rogers et al. 2013, Chap. 10 for studies of AE vowel perception by non-native listeners.) It can be argued that many of the results detailed above are a consequence of American and Canadian vowels being highly diphthongized. However, we have

argued that in addition to diphthongization, other dynamic spectral information, such as F1 temporal trajectories, as well as vowel-intrinsic syllabic duration may be important for accurate identification of AE vowels in continuous speech utterances. (See also Hillenbrand 2013, Chap. 2). We expect that these variations and others will be important in the perceptual differentiation of vowels in other languages with large vowel inventories.

To evaluate the extent to which our Dynamic Specification approach applies more generally, we tested the Silent Center paradigm using vowels from German, a language that has a large vowel inventory (14 monophthongs; 7 tense/lax pairs) but that is widely agreed to have little diphthongization and more reliance than English on duration to separate tense and lax vowels. Strange and Bohn (1998) studied North German vowels produced in /dVt/ syllables in a sentence frame. Digital editing produced the following test stimuli: *Initials*, *Final*s, *Silent Centers*, and *Vowel Centers.* In a second experiment, the availability of vocalic duration information was minimized by creating *Fixed Silent Center* stimuli where the silent period in the test syllables was set to a constant value; *Fixed Center* stimuli where the *Centers* were all trimmed to four pitch periods each, and *Tense/Lax Fixed Silent Centers* in which the durations of spectrally adjacent tense/lax pairs were set to their pair-wise average. This last condition allowed duration differences associated with vowel height to remain, but removed pair-wise duration differences for vowels adjacent in F1/F2 vowel space.

The pattern of results closely matched the findings of the studies of AE vowels. North German vowels in *Silent Centers* were identified with about 90 % accuracy, *Centers* (with duration information available) were identified with 85 % accuracy, and *Initials* and *Finals* were accurately identified less than 50 % of the time. When duration was controlled, errors of identification increased as expected: *Tense/Lax Fixed Silent Centers* were identified with about 75 % accuracy; *Fixed Silent Centers* achieved about 70 % accuracy, and *Fixed Centers* only 53 % accuracy. These results reveal the importance of spectro-temporal relationships in the onsets and offsets of syllables for North German vowel identification just as in vowels in AE, and the greater importance of vocalic duration in differentiating tense and lax North German vowels. The poverty of information in the spectral targets (*Fixed Centers*) presented alone is also noteworthy. Errors included confusions in vowel height and roundedness, as well as long/short confusions.

Acoustical analysis of the stimuli confirmed that there was very little diphthongization, especially of the tense vowels, substantiating the common description of monophthongal German vowels. For short vowels, formant movement associated with coarticulation with consonants was apparent. Comparison of the temporal patterning of the first formant in the test syllables between AE and North German was interesting. As in AE, tense vowels showed rapid and symmetrical F1 temporal onsets and offsets. Lax vowels, like their AE counterparts, displayed asymmetrical onsets and offsets; however, unlike the AE lax vowels, their pattern was the opposite; they showed relatively slow onsets and rapid offsets.

The authors conclude,

> For languages with large vowel inventories, we speculate that differences in the 'style of movement' associated with vowel gestures serve to maintain the distinctiveness of phonetic categories in the face of the ambiguity of 'target' information for vowels as they are produced in continuous speech. (Strange and Bohn 1999 p. 503)

Further work by Bohn and Polka (2001) demonstrated that infants learning German can use either dynamic spectral information or spectral target information to discriminate contrasting vowels. They also found that infants weight duration information even more heavily than adults in the discrimination task. Bohn (1997) reviews these developmental data in arguing for a Dynamic Specification theory of native and non-native vowel perception. (See also Nittrouer 2007 for a developmental study of the perception of American vowels by 3- to 7 year-old English learners.)

7.2 Cross-Language Investigations of Contextual Variability of Vowels

In more recent work, we have explored the variation in vowel acoustics due to phonetic and prosodic contextual influences across languages with relatively large vowel inventories (German, French, and English) and languages with relatively small vowel inventories (Japanese, Russian, and Spanish). In a study of North German, Parisian French and New York English (AE), corpora from nearly monolingual men and women were collected in which the oral vowels were produced in "neutral" citation-form disyllables to establish canonical formant values, and in nonsense di/trisyllables imbedded medially in carrier sentences with both labial and alveolar preceding and following stop consonants (Strange 2007; Strange et al. 2007). Speakers produced the sentences in four prosodic/rate conditions: (1) normal rate with no narrow sentence focus; (2) rapid rate with no focus; (3) normal rate with focus on the target word; and (4) normal rate with focus on the word preceding the target word (i.e., with the target word in post-focus position). In these studies, speakers were instructed to "talk as if you were talking to a familiar native speaker of the language" and they were given sufficient practice with the nonsense utterances so that they could utter the sentences fluently at their self-selected normal rate of speaking and then at a rate "as fast as you can go without leaving out any of the parts of the sentence."

Contextual variation in target undershoot (Lindblom 1963, 1983), as well as differences in relative syllable duration and fundamental frequency across these conditions were compared within and across languages. Formants were measured at 25, 50 and 75 % temporal points in the syllables in order to investigate spectral change. However, in the sentence materials, formant movements in the middle half of AE syllables were so influenced by consonantal coarticulation that only comparisons of contextual changes in 50 % formant values across languages and phonetic/prosodic contexts were reported.

Of particular interest was a cross-language comparison of the contextual fronting of back, rounded vowels (i.e., variations in F2/F3 values) in coronal contexts in languages with and without phonologically contrastive front, rounded vowels (North German and Parisian French vs. AE), and the amount of contextual raising of low and mid-low vowels (F1 values) in languages that do and do not use alternations in vocalic duration/tenseness in differentiating vowels of different heights (AE and North German vs. Parisian French). Both kinds of contextual variability differed across the three languages in notable ways, suggesting that "target undershoot" was due not only to biomechanical principles, but rather was a function of language-specific constraints that served to maintain articulatory and acoustic distinctiveness.

With respect to fronting, as expected, AE vowels in coronal context showed the most fronting of back, rounded vowels, such that AE /u, ʊ, o/ were more similar to North German front than to back, rounded vowels in this context. That is, English, which does not have *distinctive* front, rounded vowels, is free to adopt allophonically fronted variations of back, rounded vowels in coronal consonantal context. The evidence that this is an allophonic variation, as opposed to biomechanical undershoot, comes from the fact that the amount of fronting differs very little as a function of variations in sentence focus and speaking rate, even though these conditions led to differences in vocalic duration (See Fig. 6 American English panel). That is, the *targets* for AE /u/, /ʊ/, /o/ appear to be specified as front vowels in coronal contexts.

Canonical forms of Parisian French front, rounded vowels were more front than were North German front, rounded vowels (i.e. higher F2 values). In coronal contexts, Parisian French back, rounded vowels were contextually more fronted than for North German back vowels, with concomitant (slight) fronting of front, rounded Parisian French vowels (See Fig. 6 panels for German and French). In both languages, these patterns of variation served to maintain distinctiveness between front and back, rounded vowels in the F2/F3 plane of the acoustic vowel space.

With respect to the raising of low vowels, Parisian French /a/ was more raised than either North German or AE low vowels, except when it occurred in Focus position in the sentence. This was due in large part to the fact that Parisian French vowels in non-focus sentence contexts were much shorter than in citation-form context. With respect to vowel height distinctions in North German and AE vowels, Fig. 6 shows that only four vowel heights (F1 values) were acoustically distinct in both citation and connected speech utterances. The so-called mid, long vowels /eː/, /oː/ and the so-called mid-high short vowels /ɪ/, /ʊ/ overlapped in F1 target values, but were differentiated by position (F2), with the short vowels more centralized and long vowels more peripheral.

Finally, relative duration differences of short vs. long vowels were more distinct in North German vowels than in AE vowels in rapid and post-focus contexts. In North German, increases in fundamental frequency and amplitude were employed to a greater extent to signal sentence focus. In Parisian French, because intrinsic vowel duration is neither phonologically contrastive (as in German by some

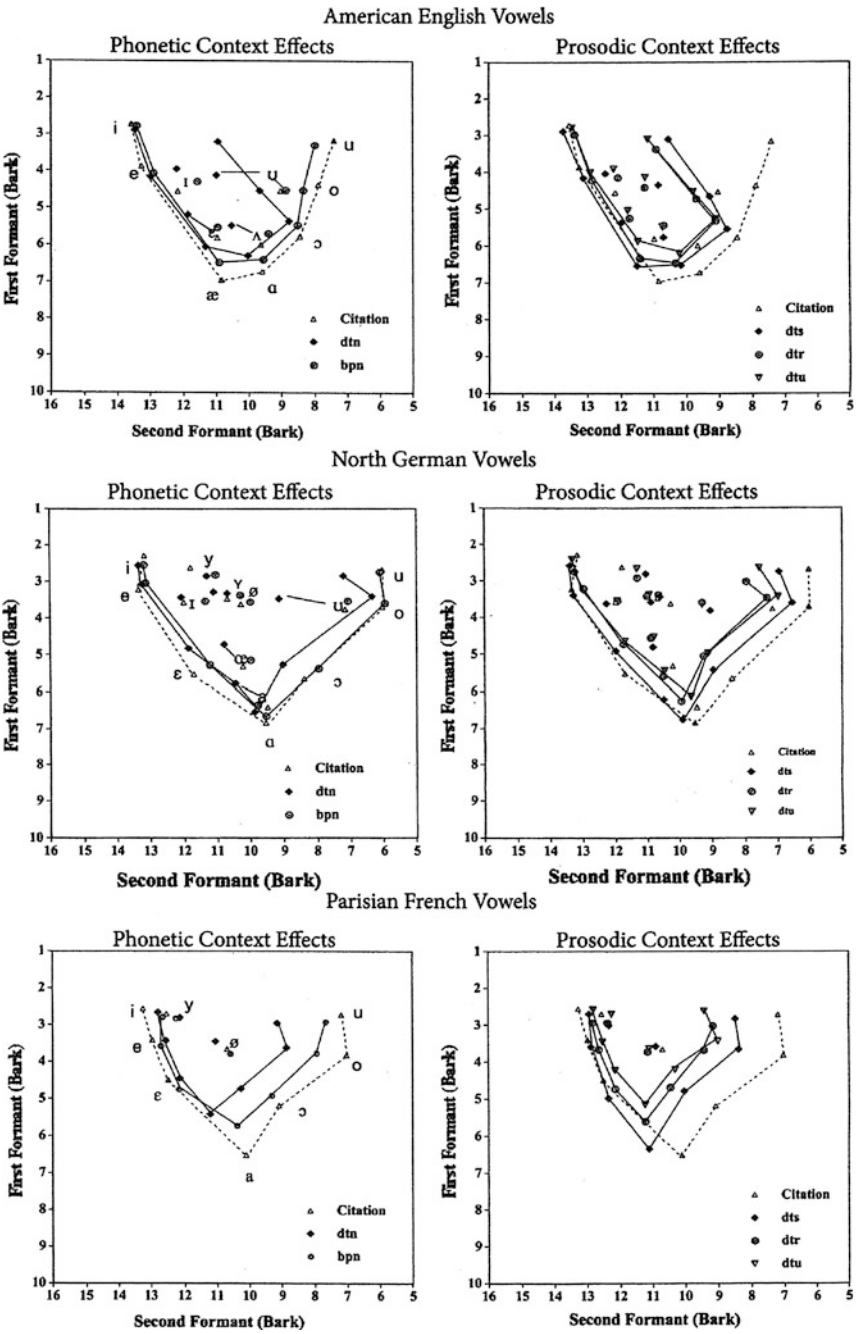


Fig. 6 German, French and English vowel spaces from Strange (2007)

linguistic analyses) nor phonetically systematic (as in both North German and AE), vowel durations varied most with differences in sentence focus (Focus/Postfocus ratios: 1.31 for Parisian French, 1.13 for AE, 1.06 for North German).

In our study of languages with small vowel inventories (Law et al. 2006), we have investigated the contextual variability of vowel target information, as well as the use of dynamic spectral patterns such as palatalization and relative vocalic duration to further distinguish vowels. In Japanese, vowel length is phonologically contrastive: long (two-mora) vs. short (one-mora) vowels remain very different temporally (vocalic duration almost 2 to 1) even in rapidly produced disyllables positioned medially in sentences. Long vowels show relative small amounts of target undershoot while short vowels show more, as would be expected. However, the three vowel heights are maintained with little overlap even in short vowels coarticulated with consonants in rapid speech. In addition, Japanese also has three palatalized versions of long and short /a/, /aa/, /o/, /oo/, /ɯ/, /ɯɯ/. Russian does not use vocalic duration to contrast vowels, but does contrast palatalized and non-palatalized versions of 4 vowels /e/, /a/, /o/, /u/. Finally, Spanish has the very frequently attested five-vowel system (Maddieson 1984) with no phonologically contrastive length, palatalization, or diphthongization. In sentential materials spoken at normal and rapid rates, and with target disyllables in focus and in post-focus sentence contexts, distributions of target formant values show very little overlap.

7.3 Adaptive Dispersion Theory and Vowel Distinctiveness

Lindblom and colleagues have developed the Theory of Adaptive Dispersion and Hyper & Hypo Theory to account for preferred vowel inventories in the languages of the world and for variations in coarticulatory undershoot of vowel targets in various speech registers (e.g., "clear" vs. "conversational" speech). In later versions of these theories (Diehl and Lindblom 2004), the concept of *sufficient distinctiveness* is introduced to characterize the dynamic interplay between talker-oriented *articulatory ease* and listener-oriented *perceptual distinctiveness* in accounting for differences in coarticulatory undershoot of vowel targets in various speaking registers. (See also Kluender et al. 2013, Chap. 6 for a similar approach to vowel specification). Another concept, utilized to motivate preferred vowel inventories, is the *auditory enhancement hypothesis* offered by Diehl and Kluender (1989). Under this hypothesis, articulatory gestures that lead to similar acoustic consequences tend to be jointly specified as preferred phonological segments. Thus, vowels with lip rounding (which lowers formants) and back tongue position (which also lowers upper formants) tend to be preferred over back, unrounded or front, rounded vowels.

Models of preferred target vowel specifications in a static F1/F2/F3 space (or alternatively, an F1/F2' space in which F3 weights F2 parameters) were predicted for inventories of up to 11 vowels (see Fig. 3.15 in Diehl Lindblom 2004). In these

predicted inventories, peripheral (front and back) vowels are preferred over central vowels; and front, unrounded and back, rounded vowels are preferred over front, rounded vowels. In addition, back vowels are differentiated by only three levels of vowel height (F1 dimension) for inventories up to nine vowels and four vowel heights for 10–11 vowel inventories, whereas front vowels are differentiated by three levels of height for inventories up to eight vowels, four levels for a 10-vowel inventory, and five levels for an 11-vowel inventory.

The data on canonical values and contextual variation of mid-syllable formant frequencies of German, French, and American vowels summarized above (Strange et al. 2007, see Fig. 6) both support and challenge the principles of the Theory of Adaptive Dispersion model. All three languages have vowel inventories of nine or more (oral) vowel targets that appear to violate some of the constraints specified in the Theory of Adaptive Dispersion and the Auditory Enhancement Hypothesis. Both French and German have front, rounded vowels, with F2 values that tend to be closer to front, unrounded vowels than to back, rounded vowels. Thus, they fall in the F1/F2 plane of the vowel space between peripheral front and back, high to mid vowels predicted from the Theory of Adaptive Dispersion models for inventories of eight or more vowels. However, they are not made with a central tongue position, but rather have mid F2 values because of the offsetting raising of F2 by the front tongue position and lowering of F2 by lip rounding.

With respect to distinctiveness in vowel height (F1 values), French, (9 oral vowels) has four vowel heights for both front, unrounded and back, rounded vowels and these vowels are not further differentiated by diphthongization or vowel duration differences (but see discussion below). North German (14 vowels) and (New York) American English (11 vowels) have front and back vowel series that are sometimes described as being differentiated by five heights (high, mid–high, mid, mid–low, low). However, F1 values collapse into four levels even in canonical forms, with /eː/, /ɪ/ and /oː/, /ʊ/ (and /oː/, /Y/ in North German) having very similar F1 values. To maintain distinctiveness of these pairs, the two languages use different kinds of dynamic spectro-temporal information. In this (and other dialects of AE) these vowel pairs are differentiated by spectral change, at least in canonical forms, and by shortening and centralization (F2 values) of the short, lax vowels in continuous speech materials, as well as by F1 temporal trajectory differences. In North German, all three vowel pairs are differentiated by F1 temporal trajectories, vocalic duration, and centralization (F2 values), but not diphthongization (VISC).

Thus, in continuous speech contexts, the AE 11-vowel system and the North German 14-vowel NG system appear to maintain four distinctive vowel heights (F1), with dynamic spectro-temporal parameters further differentiating adjacent vowels in the front and back series. The one exception to this is that in English, /ɔː/ is a long, tense vowel, whereas in German it is a short, lax vowel. This may be the reason why the /ɑː/, /ɔː/ distinction has been neutralized in many dialects of AE. In New York and other North Eastern US dialects, /ɔː/ is raised and diphthongized to maintain contrastiveness with monophthongal /ɑː/.

To summarize, from the data on the contextual variability of mid-syllable formant frequencies in North German, Parisian French, and New York English vowels, we can see how the maintenance of *sufficient distinctiveness* of vowel position (front vs. back) and height (high, mid, low) plays out differently across languages with relatively large inventories. When back vowels are "free" to be fronted in coronal contexts in the service of articulatory ease, they are. In languages with contrastive front and back, rounded vowels, the extent of coarticulatory fronting appears to be constrained by the relative positions of the front, rounded vowels along the F1/F2 dimension. The maintenance of four vowel heights is aided (in North German and AE) by additional dynamic spectro-temporal patterns in continuous speech contexts in which tongue/jaw coarticulatory raising may reduce the acoustic distinctiveness of F1 targets. These data are in line with the analysis of German (and perhaps American English) as having three vowel heights (high, mid, low) with tenseness (closeness, length) as additional distinctive features.

7.4 Dynamic Dispersion Hypothesis

On the basis of these cross-language explorations of contextual variability of vowels, we offer the following hypothesis. Due to the coarticulatory constraints in ongoing speech production, sufficient perceptual distinctiveness of four vowel heights is difficult to maintain. That is, because all consonants require a higher jaw position than low vowels in order to complete the necessary constrictions of the vocal tract, there is a tendency for the height dimension of the vowel space (F1) to "contract" in continuous speech contexts, leading to the acoustic "raising" of low and mid-low vowels. Therefore, for languages with large vowel inventories to maintain distinctiveness in height of greater than the "preferred" three (characterized by the features: +high, −high/-low (mid), +low,) variations in either duration and/or in the dynamic vowel gestures into and out of the vowel targets may be employed. Acoustically, these gestural styles are signaled by dynamic spectro-temporal parameters, such as vocalic duration, spectral change (diphthongization), F1 temporal trajectories (tenseness), and/or palatalization. In rapid speech, some of these parameters may be "hidden" by gestural overlap with consonants to a greater extent than others; however, the formant trajectories throughout the syllable nucleus provide rich information such that "sufficient distinctiveness" is maintained.

With respect to this hypothesis, Parisian French is a case in point. Because French vowels do not vary distinctively in tenseness or length, the distinction between mid and mid-low vowels appears to be under some pressure (from perceptual constraints) toward neutralization. In current speakers of this dialect, the distinction between front, rounded mid and mid-low /ø/, /œ/ has largely been neutralized (Tranel 1987), with allophonic variation accounting for most of the variability in formant targets. Front, unrounded /e/, /ɛ/ and back, rounded /o/, /ɔ/ also tend to occur largely in different syllable contexts, according to the "law of

position" in French (Monin 1988). However, the front vowels can both occur in syllable final position (open syllables). Indeed, in this context, /e/, /ɛ/ distinguish morphophonemic markers of tense (e.g. "parler*ai*" /e/ [1st pers. sing. future] "I will speak" vs. "parler*ais*" /ɛ/ [1st pers. sing. conditional] "I would speak"). The back vowels /o/, /ɔ/ both appear in closed syllables in just a few minimal pairs that may become homophones for some speakers. In 1988, Gottfried and Beddor reported that, whereas small duration differences were maintained for /ɔ/, /o/ by native speakers of Parisian French, they appeared not to attend to these duration differences perceptually. In our corpora, duration differences for these vowels in closed syllables were very small (/o/ only about 10 % longer than /ɔ/ on average). Thus, except for /e/, /ɛ/ in open syllables, which contrast in syntactic, as well as lexical distinctions, the mid vs. mid-low vowels of PF appear to be moving toward becoming allophonically conditioned variants exclusively. (See Law II 2009 for further research on the maintenance of distinctive /e/ vs. /ɛ/ in lexical and grammatical morphemes in Parisian and Canadian French.)

These tendencies in Parisian French suggest that differentiating more than three vowels in terms of target F1 values (height) is difficult to maintain unless further differences in the gestures into and out of the target position are incorporated. Future research on the contextual variation of vowels in languages with large vowel inventories will shed light on the extent to which "preferred" vowel spaces should be specified in terms of both static target values (as in the current Theory of Adaptive Dispersion) *and* dynamic gestural differences that we have dubbed "styles of movement" (see Steinlen 2005). The view we offer here is that by characterizing vowels as dynamic gestures with complex spectro-temporal "signatures" and by investigating their production and perception in speech materials that more closely resemble "real world" utterances, we will be able to answer lingering questions about what acoustic information distinguishes vowels in the languages of the world, and how that information is processed by native and non-native speakers of the language (see Rogers et al. 2013, Chap. 10).

8 Conclusions

More than 35 years of research by ourselves and others has resulted in a very different view of the acoustic specification of North American English vowels and their perception by native listeners. Much of this research is presented in the chapters of this volume. Here, we summarize our contributions to this new understanding of vowel perception and comment on their extension to vowel production and perception in other languages:

1. North American English (AE) vowels coarticulated with consonants in syllables, produced either as citation-form single words or in sentence materials, are perceived highly accurately by adult native listeners, despite the ambiguity of static target spectral information (overlap in F1/F2 space) caused by variability

across speakers, contexts, speech styles, and speech rates. Under most testing conditions, identification of coarticulated vowels is significantly better than for isolated vowels spoken by the same speakers, even though differentiation of static target information is greater for the latter utterances.

2. Acoustic patterns of CVC syllables containing AE vowels provide rich, time-varying information about vowel identity, including vowel inherent spectral change (VISC), characteristic tense (symmetrical) vs. lax (asymmetrical) F1 temporal trajectories, and differences in intrinsic vocalic duration.
3. Using the Silent-Center testing paradigm, it was demonstrated that native listeners utilize these sources of dynamic information in identifying vowels. Static target information is neither necessary nor sufficient to specify coarticulated AE vowels unambiguously. The information in syllable onsets and offsets, taken together, is critical for accurate identification of spectrally-adjacent tense and lax AE vowels.
4. Research with North German vowels extended the findings of studies of AE vowels to another language with a relatively large vowel inventory. This was important because North German vowels are characterized by very little vowel inherent spectral change even in canonical productions (little diphthongization). Results of perceptual studies showed that German listeners depend more than AE listeners on vocalic duration to disambiguate spectrally-adjacent German vowels, but that both language groups use dynamic spectro-temporal information specified by syllable onsets and offsets, taken together, in identifying vowels.
5. Acoustic analyses of coarticulated vowels produced in sentence-length utterances in several languages have shown *language-specific* distinctive patterns of contextual variation as a function of phonetic and prosodic environmental influences. This supports the hypothesis that coarticulatory variation is constrained by the need to maintain "sufficient distinctiveness" of vowels. Dynamic spectro-temporal information is critical in differentiating vowels in languages with large vowel inventories.
6. Future research is needed to specify more precisely what dynamic spectral information remains available in continuous speech contexts, and how native listeners utilize such information in identifying vowels. A cross-language approach is valuable in distinguishing language-specific constraints for maintaining distinctiveness from universal biomechanical influences on the articulation and resulting acoustic specification of vowels. Research on the development of vowel perception and on vowel perception by non-native listeners will also shed light on how perceivers learn the language-specific constraints that are used to maintain sufficient distinctiveness of vowels in both large and small vowel inventory languages.

Acknowledgments The research work reported in this chapter has been supported by the Center for Research in Human Learning at the University of Minnesota, the National Institutes of Health, the National Science Foundation and the Universities of Minnesota and South Florida as well as the Graduate Center of the City University of New York. We are grateful to all these institutions for their support.

References

Andruski, J.E., Nearey, T.M.: On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables. J. Acoust. Soc. Am. **91**, 390–410 (1992). doi:10.1121/1.402781

Assmann, P.E., Nearey, T.M., Hogan, J.: Vowel identification, orthographic, perceptual and acoustic aspects. J. Acoust. Soc. Am. **71**, 975–989 (1982). doi:10.1121/1.387579

Bell, A.G.: The Mechanism of Speech, 5th edn. Funk and Wagnall's, New York (1911)

Bohn, O.-S.: There's more to learning non-native vowels than establishing gestural and acoustic "targets". In: Leather, J., James, A. (eds.) New Sounds 97: Proceedings of the Third International Symposium on the Acquisition of Second-Language Speech. Universitatsverlag, Klagenfurt, pp 38–46 (1997)

Bohn, O.-S., Polka, L.: Target spectral, dynamic spectral, and duration cues in infant perception of German vowels. J. Acoust. Soc. Am. **110**, 504–515 (2001). doi:10.1121/1.1380415

Carney, A.E., Edman, T.R., Strange, W., Jenkins, J.J.: Advantage of speaker as listener in a vowel identification task. J. Acoust. Soc. Am. **73**, 2222–2223 (1983). doi:10.1121/1.389550

Cooper, F.S.: Spectrum analysis. J. Acoust. Soc. Am. **22**, 761–762 (1950). doi:10.1121/1.1906683

Diehl, R.L., Kluender, K.R.: On the objects of speech perception. Ecol. Psychol. **1**, 121–144 (1989). doi:10.1207/s15326969eco0102_2

Diehl, R.L., Lindblom, B.: Explaining the structure of feature and phoneme inventories: the role of auditory distinctiveness. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R. (eds.) Speech processing in the auditory system. Springer, New York, pp 101–162. doi:10.1007/0-387-21575-1_3 (2004)

Diehl, R.L., McCusker, S.B., Chapman, L.S.: Perceiving vowels in isolation and in consonantal context. J. Acoust. Soc. Am. **69**, 239–248 (1981). doi:10.1121/1.385344

Edman, T.R., Soskin, R.: Perceptual learning of vowel identity. J. Acoust. Soc. Am. **61**, 39. doi:10.1121/1.2015630

Gottfried, T.L., Beddor, P.S.: Perception of temporal and spectral information in French vowels. Lang. Speech **32**, 57–75 (1988). doi:10.1177/002383098803100103

Gottfried, T.L., Jenkins, J.J., Strange, W.: Categorial discrimination of vowels produced in syllable context and in isolation. Bull. Psychon. Soc. **23**, 101–104 (1985)

Gottfried, T.L., Strange, W.: Identification of coarticulated vowels. J. Acoust. Soc. Am. **68**, 1626–1635 (1980). doi:10.1121/1.385218

Helmholtz, H.L.F.: On the sensations of tone. (Translated by A.J. Ellis. Reprinted in 1954). Dover, New York (1885/1954)

Hillenbrand, J.M.: Static and dynamic approaches to understanding vowel perception. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change, Chap. 2. Springer, Heidelberg (2013)

Hillenbrand, J.M., Clark, M.J., Nearey, T.M.: Effect of consonant environment on vowel formant patterns. J. Acoust. Soc. Am. **109**, 748–763 (2001). doi:10.1121/1.1337959

Jacewicz, E., Fox, R.A.: Cross-dialectal differences in dynamic formant patterns in American English vowels. In: Morrison, G.S., Assmann P.F. (eds.) Vowel Inherent Spectral Change, Chap. 8. Springer, Heidelberg (2013)

Jenkins, J.J., Strange, W.: Perception of dynamic information for vowels in syllable onsets and offsets. Percept. Psychophysics **61**, 1200–1210 (1999). doi:10.3758/BF03207623

Jenkins, J.J., Strange, W., Edman, T.R.: Identification of vowels in "vowel-less" syllables. Percept. Psychophysics **34**, 441–450 (1983). doi:10.3758/BF03203059

Jenkins, J.J., Strange, W., Miranda, S.: Vowel identification in mixed-speaker silent-center syllables. J. Acoust. Soc. Am. **95**, 1030–1043 (1994). doi:10.1121/1.410014

Jenkins, J.J., Strange, W., Trent, S.A.: Context-independent dynamic information for the perception of coarticulated vowels. J. Acoust. Soc. Am. **106**, 438–448 (1999). doi:10.1121/1.427067

Johnson, K.: Acoustic and Auditory Phonetics. Blackwell, Cambridge (1997)
Johnson, T.L., Strange, W.: Perceptual constancy of vowels in rapid speech. J. Acoust. Soc. Am. **72**, 1761–1770 (1982). doi:10.1121/1.388649
Joos, M.A.: Acoustic phonetics, *Language* Supplement. *24*, pp. 1–136. Stable URL: http://www.jstor.org/stable/522229 (1948)
Kluender, K.R., Stilp, C.E., Kiefte, M.: Perception of vowel sounds within a biologically realistic model of efficient coding. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change, Chap. 6. Springer, Heidelberg (2013)
Ladefoged, P., Broadbent, D.E.: Information conveyed by vowels. J. Acoust. Soc. Am. **29**, 98–104 (1957). doi:10.1121/1.1908694
Law II, F.F., Gilichinskaya, Y.D., Ito, K., Hisagi, M., Berkowitz, S., Sperbeck, M.N., Monteleone, M., Strange, W.: Temporal and spectral variability of vowels within and across languages with small vowel inventories: Russian, Japanese, and Spanish (A). J. Acoust. Soc. Am. **120**, 3296 (2006)
Law II, F.F., Strange, W.: Maintenance of /e-ɛ/ in word-final position as a phonemic and morphemic contrast in Canadian French (A). J. Acoust. Soc. Am. **125**, 2757 (2009)
Lehiste, I., Peterson, G.E.: Transitions, glides and diphthongs. J. Acoust. Soc. Am. **33**, 268–277 (1961). doi:10.1121/1.1908638
Lieberman, P., Crelin, E.S., Klatt, D.H.: Phonetic ability and related anatomy of the newborn and adult human, Neanderthal man and the chimpanzee. Am. Anthropologist **74**, 287–307. doi:10.1525/aa.1972.74.3.02a00020 (1972)
Lindblom, B.: Spectrographic study of vowel reduction. J. Acoust. Soc. Am. **35**, 1773–1781 (1963). doi:10.1121/1.2142410
Lindblom, B.: Economy of speech gestures. In: MacNeilage, P.F. (ed.) Speech Production, Springer-Verlag, New York pp. 217–245 (1983)
Lindblom, B., Studdert-Kennedy, M.: On the role of formant transitions in vowel recognition. J. Acoust. Soc. Am. **42**, 830–884 (1967). doi:10.1121/1.1910655
Macchi, M.J.: Identification of vowels spoken in isolation versus vowels spoken in consonantal context. J. Acoust. Soc. Am. **68**, 1636–1642 (1980). doi:10.1121/1.385219
Maddieson, I.: Patterns of Sound. Cambridge University Press, Cambridge (1984). doi:10.1017/CBO9780511753459
Miller, J.D.: Auditory-perceptual interpretation of the vowel. J. Acoust. Soc. Am. **85**, 2114–2134 (1989). doi:10.1121/1.397862
Monin, Y.-C. Loi de position? Revue Québécoise de Linguistique, (1988) 17, pp. 237–243. Stable Url: http://id.erudit.org/iderudit/602622ar
Nearey, T.M.: Static, dynamic, and relational factors in vowel perception. J. Acoust. Soc. Am. **85**, 2088–2113 (1989). doi:10.1121/1.397861
Nearey, T.M.: Vowel inherent spectral change in the vowels of North American English. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel inherent spectral change, Chap. 4. Springer, Heidelberg (2013)
Nittrouer, S.: Dynamic spectral structure specifies vowels for children and adults. J. Acoust. Soc. Am. **122**, 2328–2339 (2007). doi:10.1121/1.2769624
Peterson, G.E., Barney, H.L.: Control methods used in a study of vowels. J. Acoust. Soc. Am. **24**, 175–184 (1952). doi:10.1121/1.1906875
Potter, R.K., Kopp, G., Green, H.: Visible speech. Van Nostrand-Reinhold, New York (1947)
Rogers, C.L., Glasbrenner, M.M., DeMasi, T.M., Bianchi, M.: Vowel-inherent spectral change and the second-language learner. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change, Chap. 10. Springer, Heidelberg, (2013)
Shankweiler, D.P., Strange, W., Verbrugge, R.R.: Speech and the problem of perceptual constancy. In: Shaw, R.E., Bransford, J. (eds.) Perceiving, Acting and Knowing: Toward an Ecological Psychology, pp. 315–345. Erlbaum Associates, Hillsdale (1977)
Stack, J.W., Strange, W., Jenkins, J.J., Clarke III, W.D., Trent, S.A.: Perceptual invariance of coarticulated vowels over variations in speaking rate. J. Acoust. Soc. Am. **119**, 2394–2405 (2006). doi:10.1121/1.2171837

Steinlen, A.K.: The influence of consonantal context on native and non-native vowel production: A cross-language study. Tübingen, TübingenGunter Narr (2005)
Stevens, K.N., House, A.S.: Perturbation of vowel articulations by consonantal context: An acoustical study. J. Speech Hear. Res. **6**, 111–128 (1963)
Strange, W.: Evolving theories of vowel perception. J. Acoust. Soc. Am. **85**, 2081–2087 (1989a). doi:10.1121/1.397860
Strange, W.: Dynamic specification of coarticulated vowels spoken in sentence context. J. Acoust. Soc. Am. **85**, 2135–2153 (1989b). doi:10.1121/1.397863
Strange, W.: Cross-language phonetic similarity of vowels: theoretical and methodological issues. In: Bohn, O.-S., Munro, M.J. (eds.) Language experience in second language learning: In honor of James Emil Flege, pp. 35–55. John Benjamins, Philadelphia (2007)
Strange, W., Bohn, O.-S.: Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies. J. Acoust. Soc. Am. **104**, 488–504 (1998). doi:10.1121/1.423299
Strange, W., Edman, T.R., Jenkins, J.J.: Acoustic and phonological factors in vowel identification. J. Exp. Psychol. Hum. Percept. Perform. **5**, 643–656 (1979). doi:10.1037/0096-1523.5.4.643
Strange, W., Gottfried, T.L.: Task variables in the study of vowel perception. J. Acoust. Soc. Am. **68**, 1622–1625 (1980). doi:10.1121/1.385217
Strange, W., Jenkins, J.J., Johnson, T.L.: Dynamic specification of coarticulated vowels. J. Acoust. Soc. Am. **74**, 695–705 (1983). doi:10.1121/1.389855
Strange, W., Verbrugge, R.R., Shankweiler, D.P., Edman, T.R.: Consonant environment specifies vowel identity. J. Acoust. Soc. Am. **60**, 213–224 (1976). doi:10.1121/1.381066
Strange, W., Weber, A., Levy, E.S., Shafiro, V., Hisagi, M., Nishi, K.: Acoustic variability within and across German, French and American English vowels: Phonetic context effects. J. Acoust. Soc. Am. **122**, 1111–1129 (2007). doi:10.1121/1.2749716
Tranel, B.: The Sounds of French. Cambridge University Press, New York (1987)
Verbrugge, R.R., Rakerd, B.: Evidence of talker-independent information for vowels. Lang. Speech **29**, 39–57 (1986). doi:10.1177/002383098602900105
Verbrugge, R.R., Shankweiler, D.P.: Prosodic information for vowel identity. J Acoust. Soc. Am. **61** (1), S39. (1977) doi:10.1121/1.2015621
Verbrugge, R.R., Strange, W., Shankweiler, D.P., Edman, T.R.: What information enables a listener to map a talker's vowel space? J. Acoust. Soc. Am. **60**, 198–212 (1976). doi:10.1121/1.381065
Von Son, R.J.J.H., Pols, L.C.W.: Formant frequencies of Dutch vowels in a text, read at normal and fast rate. J. Acoust. Soc. Am. **88**, 1683–1693 (1990). doi:10.1121/1.400243
Von Son, R.J.J.H., Pols, L.C.W.: Formant movements of Dutch vowels in a text, read at normal and fast rate. J. Acoust. Soc. Am. **92**, 121–127 (1992). doi:10.1121/1.404277

Perception of Vowel Sounds Within a Biologically Realistic Model of Efficient Coding

Keith R. Kluender, Christian E. Stilp and Michael Kiefte

Abstract Predicated upon principles of information theory, efficient coding has proven valuable for understanding visual perception. Here, we illustrate how efficient coding provides a powerful explanatory framework for understanding speech perception. This framework dissolves debates about objects of perception, instead focusing on the objective of perception: optimizing information transmission between the environment and perceivers. A simple measure of physiologically significant information is shown to predict intelligibility of variable-rate speech and discriminability of vowel sounds. Reliable covariance between acoustic attributes in complex sounds, both speech and nonspeech, is demonstrated to be amply available in natural sounds and efficiently coded by listeners. An efficient coding framework provides a productive approach to answer questions concerning perception of vowel sounds (including vowel inherent spectral change), perception of speech, and perception most broadly.

Abbreviations

C	Consonant
CV	Consonant–vowel
CVC	Consonant–vowel-consonant

K. R. Kluender (✉)
Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN, USA
e-mail: kkluender@purdue.edu

C. E. Stilp
Department of Psychological and Brain Sciences, University of Louisville, Louisville, KY, USA

M. Kiefte
School of Human Communication Disorders, Dalhousie University, Halifax, NS, Canada

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_6,

CSE	Cochlea-scaled spectral entropy
ERB	Equivalent rectangular bandwidth
f0	Fundamental frequency
F1	First formant
F2	Second formant
F3	Third formant
JND	Just noticeable difference
PCA	Principal component analysis
r	Pearson product-moment correlation coefficient
TIMIT	Texas Instruments/Massachusetts Institute of Technology
V	Vowel
VC	Vowel-consonant
VISC	Vowel inherent spectral change

1 Introduction

In this contribution, we offer a perspective on perception of speech, and of vowel sounds in particular, that is motivated by broad principles that apply to sensation and perception across all modalities. Our approach, referred to as 'efficient coding' in models of visual perception, is situated within a half-century tradition of information-theoretic approaches that remain fruitful today (e.g., Attneave 1954; Barlow 1961; Fairhall et al. 2001; Simoncelli 2003; Clifford et al. 2007). This information-theoretic perspective dissolves some longstanding debates and distinctions while revealing new insights into processes of speech perception that have been neglected or obscured by more traditional approaches to auditory perception.

The organization of this chapter begins with consideration of speech perception within a broad perspective on perception spanning all modalities. Within this framework, vowel perception can provide both examples of efficient coding and tests of the power of such an approach.

2 Objects of Perception

An enduring distraction for investigators studying speech perception has concerned the extent to which objects of speech perception are defined as articulatory gestures or patterns (e.g., Liberman and Mattingly 1985; Fowler 1986), acoustic patterns, patterns of sensory stimulation (e.g., Diehl and Kluender 1989), or some combination (e.g., Stevens and Blumstein 1981; Nearey 1997). Oftentimes, controversies concerning appropriate objects of perception generated more heat than light. We suggest that debates concerning objects of perception cannot be resolved because the question itself is ill-posed, if not outright misleading.

Unconvinced by arguments exclusively for gestures or acoustics as proper objects of speech perception, Nearey (1997) proposed what he described as a "double-weak" model of speech perception—a judicious combination of both articulatory and acoustic/auditory characterizations of speech perception. In the present approach founded upon information-theoretic principles, we go one step further. We make the claim that there are no *objects* of perception, neither for speech nor for perception in general. Instead, there is an *objective* for perception, which is to maintain adequate agreement between an organism and its world in order to facilitate adaptive behavior.

Within this functional framework, perceptual success does not require recovery or representations of the world *per se*. Perceivers' subjective impressions may be of objects and events in the world, and the study of perceptual processes benefits from inspection of real-world objects and events, patterns of light, sound pressure waves, transduction properties, and neural responses. By and large, however, viewing perception with a focus on either distal or proximal properties falls short of capturing the essential functional characteristic of perception—the relationship between an organism's world and its behavior.

If there are no objects *of* perception, how should one think about information *for* perception? Information for perception does not exist in the objects and events in the world, nor does it exist in the brain of the perceiver. Instead, information exists in the relationship between an organism and its world. It may be useful to consider the contrast between information *about* and information *for*. When one discusses objects of perception, it is information *about* that is typically inferred. Implicit in such efforts is the notion that one needs to solve the inverse problem; how does one satisfactorily represent the world within one's brain? By contrast, if the objective of a successful perceptual system is to maintain adequate agreement between an organism and its world in order to facilitate adaptive behavior, then information *for* successful perception is nothing more or less than information that resides in this relationship (or agreement).

3 Shannon Information Theory

This way of viewing information is consistent with one of the fundamental characteristics of Shannon information theory (Shannon 1948; Weiner 1948). Some readers may be familiar with Fletcher's pioneering applications of information theory to speech (Fletcher 1953/1995). However, our application here will be more akin to early approaches of vision researchers such as Attneave (1954, 1959) and Barlow (1961), as these efforts continue to be productive in contemporary theories of 'efficient coding' (e.g., Barlow 1997, 2001; Simoncelli and Olshausen 2001; Schwartz and Simoncelli 2001; Simoncelli 2003; Stilp et al. 2010c). Developed at Bell Laboratories for practical application to telephone bandwidth, one fundamental premise of Shannon's information theory is that information exists only in the relationship between transmitters and receivers.

Information does not exist in either *per se*, and information does not portray any essential characteristics about either transmitters or receivers. In the same fashion, perceptual information exists in the *relationship* between organisms and their environments.

Information is transmitted when uncertainty is reduced and agreement is achieved between receivers and transmitters, or in the case of perception, between organisms and their world. Within a sea of alternative perceptual endpoints, agreement between the organism and environment is functionally successful to the extent that the organism arrives at the alternative that gives rise to adaptive behavior. The greater the number of alternatives there are (uncertainty, unpredictability, variability, or entropy), the greater the amount of information that potentially can be transmitted. There is no information when there is no variability. When there is no variability, there is total predictability and hence, no information.

Given these facts about information, it is true and fortunate that sensorineural systems respond only to change relative to what is stationary or predictable (Kluender et al. 2003). Perceptual systems do not record absolute levels whether loudness, pitch, brightness, or color. Relative change is the coin of the realm for perception, a fact known at least since Ernst Weber in the mid-18th century. Sacrifice of absolute encoding has enormous benefits along the way to optimizing information transmission. For example, biological transducers have impressive dynamic range given their evolution via borrowed parts (e.g., gill arches to middle ear bones); however, this dynamic range is always dwarfed by the physical range of absolute levels available from the environment. The beauty of sensory systems is that, by responding to relative change, a limited dynamic range shifts to optimize the amount of change that can be detected in the environment at a given moment. There are increasingly sophisticated mechanisms supporting sensitivity to change with ascending levels of processing, and several will be discussed in this chapter.

Relative change, of course, requires context from which to change. Context itself is relatively uninformative; it is what already exists or can be predicted. Context can be very brief—the present or immediate past from which change arises. Context can be extended, such as predictable characteristics of listening conditions, such as acoustics of concert halls or across trials in an experimental session. Context can be measured in milliseconds, minutes, months, or even a lifetime of experience with predictable properties of a structured world. In all cases, perceptual systems are more efficient to the extent that predictable elements of context are registered in ways that enhance sensitivity to that which is less predictable and more informative.

By adopting this way of viewing context and information for perception more generally, traditional distinctions between sensation, perception, and learning diffuse along a series of processes that operate over broader ranges of time and experience. From peripheral sensory transduction through cortical organization consequent to experience, a series of successively more sophisticated processes extract predictability to make unpredictable (informative) changes easier to detect.

4 Potential Information and Intelligibility

We begin by considering the importance of sensory change at the lowest levels of the auditory system. Stilp and Kluender (2010) recently evaluated the extent to which measures of sensory change, tailored by the cochlea, may serve to explain intelligibility of connected speech. Earlier efforts employed orthodox descriptions of speech signals (strings of consonant and vowel sounds) in studies conducted to evaluate relative contributions of vowel versus consonant portions of the speech stream. When intervals corresponding to consonants or vowels were replaced with noise, vowels appeared to provide more information (contribute more to sentence intelligibility) than consonant sounds (Cole et al. 1996; Kewley-Port et al. 2007; Fogerty and Kewley-Port 2009, but see Owren and Cardillo 2006). Because listeners were better at understanding sentences with consonants replaced than with vowels replaced, a "vowel superiority effect" had been suggested to exist.

Interpreting putative vowel superiority effects is not straightforward. Experiments in which consonants or vowels are replaced by noise rely upon operationally defined temporal boundaries provided by phoneticians to demarcate consonants and vowels in the Texas Instruments/Massachusetts Institute of Technology (TIMIT) database of sentences (Garofolo et al. 1990). While neither the phoneticians who provide these boundaries nor the investigators who use them in experiments are naïve about shortcomings of such demarcations, the fact that no single point in time exists when a consonant ends and a vowel begins in connected speech presents a challenge to interpreting data suggesting vowel superiority. Production of every consonant and vowel overlaps in time (and in the vocal tract) with speech sounds that precede and follow. In addition, acoustic characteristics within vowels provide rich evidence concerning neighboring consonants and *vice versa* (e.g., Liberman et al. 1957; Jenkins et al. 1983; Sussman et al. 1991; Kiefte 2000).

A second complication arises from the fact that, at least as delineated in TIMIT, vowel intervals of sentence waveforms are roughly one-third longer than those for consonants. Recently, Lee and Kewley-Port (2009) provided evidence that there may be no intelligibility differences between replacing different combinations of consonant (C) and vowel (V) portions with noise after accounting for duration. Stilp and Kluender (2010) also found that replacement of Vs, CVs, and VCs resulted in closely similar declines in performance, although, they found that relatively greater proportions of Cs could be replaced, relative to other conditions, for performance to decline.

It seems unlikely that most speech intervals are perceptually equivalent in terms of information, and that only duration—not acoustic composition—matters. Stilp and Kluender (2010) tested whether a metric other than consonants versus vowels could better account for intelligibility data. They began by abandoning phonetic designations. Instead of considering consonants, vowels, or their combinations, they tested the degree to which speech intelligibility depends upon amount of potential information, defined psychoacoustically, that is removed and replaced by noise. The simple fundamental principle that perceptual systems respond primarily to change

has the formal consequence of enhancing information transmission. Because there is no new information when events either do not change or are predictable, one can employ relative change as an approximate measure of potential information. Cochlea-scaled spectral entropy (CSE) is a measure of relative spectral change across time, operationalized as the extent to which successive spectral slices differ (i.e., cannot be predicted) from preceding spectral slices.

CSE is quantified as Euclidean distances between adjacent psychoacoustically-scaled spectral slices (Fig. 1). Sentences were RMS-intensity-normalized and divided into 16 ms frames independent of TIMIT segmentations. Slices were passed through 33 filters that capture nonlinear weighting and frequency distribution along the cochlea (Patterson et al. 1982). Filters were spaced one equivalent rectangular bandwidth (ERB) apart up to 8 kHz. ERBs provide a close approximation to tonotopic distribution along the cochlea (Greenwood 1990) and psychophysical auditory filters derived from normal-hearing listeners (Glasberg and Moore 1990). Consequently, each filter corresponds to an equivalent number of inner hair cells and neurons in the auditory nerve.

The ERB scale is roughly logarithmic except in low frequencies where it is more nearly linear. Euclidean distances between adjacent 16 ms slices were calculated across the 33 filter output levels. Distances were then summed in boxcars of either five (80 ms, approximate mean consonant duration) or seven successive

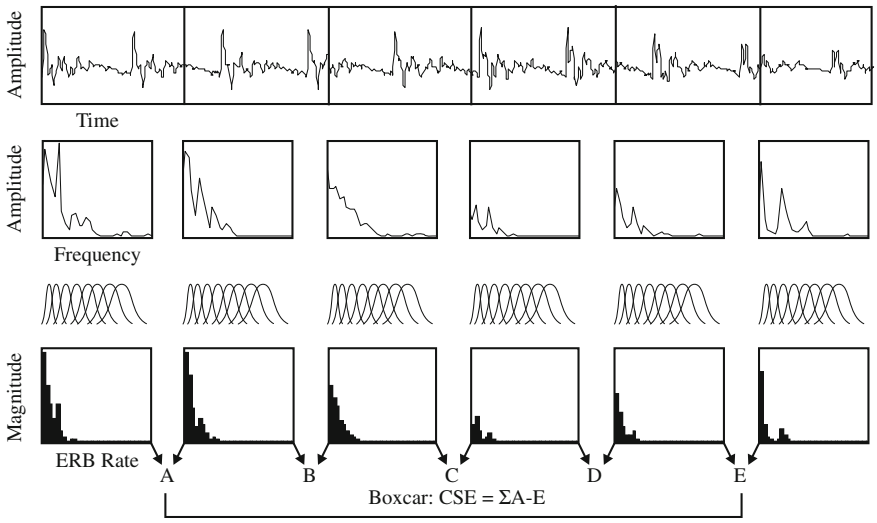


Fig. 1 Calculation of CSE using an 80 ms boxcar. *First row* Original speech waveform, with amplitude plotted as a function of time. Each box corresponds to a 16 ms slice. *Second row* Fourier transform of each 16 ms slice, with amplitude plotted as a function of frequency. *Third row* Auditory filter bank. *Fourth row* Weighted outputs of auditory filter processing with magnitude plotted as a function of ERB rate. Euclidean distances are calculated between each slice, then summed in boxcars

slices (112 ms, approximate mean vowel duration). Cumulative Euclidean distances within a boxcar were taken as measures of spectral entropy and served as a psychoacoustic metric of potential information.

After convolving boxcars of summed distances across entire sentences, entropy measures were sorted into ascending order (Low Entropy condition), descending order (High Entropy), or ascending absolute difference from median boxcar value (Medium Entropy). The boxcar ranked first (lowest, highest, or median CSE) was replaced by speech-shaped noise matched to mean sentence level. Eighty milliseconds before and after selected boxcars were preserved to avoid boxcars overlapping, and the first 80 ms of every sentence was always left intact. The procedure continued iteratively to the next-highest-ranked boxcar, which was replaced only if its content had not already been replaced or preserved.

As expected, replacing longer (112 ms) segments impaired performance more on average than replacing shorter (80 ms) segments. Amount of potential information (CSE), however, plays a much larger role in predicting performance (Fig. 2). Intelligibility closely follows measures of entropy replaced with noise ($r^2 = 0.80$, $p < 0.01$).

5 Auditory Entropy and the Sonority Hierarchy

There are informative systematicities in the types of speech sounds that have higher and lower CSE (Stilp and Kluender 2010). More consonants than vowels are replaced in low-entropy conditions, and more vowels than consonants are replaced in high-entropy conditions. Although significantly more vowels were replaced with each increase in CSE, proportion of vowels or consonants replaced are not significant predictors of intelligibility ($r^2 = 0.55$, n.s.).

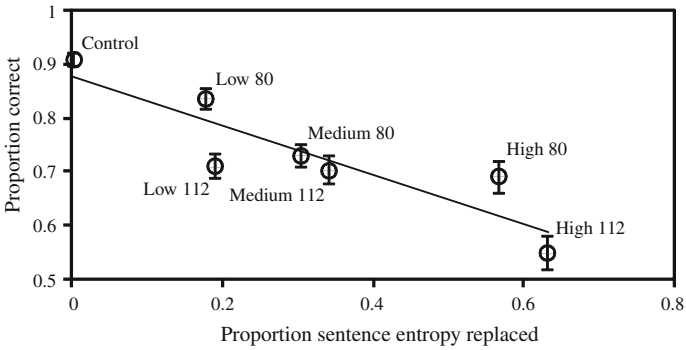


Fig. 2 Results of Stilp and Kluender (2010). Sentence intelligibility (proportion of words correctly identified) is plotted as a function of proportion of sentence entropy (CSE) replaced by noise. Data points are labeled by level of CSE replaced by noise (*Low*, *Medium*, *High*) and boxcar duration (80, 112 ms). Potential information is a significant predictor of sentence intelligibility ($r^2 = 0.80$, $p < 0.01$). *Error bars* represent standard error of the mean

Phonetic compositions of replaced speech segments were analyzed according to vocal tract configuration (vowels) or manner of articulation (consonants). Vowels were subdivided into high (close) versus low (open), front versus central or back, and diphthongs. As one would expect, acoustic segments traditionally labeled as diphthongs were replaced most often in high-entropy conditions and least often in low-entropy conditions. Low vowels are replaced most often in high-entropy conditions and least often in low-entropy conditions, while high vowels show the opposite pattern. There were no systematic differences between front, central, or back vowels.

Consonants were classified as plosives, affricates, closure silence (preceding plosives and affricates), fricatives, laterals/glides, and nasals. Consonants replaced by noise differed substantially across conditions. The most vowel-like consonants—nasals and laterals/glides—were most-often replaced in high-entropy conditions and least-often replaced in low-entropy conditions. The least-vowel-like plosives and affricates showed the complementary pattern. Only modest changes were observed for fricatives and closures across conditions.

This pattern of CSE decreasing from non-high vowels, to high vowels, to laterals/glides and nasals, to fricatives, to affricates and finally plosives closely parallels the sonority hierarchy. The linguistic construct of sonority (or vowel-likeness) is useful for describing phonological systematicity, especially syllable composition; however, to date sonority has been resistant to clear definition in acoustics or articulation (Ohala 1990; Clements 2009).

As an explicit test of the extent to which CSE corresponds to sonority, CSE was measured in VCV recordings from six adult talkers (three female, three male). Twenty American English consonant sounds (/b, d, g, p, t, k, m, n, l, r, w, f, v, θ, ð, s, z, ʃ, tʃ, dʒ/) were flanked by each of three vowels (/ɑ, i, u/), generating 60 VCVs per talker. Recordings were truncated to maintain consistent overall duration (≈460 ms). Waveforms were normalized so that amplitudes of the vowels reflected differences in naturally spoken vowels, and onsets and offsets of VCVs (64 ms) were excluded from analysis. Averaged across talkers and consonants, there was more CSE in low-vowel contexts (/ɑCɑ/) than in high-vowel contexts (/iCi/, /uCu/; Fig. 3a). Across talkers and all VCVs, laterals and glides had the highest CSE, followed by nasals, fricatives, affricates, and plosives (Fig. 3b). Patterns of CSE for both vowel and consonant analyses follow the sonority hierarchy, corroborating findings from the sentence intelligibility experiment.

Phonetic composition of high- versus low-entropy segments initially may be surprising. One might expect that consonants, created by vocal tract constrictions with correspondingly rapid acoustic changes, would have higher entropy than vowels. However, vowels and vowel-like sounds have greater cochlea-scaled entropy because roughly logarithmic psychoacoustic and physiologic indices weight lower-frequency spectral prominences (formants, F_1 and F_2) more. This explanation depends, however, upon sufficient levels of spectral change in vowel and vowel-like sounds. Multiple contributions to the present volume provide ample testament to the fact that, indeed, vowel formants change substantially over time with few exceptions (Nearey and Assmann 1986), and this vowel inherent

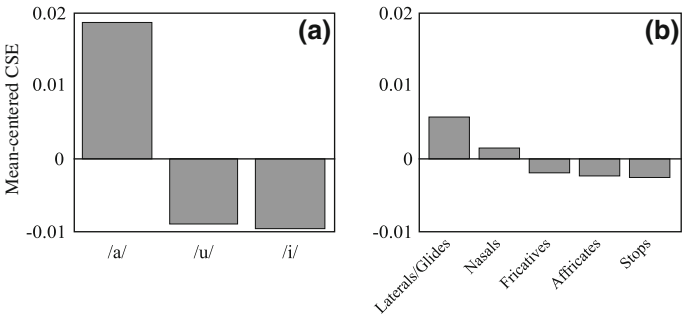


Fig. 3 CSE analyses of VCVs in Stilp and Kluender (2010) corroborating the relationship between CSE and sonority. **a** Measures of vowel CSE averaged across consonants. **b** Measures of consonant CSE averaged across vowels

spectral change (VISC) is important to listeners (Hillenbrand and Nearey 1999; Bunton and Story 2010).

Similarly, Kiefte and Kluender (2005) found that perception of synthetic monophthongs, relatively poor renditions of real speech, was unrepresentative of normative speech perception. They showed how listeners could rely on putatively unreliable gross acoustic properties to identify monophthongs. Under some conditions, listeners identified monophthongs on the basis of spectral tilt, which is highly vulnerable to transmission channel characteristics such as room acoustics (Klatt 1982). In such cases, listeners responded to stimuli in a manner similar to hearing-impaired listeners (Lindholm et al. 1988; Alexander and Kluender 2008). By contrast, Kiefte and Kluender (2005) found that listeners do not use tilt when provided the opportunity to use changing spectral composition, as formant transitions may serve to enhance perceptibility of formants themselves.

Some correspondence between CSE and VISC is expected, but there are important differences between the two measures. Because CSE is an accumulation of brief slice-to-slice changes across the full cochlea-scaled spectrum, it is not expected to map perfectly onto simpler characterizations of F_1 and F_2 center frequencies at only two points in time. Nevertheless, vowels displaying inherent spectral change (diphthongs and low but not high vowels) displayed the highest measures of CSE in Stilp and Kluender (2010), so the metric appears to be sensitive to some aspects of VISC. To evaluate the relationship between CSE and VISC, we measured cumulative CSE across all vowels spoken by men, women, and child talkers in the database (Hillenbrand et al. 1995) used by Hillenbrand and Nearey (1999). As an index of VISC, we calculated vector lengths in the ERB-scaled F_1–F_2 plane for each English vowel (square root of the sum of squared changes in F_1 and F_2 frequencies between 20 and 80 % time points) using formant center frequencies reported by Hillenbrand et al. (1995). Longer vector lengths correspond to more VISC between these two time points. Cumulative CSE was calculated over the same (20–80 %) intervals. Vector length and CSE share a significant correlation ($r = 0.45$; $p < 0.01$). Given this relationship between VISC

and CSE, and the extent to which CSE accounts for sentence intelligibility (Stilp and Kluender 2010), VISC may be important beyond characterization of vowels *per se*.

6 VISC as Adaptive Dispersion

Our experiments employing noise replacement highlighted the importance not only of vowels, but of changes in spectral composition of vowels for sentence understanding. The more vowel-like a segment of the signal was, the more essential that segment was for intelligibility. This finding explicitly relies on the fact that vowel spectra, and their cochlear consequences, change across time. Further inspection of the nature of these vowel inherent spectral changes reveals that, in addition to increasing psychoacoustic potential information, VISC may serve to make it easier for listeners to detect differences between vowel sounds.

Different languages use different inventories of vowel sounds, and languages use subsets of vowels that are most easily discriminated from one another. For example, those vowels favored by languages with five vowels are ones that are as acoustically distinct as possible from one another. As a general rule, the set of vowels selected by a language, whether it uses three or ten vowels, is comprised of sounds that tend toward maximal distinctiveness (Liljencrants and Lindblom 1972; Bladon and Lindblom 1981). Lindblom (1986) refers to this fact as 'adaptive dispersion.' However, these demonstrations suffer from considering vowel sounds only as relatively static entities by employing steady state formants. To what extent are realistic portrayals of vowels, possessing inherent spectral change, consistent with adaptive dispersion?

To address this question, we begin by plotting formant measurements in cochleotopic (ERB) coordinates for average F_1 and F_2 at 20 % time points drawn from the Hillenbrand et al. (1995) database for western Michigan male talkers (Fig. 4, left). We make a rough division of these F_1/F_2 points into front/back and high/low regions of the vowel space. Next, we capture VISC by plotting directional vectors (length proportional to spectral change) with the origin being 20 % measurements in F_1/F_2 (Fig. 4, right). In general, formants with starting points closer to one another (in the same F_1/F_2 quadrant, left) tend to disperse such that the direction of spectral change is distinct from that for other vowels with neighboring starting points. This is most obviously true for closely neighboring vowels /ɪ/ and /e/, /ɛ/ and /æ/, and /o/ and /ʊ/. Kinematic dispersion is less substantial in this quadrant-based analysis for /Λ/ versus neighboring /ɔ/. Neither point vowel /i/ nor /u/ exhibits much VISC; however, they are quite distant from each other and have no close neighbors (Fig. 4, left). Finally, unlike the relatively stationary /i/ and /u/, spectral composition of the third point vowel /ɑ/ moves toward the center of the vowel space in a way not predicted by dispersion, at least not for F_1 and F_2 alone.

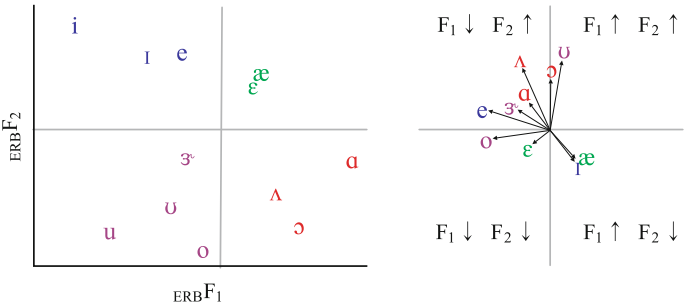


Fig. 4 (*Left*) Mean F_1/F_2 values of vowels spoken by male talkers in Hillenbrand and Nearey (1999). Phonetic symbols are placed at formant values measured at 20 % of overall vowel duration. *Grey lines* represent coarse distinctions between *high/low* and *front/back*. (*Right*) VISC measures in the same vowels. *Arrow tails* are formant measurements at 20 % of overall duration, and are aligned at the origin of the graph. *Arrow heads* are formant measurements at 80 % of overall duration, and *arrow length* is proportional to VISC. Quadrants are labeled according to F_1 and F_2 change over time. Vowels /i/ and /u/ are not included as both display minimal VISC and would lie very close to the origin

Perceptual data support the importance of VISC for distinguishing vowel sounds from one another. For example, Nearey and Assmann (1986) presented listeners three combinations of vowel fragments separated by 10 ms of silence. These were: *natural* (segment of vowel 'nucleus' followed by offglide); *repeated* (nucleus followed by itself); and, *reverse* (offglide followed by nucleus). Most telling are cases for which kinematics are reversed. Listeners were more likely to identify reversed /ɪ/ as /e/ and reversed /e/ as /ɪ/, and reversed /ɛ/, /æ/, /o/, and /ʊ/ were among the most likely vowels to be confused with others. When Hillenbrand and Nearey (1999) synthesized vowels with either natural kinematics (VISC) or with flat formants fixed at the frequencies of the least kinematic (quasi 'steady-state') portion of vowel sounds, the vowel sounds most likely to be misidentified were /e/, /o/, /ʊ/, /u/, and /Λ/. Vowels synthesized based upon /e/ were most likely to be confused with /ɪ/ and /ɛ/, /o/ with /ɔ/, /ʊ/ with /Λ/ and /o/, /u/ with /o/, and /Λ/ and /ʊ/. Most recently, Bunton and Story (2010) synthesized vowels using both static and time-varying speaker-dependent area functions based on magnetic resonance imaging measurements of the vocal tract (see also Story and Bunton 2012 this volume). Performance was much more consistent for vowel sounds created with time-varying spectra, a pattern of data very similar to that found by Hillenbrand and Nearey (1999). Static vowels synthesized based upon /ɪ/ were most likely to be confused with /e/ and /ɛ/, /e/ with /ɛ/ and /ɪ/, /ɛ/ with /æ/, /Λ/ with /ʊ/ and /ɑ/, /ɑ/ with /ɔ/, /ɔ/ with /ɑ/ and /o/, /o/ with /ʊ/, and /ʊ/ with /u/ and /o/.

Taken together, VISC serves to make vowels more perceptually distinctive in at least two ways. First, spectral change in itself capitalizes on a fundamental design aspect of sensorineural systems, this being near exclusive sensitivity to change. As demonstrated earlier, the importance of spectral change in vowel sounds is all the more remarkable when one captures this change with respect to cochlear

organization. Second, vowel inherent spectral change is not change for the sake of change itself, however important change is to sensorineural processing. Instead, VISC is by and large sensible when one conceptualizes the most important property of vowel sounds to be the extent to which distinctions between vowel sounds are enhanced. Vowel inherent spectral change enhances distinctiveness by providing an additional means of distinguishing vowels from one another. In most instances, when vowels begin relatively close in perceptual space, they quickly proceed to increase 'otherness' in trajectories of spectral shape change.

7 Information and Rate 'Normalization'

When one adopts relative change as a metric for perceptual information, some other traditional measures, beyond linguistic constructs such as consonants and vowels, are left behind. For example, relative change in the cochlea-scaled spectrum does more than simply morph the frequency axis. When relative change is employed, units corresponding to absolute frequency and intensity are discarded. To the extent that change is a fundamentally unitless measure, absolute time also ought not matter within limits. In Stilp and Kluender's (2010) experiment above, absolute signal duration replaced by noise did not explain performance. To the extent that relative change is the most useful measure of potential information for perception, one should be able to warp time (slower or faster) and intelligibility should be predicted on the basis of amount of relative change more or less independent of the time course over which that information is accrued.

Stilp et al. (2010b) conducted a series of experiments in which measures of relative change were used to predict sentence intelligibility across wide variation in rate of speech. They synthesized 115 seven-syllable sentences from the Hearing In Noise Test (HINT; Nilsson et al. 1994) at three different speaking rates: slow, medium, and fast (2.5, 5.0, and 10.0 syllables per second, respectively). Next, they time-reversed equal-duration segments (20, 40, 80, and 160 ms) at the nearest zero crossings for every sentence (see e.g., Saberi and Perrott 1999). As seen in Fig. 5 (right), listener performance across conditions is very well predicted based upon proportion of the utterance distorted (reversal duration) and not absolute duration (left).

CSE functions peak at roughly two-thirds of mean syllable duration (64 ms for fast sentences, 128 ms for medium, and 256 ms for slow) reflecting the fact that acoustic realizations of consonant and vowel sounds are largely conditioned by preceding vowels or consonants until they begin to assimilate to the next speech sound (Fig. 6, right). For English VCVs, the identity of the second vowel is largely independent of the first vowel, and identities of vowels in successive syllables are also largely independent. Consequently, beyond these relative maxima, distances regress toward the mean Euclidean distance of any spectral sample to the long-term spectrum of speech from the same talker. This simple, limited measure of

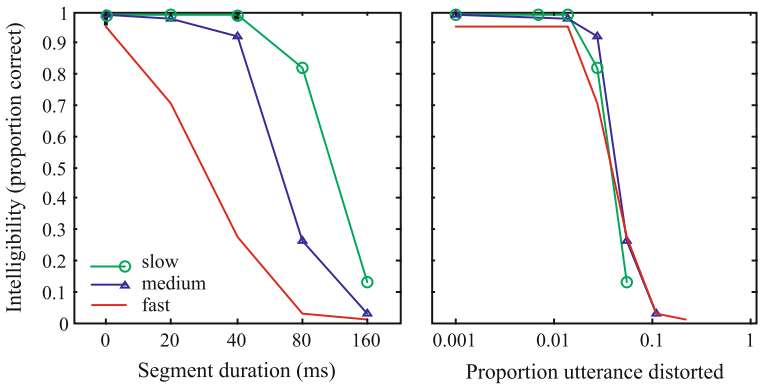


Fig. 5 Results from Stilp et al. (2010b). (*Left*) Intelligibility of sentences at a wide range of speaking rates when fixed-duration segments were temporally reversed. Performance declines *faster* (*slower*) for sentences at *faster* (*slower*) rates relative to *medium*-rate speech. (*Right*) Data converge to a common function when plotted using the relative measure of proportion of utterance distorted (segment duration divided by mean sentence duration)

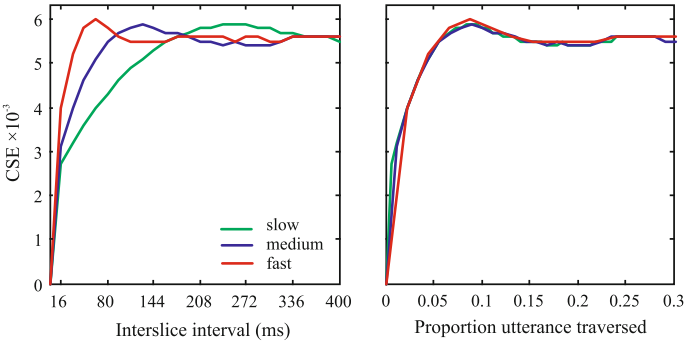


Fig. 6 CSE analyses of variable-rate sentences from Stilp et al. (2010b). (*Left*) CSE measured in 16 ms slices with increasing interslice intervals. *Fast*-rate speech peaks first (64 ms), followed by *medium* (128 ms), then *slow* (256 ms). Distances regress to the mean spectral distance between any two slices spoken by the same talker. (*Right*) Like behavioral data, CSE functions converge when plotted using the relative measure of proportion utterance traversed (interslice interval divided by mean sentence duration). All functions peak at approximately two-thirds of mean syllable duration

information conveyed by spectral change accounts for a substantial proportion of variance in listener performance across all rate conditions ($r^2 = 0.89$; $p < 0.001$).

We now see that another attractive property of CSE is that it requires no explicit rate normalization. This measure of potential information naturally accommodates variable-rate speech materials. There have been substantial efforts to better understand how listeners normalize across speaking rate when identifying individual consonants (e.g., Miller and Liberman 1979; Miller 1981), vowels (e.g., Ainsworth 1972, 1974; Gottfried et al. 1990), or words (e.g., Miller and Dexter

1988), and all of these efforts have concentrated upon absolute physical changes in frequency and time. To the extent that potential information, not time or frequency *per se*, accounts for perception, concerns about normalization of time or frequency toward some iconic standard dissolve. While durations and frequencies may vary, potential information remains relatively constant and requires no such normalization.

8 When Acoustics are Predictable

To this point, discussion has concerned relatively local effects of context, as we have addressed changes in frequency on the order of hertz and time over the course of milliseconds. Over the past decade, multiple studies have been conducted to investigate processes by which the auditory system compensates for reliable spectral and temporal characteristics of a sound source under different listening conditions. Here, attention is drawn to the full spectrum across much longer intervals. It has become quite clear that the auditory system calibrates to different listening environments by perceptually compensating for or absorbing reliable, redundant spectral features of the acoustic input (Watkins 1991; Watkins and Makin 1994; Kiefte and Kluender 2008; Alexander and Kluender 2010). The advantage of such processes is that they optimize sensitivity to new (unpredictable) information.

Kiefte and Kluender (2008) conducted experiments designed to assess relative contributions of spectrally global (spectral tilt) versus local (spectral peak) characteristics of a listening context. They varied both spectral tilt and center frequency of F_2 to generate a matrix of steady-state vowel sounds that perceptually varied from /u/ to /i/. Listeners identified these sounds following filtered natural or time-reversed precursor sentences. When either set of precursor sentences was filtered to share the same long-term spectral tilt as the target vowel, tilt information was neglected and listeners identified vowels principally on the basis of F_2. Conversely, when precursors were filtered with a single pole centered at the F_2 frequency of the target vowel, perception instead relied upon tilt. These results demonstrate calibration to reliable global and local spectral features across both intelligible and unintelligible speech-like contexts, as listeners predominantly used the less predictable spectral property that was not consistent within the precursor.

Stilp and colleagues (Stilp et al. 2010a) demonstrated that such compensation appears to be relatively automatic and naïve to the source of sounds. They found that listeners compensate for reliable spectral properties of a listening context when identifying musical instruments. Perception of tenor saxophone versus French horn adjusts following a filtered passage of speech or of other musical instruments (Schubert string quintet). This adjustment for spectral characteristics of a listening context is closely analogous to visual color constancy through which colors appear relative constant despite widely varying reflectances under different types of illumination (Stilp et al. 2010a).

9 Efficient Coding of Second-Order Statistics

It is clear that the auditory system absorbs predictability within a listening context, and the result is increased sensitivity to information which, by definition, is something that is not predictable. Much predictability in sounds is not about listening contexts, but instead is inherent in the ways sounds are structured. Natural sounds are complex and are typically changing along multiple acoustic dimensions. For sounds created by real structures including musical instruments and vocal tracts, changes in different acoustic dimensions cohere in accordance with physical laws governing sound-producing sources. For example, articulatory maneuvers that produce consonant and vowel sounds give rise to multiple acoustic attributes. This redundancy across attributes contributes to robust speech perception despite substantial signal degradation (Assmann and Summerfield 2004; Kluender and Kiefte 2006; Kluender and Alexander 2007). For example, because the relationship between formant frequency and amplitude is determined entirely by the physics of speech production (Fant 1956; Stevens 1998), it is to be expected that listeners are relatively insensitive to variations in one property when the other provides more reliable information regarding vowel identity. In support of this, Kiefte et al. (2010) have shown that listeners are indeed insensitive to variations in formant amplitude—a relatively unstable acoustic property across different listening contexts (e.g., Kiefte and Kluender 2005, 2008). They further show that any evidence for an important role for formant amplitude in vowel perception (e.g., Aaltonen 1985; Hedrick and Nábělek 2004) can be explained entirely by peripheral auditory effects such as simultaneous masking.

Such redundancy in speech epitomizes the general fact that objects and events in the world have structure. Attneave (1954) emphasizes how information received by the visual system is redundant because sensory events are highly interdependent in both space and time. This is simply because "the world as we know it is lawful" (p. 183). Adopters of information theory as an explanatory construct for human perception quickly came to appreciate the significance of capturing predictability among stimulus attributes in the interest of increasing sensitivity to relatively unpredictable changes between signals. Attneave argued "It appears likely that a major function of the perceptual machinery is to strip away some of the redundancy of stimulation, to describe or encode incoming information *in a form more economical* than that in which it impinges on the receptors" (p. 189, emphasis added). Within an emphasis upon neural encoding, Barlow (1959) hypothesized "It is supposed that the sensory messages are submitted to a succession of recoding operations which result in *reduction of redundancy* and increase of relative entropy of the messages which get through" (p. 536, emphasis added). By detecting and exploiting covariance in the environment (predictability), perceptual systems enhance sensitivity to new information (unpredictability, or change).

These principles lie at the heart of contemporary models of efficient coding, and there have been many supporting findings in visual perception. Some studies concern adaptation to images varying in simpler aspects such as color, orientation,

or directional movement, and extend to complex images including faces (see Clifford et al. 2007 for review). In their highly-influential paper concerning efficient coding, Barlow and Földiák (1989) argued that populations of cortical neurons should organize in a way that absorbs correlations across inputs so that perceptual dimensions are more nearly orthogonal (decorrelated) and better able to detect changes in the environment that are not predictable (more informative) based upon prior experience. Barlow and Földiák proposed that absorption of correlations: (1) makes it easier to detect newly appearing associations resulting from new causal factors in the environment; (2) provides a role for recurrent collaterals, which are a conspicuous feature of cortical neurons; and, (3) could account for part of the effects of experience during cortical development. To this list, one might add a simple, perhaps obvious, observation concerning neurons most broadly. Most neurons have many synapses along their dendrites. Whether a neuron fires depends on the joint contributions of many inputs, excitatory and inhibitory, along those dendrites. This simple fact of neural architecture requires that responses depend critically upon correlated activity across synapses. Finally, there is physiological evidence that responses of neurons at successive stages of processing become increasingly independent from one another, and such demonstrations have been clearest in the auditory system (Chechik et al. 2006).

There are multiple recent findings concerning the ways through which the visual perceptual system exploits redundancies among optical attributes. For example, models that capture edge co-occurrence in natural images precisely predict observer performance in a contour grouping task (Geisler et al. 2001). Perhaps the most impressive instances of efficient coding are perceptual "metamers" which are composite stimuli that cannot be discriminated even though individual properties can be discriminated when presented in isolation (Backus 2002). For example, perception of visual slant is cued by both binocular disparity and texture gradient among other cues (Hillis et al. 2002). In nature, these two cues are highly correlated, and when altered in an experimental setting, observers are incapable of discriminating either binocular disparity or texture gradient independent of the other cue.

10 Second-Order Statistics in Speech

It is well-attested that all contrasts between speech sounds are multiply specified. For example, the distinction between medial /b/ and /p/ includes at least sixteen different acoustic differences (Lisker 1978). No single attribute is, in itself, both necessary and sufficient to support perception of /b/ or /p/, which instead relies upon combinations of attributes. Kluender and colleagues (Kluender and Lotto 1999; Kluender and Kiefte 2006; Kluender and Alexander 2007) have argued that one way in which multiple attributes are important to perception is the extent to which they are correlated with one another, and hence, provide redundancies that are central to sensorineural encoding of speech sounds.

Perceptual sensitivity to correlations among stimulus attributes may well account in part for listeners' solution to the lack of invariance across consonantal place of articulation. For example, acoustic information specifying /d/ is dramatically different depending upon the following vowel sound. Perceiving speech despite such variation was once thought to suggest that it was unique among perceptual achievements. However, Kluender et al. (1987) demonstrated that birds could learn the mapping for /d/ versus /b/ and /g/.

In order to account for some of the variability for place of articulation across vowel contexts, Sussman and colleagues (e.g., Sussman et al. 1998) reintroduced the idea of locus equations (Delattre et al. 1955) as part of an explanation for perception of place of articulation. They made exhaustive measurements of thousands of tokens of /b/, /d/, and /g/ produced before multiple vowels by many different talkers, and found that the correlations between onset frequency of F_2 and F_2 frequency of the following vowel efficiently captured differences between /b/, /d/, and the two allophones of /g/ (front and back). Regression lines between F_2 onset and F_2 of the following vowel were distinct between /b/, /d/, and the two allophones of /g/. Correlation coefficients were relatively strong ($r = 0.75 - 0.96$).

Kiefte (2000) similarly measured formant transitions for prevocalic plosives consonants /b/, /d/, and /g/ followed by each of ten Western Canadian English vowels /i, I, e, ɛ, æ, ɑ, Λ, o, ʊ, u/ preceding syllable-final /k/, /ʤ/ and /l/. These CVCs were spoken by five male and six female adults. Correlations between F_2 values measured at the end of the second glottal pulse after onset of periodicity and F_2 60 ms later were $r = 0.98$ for /b/, $r = 0.87$ for /d/, and $r = 0.95$ for /g/. These relational patterns are consistent with Sussman and colleagues' earlier observations, and coefficients are highly similar when ERB-scaled values are used.

We asked whether vowels, like consonants, can be similarly characterized by reliable relationships between F_2 values as a function of time. Using the same data set (Kiefte 2000), we analyzed ERB formant trajectories at the same time points for the ten vowels following consonants /b/, /d/, and /g/. Directly analogous to locus equations for stop consonants, there is remarkable correlation between F_2 values for each vowel across variation in preceding plosive, averaging $r = 0.82$. Of course, owing to the fact that there are more vowels than there are consonantal places of articulation, differences between slopes of regression lines across vowels cannot be as profound as those found for /b/, /d/, and /g/, and the extent to which listeners exploit these correlations in vowel perception is unknown at present. Very recently, Nearey (2012 this volume) demonstrated that, at least for cases tested thus far (Hillenbrand et al. 2001), it is possible to decompose CVC syllables into locus constituents (CV and VC) and kinematic representations of vowels (VISC). Success thus far suggests that useful relational properties for Cs and Vs are separable and available to listeners.

There have been justifiable criticisms of the locus equation concept, perhaps most importantly the fact that other acoustic characteristics contribute to perception of place of articulation (e.g., Blumstein 1998). However, within our proposal

that redundancy between correlated stimulus attributes should be efficiently coded, there is no formal upper bound on the number of attributes that can contribute to the overall covariance structure.

11 Learning Correlated Attributes

Later in this contribution, we will present other examples of naturally occurring covariance relationships between acoustic attributes of vowel sounds as they are created by talkers and their vocal tracts. First, one should ask whether and how listeners detect and exploit redundancy between stimulus attributes. For speech, this process of perceptual organization begins early in life and presumably supports, at least in part, infants' rapid mastery of multiply-specified contrasts within their native language environment. To learn more about acquisition of sensitivity to correlations among stimulus attributes by adult listeners, Stilp et al. (2010c) designed novel complex stimuli that varied across two physically independent acoustic attributes: attack/decay (AD) and spectral shape (SS). SS was varied via summation of two instrument endpoints (French horn, tenor saxophone) in different proportions. In principle, AD and SS are relatively independent both perceptually and in early neural encoding (Caclin et al. 2006). Physically complex attributes were chosen with the expectation that more complex attributes should be more plastic relative to elemental properties such as frequency, which serves as a primitive dimension in the tonotopically organized auditory system.

A stimulus matrix was generated by crossing AD and SS series for which sounds separated by fixed distance in the stimulus space were approximately equally discriminable. Two stimulus subsets were selected from this matrix, each rotated 90° from the other, capturing near-perfect correlations between the two acoustic cues (Fig. 7a: $r_{AD,SS} = \pm 0.97$). Half of forty listeners completed a discrimination task with pairs of stimuli drawn from the distribution shown, and the other 20 listeners discriminated pairs drawn from the rotated subset. Across three blocks of 144 trials within a single experimental session, listeners discriminated sound pairs of three types: (1) consistent with the correlation (Consistent condition; blue); (2) orthogonal to experienced correlation (Orthogonal condition; red); and (3) differing by equal magnitude along only one stimulus dimension (Single-cue condition; green). Test sounds were presented in two-alternative forced choice (AXB) trials using 18 pairs of equally distant stimuli [three steps of AD (one pair), SS (one pair), or AD and SS (15 Consistent plus one Orthogonal pair).]

No feedback was provided. Performance across the three blocks of trials changed in highly informative ways (Fig. 7b). Discrimination of Consistent pairs that respect the correlation began well and remained relatively high throughout the experiment; however, performance on Orthogonal trials was significantly inferior early in testing. Discrimination of stimulus pairs that differed only in a single dimension was near chance in the first block and slowly improved across successive blocks. Perceptual performance rapidly became attuned to the correlation

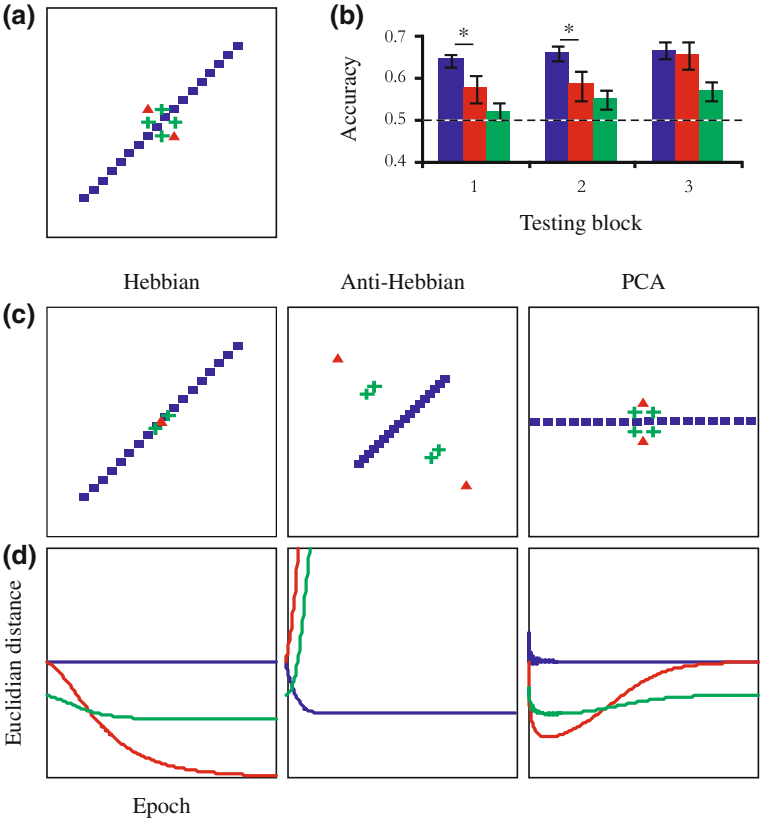


Fig. 7 Stimuli, performance, and modeling results from Stilp et al. (2010c). **a** Listeners discriminated sounds with a strong correlation between AD and SS ($r = \pm 0.97$). Consistent: *blue squares*, Orthogonal: *red triangles*, Single-cue: *green plus signs*. **b** Listener performance. *Dashed line* indicates chance performance, and error bars represent standard error of the mean. Asterisks indicate significant contrasts of interest assessed by paired-sample *t*-tests following Bonferroni correction for multiple comparisons ($p < 0.025$). **c** Model representations of stimuli at the end of simulation. Simulations of the Hebbian network are shown in the *left* column, anti-Hebbian network in the *center* column, and PCA network in the *right* column. Representations are plotted in first-output-unit-by-second-output-unit space. **d** Measures of Euclidean distance between test points throughout the simulation for each model

among stimulus attributes during the very first block of trials, and discrimination of sounds that did not share this covariance was initially impaired. Following successive blocks of trials, performance on Orthogonal test pairs improved to levels comparable to Consistent test pairs.

We employed three simple unsupervised-learning neural network models with similar architectures, but reflecting different hypotheses about how sensorineural systems exploit covariance, to assess how each accounts for listener data. First is a Hebbian model (Hebb 1949; Oja 1982) in which connection weights adjust in

proportion to the correlation between input and output node activations. Second, an anti-Hebbian (decorrelation) model (Barlow and Földiák 1989; Clifford et al. 2007) that orthogonalizes output dimensions by adjusting symmetric inhibition among output nodes proportional to their correlation. Finally, principal component analysis (PCA) was implemented in a third model (Sanger 1989). Connections to output units adjust in a Hebbian manner; however, the first output inhibits inputs to the second, effectively capturing the principal component from the input pattern and leaving the second unit to capture residual covariance. This model captures correlation across inputs (like the Hebbian model) and orthogonalizes outputs (like the anti-Hebbian model). Final solution spaces for each of the three models are shown in Fig. 7c.

As a model analog of perceptual discriminability, Euclidean distances between output activations for each pair of test sounds were computed at each pass through the 18 stimulus pairs. The Hebbian model, owing to the robust correlation in the input, assimilated Orthogonal and Single-cue stimuli to the correlation (Fig. 7d, left). Thus, this model predicts that performance on Orthogonal cues should collapse and never recover. In the anti-Hebbian model, symmetric inhibition between outputs grew in proportion to their pairwise correlation until activity of the output units was uncorrelated. As a result, the Consistent dimension was compressed, and the Orthogonal dimension strongly expanded. Thus, this model predicts that Orthogonal cues should grow more and more discriminable over time (Fig. 7d, middle). The PCA model quickly discovered the first component (the Consistent dimension) so that distances between Orthogonal and Single-cue pairs initially decreased (Fig. 7d, right). With further experience on the same materials, the PCA model gradually captured the modest variance not explained by the first component, progressively increasing distances between Orthogonal pairs to original relative distances. Listener performance violated predictions of the Hebbian and anti-Hebbian models, but matched the PCA model quite well. Continued exposure with the same highly-structured stimulus set quickly eroded, then gradually recaptured, distinctions along the Orthogonal dimension.

Data from this and numerous additional experiments by Stilp and colleagues (Stilp et al. 2010c; Stilp and Kluender 2012) all support the hypothesis that the auditory system rapidly and efficiently captures covariance (redundancy) across the set of complex stimuli. Like the PCA model, listener performance appears to initially capture the principal component of variation in the two-dimensional stimulus space at the expense of the orthogonal component, and only gradually comes to encode remaining variance. Both this initial component and the second component appear to rapidly become weighted in a way that is proportional to the amount of variance accounted for by each dimension.

The particular PCA model investigated here (Sanger 1989) is certainly oversimplified and is unlikely to exactly reflect neural learning mechanisms. It bears note that, because stimuli were normed to equivalent perceptual distances (JNDs), the perceptual space was linearized in a way that is amenable to a linear model such as PCA. The close correspondence between listener and model performance does suggest that sensorineural processes adapt to reflect experienced covariance

so that dimensions of the perceptual space are weighted in a statistically sensible fashion. Small amounts of experience with highly-correlated items provide evidence that stimuli align along a single dimension, so discriminability of differences along orthogonal components is reduced.

Brief experience with correlation between two acoustic attributes may illuminate how extended experience with natural covariance among many attributes contributes to categorical perception. Studies of categorical perception employ highly familiar complex stimuli that vary along multiple dimensions. One criterion of categorical perception—poor within-category discrimination—may arise from efficient coding of covariance structure in a high-dimensional feature space. To the extent that correlations between stimulus attributes are quite strong and there is reduction in dimensionality, one would predict that discrimination of stimulus differences that do not respect those correlations should be relatively poor. To make such a claim, however, requires extending existing efforts to more complex covariance structures. To date, these investigations have been limited to a single covariance structure with no competing correlations available to listeners. Ongoing studies include two separate correlation structures within a three-dimensional perceptual space. Our expectation is that, when separate patterns of redundancies inherent to respective correlations are learned, variability not predicted within either correlation will be enhanced.

12 Second-Order Statistics in Vowels

Given that correlations between a priori independent acoustic attributes can be learned quickly and efficiently, we now turn to speech sounds with which listeners have a lifetime of experience. We will explore ways through which predictability among acoustic attributes (patterns of fundamental and formant frequencies) may reveal important insights into how vowels are perceived. To the extent that this approach is correct, classic concerns about 'talker normalization' may be dissolved in a way related to that for 'rate normalization'.

We begin with some simple facts concerning the relationship between patterns of formants for different vowel sounds across different talkers. Spectra of vowel sounds include peaks (formants) corresponding to resonances in the vocal tract. Center frequencies of these peaks depend upon two physical properties of vocal tracts. First, formant frequencies depend upon the shape of the vocal tract. Vowels vary mostly in how high or low (close, open) and how forward (front, back) the tongue body is in the oral tract. In addition, some vowels are produced with rounded lips (e.g., /u/ as in "boot") or with different fundamental frequencies among other variations. The center frequency of the F_1 depends primarily upon how low or high the tongue and jaw are positioned. Open vowels with low tongue body such as /æ/ and /ɑ/ have higher F_1 frequencies, and close vowels with high tongue body such as /i/ and /u/ have lower F_1s. When the tongue is placed relatively forward in the vocal tract, the frequency of F_2 for front vowels such as /æ/

and /i/ is higher, but for vowels in which the tongue is placed relatively farther back such as /u/ and /ɑ/, F_2 is lower in frequency. While the center frequency of F_3 also varies across vowel sounds in perceptually significant ways, all vowel sounds in English can be depicted roughly by relative frequencies of F_1 and F_2 with the exception of /ɝ/.

The second major physical characteristic for vowel sounds is length of the vocal tract. When vocal tracts are shorter or longer, center frequencies of formants are higher or lower, respectively. It is given by the physical acoustics of tubes, vocal tracts included, that for a proportional increase or decrease in length, center frequencies of resonances decrease or increase by the same proportion (Nordström 1975). One consequence of this dependency between vocal tract length and vowel acoustics is that vowel sounds are very different across talkers. Vowels judged perceptually to be phonemically the same, such as /æ/ produced by men, women and children, differ greatly in acoustic properties according to vocal tract length. This variation across talkers is so extreme that clear renditions of any given vowel overlap considerably with different vowels by talkers with vocal tracts of different lengths.

Nearey (1989) carefully reviewed both this challenge for vowel perception and a host of potential solutions to the problem of talker-dependent overlap. Following Ainsworth (1975), Nearey divides solutions into two types: intrinsic and extrinsic. Extrinsic models recommend that perception adjusts or normalizes following development of a frame of reference based upon formant frequencies across a talker's entire vowel system (e.g., Ladefoged and Broadbent 1957; Ladefoged 1967; Gerstman 1968; Nordström and Lindblom 1975; Nearey 1978). Inclusion of extrinsic adjustments can contribute to performance of vowel pattern recognition models (Nearey 1989), but extrinsic models will not be discussed further here.

By contrast, intrinsic models assume that differences between talkers can be more elegantly accommodated if one adopts transformations that reveal underlying commonality. Confusion between vowels across talkers is suggested to ameliorate, if not disappear if some relational measure is adopted across an appropriately transformed vowel space. Intrinsic models have a very long history, extending back to Lloyd (1890a, b, 1891, 1892, *cf.* Miller 1989) who claimed that vowels with common articulations result in common perceptions of vowel quality because they share common ratios among formants. Variants of this formant-ratio theory have appeared and reappeared with regularity (e.g., Chiba and Kajiyama 1941; Okamura 1966; Minifie 1973; Broad 1976; Kent 1979; Miller 1989).

Much of talker-dependent differences in vowel sounds, or at least those accounted for by vocal-tract length, decrease following two operations. First, formant center frequencies are converted from a linear scale to more psychoacoustically realistic scales such as logarithmic, Koenig, mel, or Bark. Second, to capture systematicities across talkers, Miller (1989) and Nearey (1989) employed measures of $\log(F_2/F_1)$ and $\log(F_3/F_2)$. When one projects lines corresponding to these ratios across the vowel space, much of the variance between talkers for each vowel is captured. For both Miller and Nearey, ratios that capture this relationship

serve to normalize talker-dependent renditions of vowel sounds in a way that permits matching within a pattern recognition model.

Consistent with efforts related above, we adopt the more contemporary ERB scale (Moore and Glasberg 1983). We employed principal component analysis to measure the amount of shared covariance for 'steady state' $_{ERB}F_1$, $_{ERB}F_2$, and $_{ERB}F_3$ for each of the twelve vowels spoken by 139 men, women, and children and reported by Hillenbrand and colleagues (Hillenbrand et al. 1995). These proportions of variance accounted for by covariance between cochleotopic indices (ERB; see Fig. 8) of spectral peaks are provided in the first column of Table 1. The relationship between ERB F_1, F_2, and F_3 captures over three fourths of the substantial variability across men, women, and child talkers.

13 Redundancy Enhances Sensitivity to Phonemic Distinctions

Our efficient coding approach embraces these systematicities, but departs from previous efforts in two important ways. First, we do not claim that the perceptual utility of these systematicities across talkers is to normalize toward some iconic ideal or template. Instead, these systematicities describe redundancy inherent to formant relationships for each phonemic vowel across talkers. Second, the real perceptual effect of efficiently coding these redundancies is to increase discriminability of each vowel from any other vowel across talkers. Differences between formant patterns that respect quasi-lawful consequences of vocal tract length are

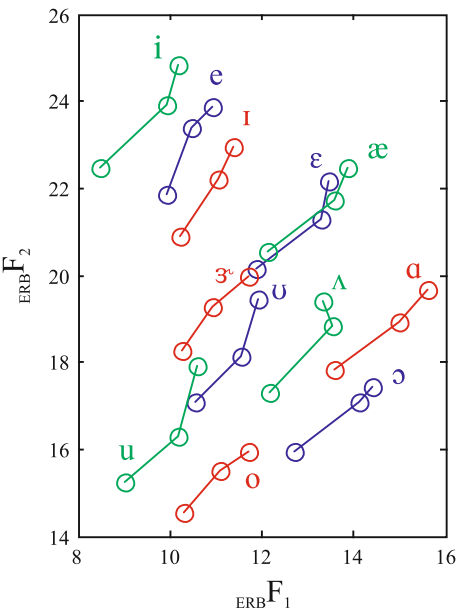


Fig. 8 Mean $_{ERB}F_1/_{ERB}F_2$ of vowels in the Hillenbrand et al. (1995) database measured at vowel midpoint (50 % of overall duration). Measures are averaged across men (*lower-left* point on each line), women (*center*), and children (*upper-right*)

Table 1 Proportions of variance accounted for across talkers (men, women, boys, girls) for American English vowels in the Hillenbrand et al. (1995) database

Vowel	F_1, F_2, F_3	*f*0, mean $F_{1–3}$	*f*0, F_1, F_2, F_3
æ	0.78	0.94	0.79
ɑ	0.76	0.90	0.75
ɔ	0.77	0.90	0.76
ɛ	0.79	0.92	0.79
e	0.76	0.93	0.78
ɝ	0.83	0.90	0.79
I	0.85	0.94	0.83
i	0.82	0.93	0.81
o	0.75	0.86	0.69
ʊ	0.80	0.87	0.74
ʌ	0.80	0.93	0.80
u	0.78	0.85	0.69
Mean	0.79	0.90	0.77

Only vowels with measures of *f*0 and all formant frequencies were analyzed. All calculations are done in ERB frequency. Columns list proportion of variance explained by the principal component (eigenvalue of the principal component divided by the sum of all eigenvalues)

predictable and uninformative with respect to decoding the linguistic message. As a consequence, detection of differences that are linguistically significant, for example 'bet' versus 'bat', is enhanced. Such an approach is, of course, consistent with patterns of perception for vowel sounds. While perception of systematic acoustic changes between vowels may not always be as compellingly categorical as for consonants, it certainly is true that detection of differences between vowel sounds is heightened when stimuli are examples of two different vowels (e.g., /i/ vs. /ɪ/) than examples of the same vowel (e.g., Lotto et al. 1998).

The relationship between formant center frequencies and fundamental frequencies (*f*0s) may be especially telling. Fundamental frequency is a consequence of vocal fold vibration, and rate of vibration depends upon both the mass and stiffness of vocal folds during phonation. While talkers have substantial control over relative stiffness/laxness, vocal fold mass is a different matter. Vocal fold mass, like vocal tract length, is given by anatomy. All things being equal, larger people generally have larger vocal folds, but there are notable exceptions (Nearey 1989). This relationship is complicated by the fact that post-pubescent males have much heftier vocal folds as a secondary sex characteristic. This change in vocal fold mass consequent to puberty results in dramatically lower *f*0 beyond expectations for body size alone. In formal considerations of speech production, source properties owing to vocal fold activity (*f*0) can be viewed as largely independent of filter effects (resonances/formants; Fant 1970).

These facts may lead one to expect that the relationship between *f*0 and formant center frequencies should be much more tenuous than the relationships between $_{ERB}F_1$, $_{ERB}F_2$, and $_{ERB}F_3$. Relationships between vocal tract resonances are inevitable consequences of vocal tract length. By contrast, *f*0 is much freer to vary;

consider singing as an obvious example. Despite the relatively weak physical relationship between vocal tract length and vocal fold mass, especially given mass-enhancing consequences of male development, $_{ERB}f0$ correlates well with mean ERB formant values. Talkers appear to enforce this systematicity despite being under no physiologic obligation to do so. Proportion of variance captured by $_{ERB}f0$ with mean ERB formant frequencies for the same vowels (Hillenbrand et al. 1995) are listed in the second column of Table 1, with a mean of 0.90. Finally, proportions of variance accounted for by covariance between cochleotopic indices of F_1, F_2, F_3, and $f0$ are presented in the third column, with a mean across vowels of 0.77.

Here, we suggest that significant relationships between ERB $f0$ and ERB formant center frequencies suggest that talkers 'know' about the correlation between $_{ERB}F_1$, $_{ERB}F_2$, and $_{ERB}F_3$. Further, talkers adjust their $f0$ when producing vowels in a way that reinforces the redundancy between $_{ERB}F_1$, $_{ERB}F_2$, and $_{ERB}F_3$ by producing an $f0$ that respects this correlation. To the extent that the auditory system seizes upon redundancies, the obligatory relationship between $_{ERB}F_1$, $_{ERB}F_2$, and $_{ERB}F_3$ together with relatively volitional adjustment of $_{ERB}f0$, distinctions between phonemically different vowel sounds are enhanced. At the same time, concerns about talker normalization dissolve. Different talkers all produce vowel sounds that share these relational systematicities or redundancies. Listeners discover these redundancies through experience with speech, and encoding of these redundancies serves to enhance discriminability of more informative differences between phonemically different vowel sounds.

Although some experiments have failed to find a substantial role of $f0$ for identification of vowels in quiet when all other acoustic properties were available to listeners (Katz and Assmann 2001), other studies have suggested that listeners are sensitive to coordination of $f0$ and formant frequencies. When Assmann and Nearey (2007) allowed listeners to adjust either $f0$ or formant center frequencies (maintained at constant log $F_{1,2,3}$ intervals), participants made adjustments that produced combinations of $f0$ and formant frequencies that match the covariance pattern observed in acoustic measurements of natural vowels. Following this study using listeners' adjustment of stimulus attributes, Assmann and Nearey (2008) employed a relatively large-scale identification task using 11 different vowels sounds varying in five or six equal-proportion (log-spaced) steps of formant frequency values and three to six log-spaced steps of $f0$. Performance declined systematically when spectral envelopes were shifted upward or downward from formant frequencies that are typical for a given $f0$, suggesting that perceptual processes engaged in vowel perception are sensitive to covariation of $f0$ and formant frequencies in natural speech. For both of these studies (Assmann and Nearey 2007, 2008), the authors report changes in performance across all vowels tested, and it would be useful to know the extent to which, if any, listeners are sensitive to vowel-specific relationships between ERB $f0$ and ERB formant patterns.

Finally, one recent finding (Katseff et al. 2010) provides further evidence that talkers are implicitly sensitive to relational systematicities among different acoustic attributes of vowel sounds. There is a long history of studying changes in

vocal output when talkers receive acoustically altered auditory feedback. Through the use of clever signal processing methods, some recent studies have investigated if and how listeners adjust their productions when receiving near-simultaneous feedback of their own voice with only F_1 adjusted away from its original peak center frequency (Houde and Jordan 2002; Purcell and Munhall 2006, 2008). The authors of these studies report how talkers adjust their F_1 frequency (higher, lower) in the direction that would compensate for the altered rendition (lower, higher F_1) of their speech.

Katseff and colleagues extended this work by also measuring whether talkers adjusted other formants, specifically F_2, following perturbed auditory feedback of F_1, and whether production of F_1 adjusted in response to perturbation of F_2 feedback. They employed CVC words with the vowels /ɛ/ and /Λ/. Talkers were very unsystematic in their productions of /Λ/ in the presence of altered-formant feedback, but productions of /ɛ/ (depicted as a line in Fig. 9) were systematic and readily interpretable. Consistent with previous studies, when auditory feedback displaced F_1 higher than originally spoken (dashed arrow), talkers decreased F_1 in production a compensatory fashion; the same effect was observed for altered F_2 feedback. However, talkers' adjustments of the unaltered formant are especially telling. When F_1 was altered upward in frequency, talkers increased F_2 in frequency, and when F_2 was altered upward, talkers increased the center frequency of F_1.

If, as we propose, phonemically different vowels are defined by relational properties between formant frequencies (a line in $_{ERB}F_1$–$_{ERB}F_2$ space), there are two ways to restore this relationship when the center frequency of one formant is altered. First, talkers can and do compensate by changing the frequency of the adjusted formant in the direction opposite the alteration. Second, talkers adjust the center frequency of the nonaltered formant in the same direction as the perturbation (net effect depicted in bold arrow), thus better preserving the desired

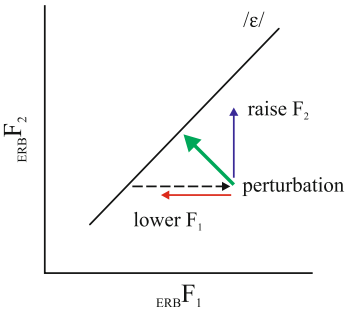


Fig. 9 Schematized results of the altered-feedback experiment by Katseff et al. (2010). As in Fig. 8, /ɛ/ is depicted as a line in $_{ERB}F_1$/$_{ERB}F_2$ space. Listeners hear their own vocalizations of /ɛ/ with F_1 shifted upward (*dashed line*). Listeners compensate for this alteration by simultaneously lowering F_1 and raising F_2 (*solid arrows*). These compensatory shifts have the net effect of taking the shortest path possible to get back to line depicting /ɛ/ (*bold arrow*)

relational properties across both formants. This is another instance for which one can interpret talker's productions in terms of detecting and reinforcing redundancies in ways that enhance discriminability of more informative differences between phonemically different vowel sounds.

14 Discriminative, Not Generative, Models

The reader will note that, for each of the topics discussed thus far, the emphasis has been upon enhancing detection of change. Our information-theoretic premise is that sensorineural systems optimize sensitivity to change. Discovering redundancy between formant peaks and *f*0 enhances phonemically significant differences. Thus, unlike notions of phonetic prototypes, consonants and vowels are revealed much more by what they are not than by how well they approximate some ideal. In this way, our conceptualization is consistent with discriminative, versus generative, models of classification. For example, a generative model of character recognition, such as for reading the address on an envelope, would attempt to capture defining characteristics (*à la* prototype) of each letter (e.g., 'd') across variations such as font and size. By contrast, the discriminative model discovers the ways in which 'd' is distinguished from 'a', 'b', 'c', 'e', 'f', etc. across changes in font and size. In the discipline of pattern classification, discriminative models are greatly preferred over generative models for multiple reasons (Vapnik 1998), not the least of which is that they typically prove more successful (fewer errors) as the size of training sets (experience) grows larger (Ng and Jordan 2002). Further discussion of discriminative versus generative models is beyond the scope of this contribution (see e.g., Vapnik 1998); however, one can capture the main idea simply by thinking about speech perception with respect to confusion matrices (e.g., Miller and Nicely 1955). Correct responses (diagonal) are correct to the extent that distinctions from other stimuli (off diagonal) are detected.

Contrasts between sounds, not commonalities, are emphasized in speech perception. Through experience, perceptual processes come to register predictable patterns of covariance, and by doing so, become especially sensitive to less predictable acoustic changes that distinguish different consonants and vowels. What matters are distinctions between speech sounds, not consonants and vowels *per se*. Listeners hear the sounds of a language by virtue of learning how they are distinguished from all other consonants and vowels. This way of conceptualizing phonetic distinctions harkens back at least to Trubetzkoy (1939/1969) and persists in contemporary contrastive phonology, especially at University of Toronto (e.g., Dresher and Rice 1994). Linguists Roman Jacobson and Morris Halle stated this most starkly in their classic book *Fundamentals of Language* (1971): "All phonemes denote nothing but mere otherness" (p. 22).

15 Learning to Talk

Here, it may be worthwhile to briefly consider implications for young children who are learning to talk. Our claim is that infants learn distinctions between sounds, not consonants and vowels as entities *per se*. Infants can distinguish speech sounds long before they can produce them, as the ways in which they detect differences between sounds become molded to the statistics of their native language sound environment during their first year of life. Information transmission is optimized by maximizing sensitivity to differences; this is the benefit of consolidating redundant attributes. Emphasizing the ways that sounds are different, versus how they are the same, helps illuminate issues concerning learning how to produce speech sounds.

Owing to the developmental course of supralaryngeal anatomy and control, it is impossible for small developing vocal tracts to produce adult-like sounds of a language (e.g., Kent and Miolo 1995; Vorperian et al. 2005). The infant vocal tract begins more as a single tube not unlike that of a chimpanzee. While this configuration facilitates simultaneous drinking and breathing, it impedes production of many speech sounds. The larynx begins too high with a vocal tract too short, and laryngeal and articulatory structures undergo drastic restructuring across the first 6 years (Vorperian et al. 1999, 2005). What is a neotalker to do?

Mimicking speech sounds of the adult is not an option. Resonances are too high and some vocal-tract configurations (e.g., high back /u/) are physiologically impossible. However, it is possible for the developing vocal tract to produce sounds that are different in ways similar to how adult speech sounds differ. Different vocal tract architectures make it fruitless for young children to try to make a veridical match to articulatory or auditory targets. However, the child is able to preserve acoustic contrasts in speech proportional to those heard from adult talkers. In perceptual systems that have little or no access to absolute measures of anything, this quality is both attractive and essential.

16 Conclusions

In this contribution, we first established some first principles that motivate our perspective on speech perception and perception most broadly. We adopted an information-theoretic framework that has a long and productive history in the study of vision and is now more commonly described by the contemporary moniker ‘efficient coding.’ There are two substantial consequences of adopting this information-theoretic framework to questions concerning speech perception. First, distractions concerning objects of perception (gestures versus sounds) are removed. Second, we reframe speech perception as a series of processes through which sensitivity to information—that which changes and/or is unpredictable—becomes increasingly sophisticated and shaped by experience.

We showed how a simple measure of change in the auditory periphery (CSE) proved to be a remarkable predictor of speech intelligibility. Moreover, intelligibility is shown to be critically dependent upon spectral change in vowels and vowel-like speech sounds. Further, we illuminated the ways through which vowel inherent spectral change (VISC) serves to enhance distinctions between vowel sounds, especially those for which relatively static compositions are acoustically or auditorally proximate.

Next, we demonstrated how adopting measures of psychoacoustic change helps to dissolve some traditional concerns about perception across variation in speaking rate that putatively required some process of normalization. Providing further evidence that perceptual processes operate in ways that factor out predictability in order to emphasize spectral change or information, we reviewed studies that show how reliable spectral characteristics of a listening context are factored out of perception entirely.

We then considered how listeners may efficiently code reliable covariance between acoustic attributes of speech sounds as they are structured by lawful properties of the vocal tract. We provided evidence that listeners very quickly learn correlations among stimulus attributes in complex nonspeech sounds, and this has remarkable consequences for discriminability of sounds depending upon whether they respect or violate experienced covariance. We then re-examined well-known relational characteristics among spectral peaks in vowels as a function of talker differences. In this case, efficient coding of these predictable relationships serves both to dissolve concerns about talker normalization and to enhance distinctiveness between renditions of phonemically different vowel sounds.

We suggest that adopting an efficient coding framework provides a productive way to address questions concerning perception of vowel sounds, perception of speech, and perception most broadly.

Acknowledgments We wish to thank Ray Kent, Peter Assmann, and Catherine Rogers for helpful insights from previous drafts of this chapter. Funding has been provided by NIDCD (first and second authors) and SSHRC (third author).

References

Aaltonen, O.: The effect of relative amplitude levels of F2 and F3 on the categorization of synthetic vowels. J. Phon. **13**, 1–9 (1985)

Ainsworth, W.A.: Duration as a cue in the recognition of synthetic vowels. J. Acoust. Soc. Am. **51**, 648–651 (1972). doi:10.1121/1.1912889

Ainsworth, W.A.: The influence of precursive sequences on the perception of synthesized vowels. Lang. Speech **17**, 103–109 (1974). doi:10.1177/002383097401700201

Ainsworth, W.A.: Intrinsic and extrinsic factors in vowel judgments. In: Fant, G., Tatham, M. (eds.) Auditory Analysis and Perception of Speech, pp. 103–113. Academic, London (1975)

Alexander, J.M., Kluender, K.R.: Spectral tilt change in stop consonant perception. J. Acoust. Soc. Am. **123**, 386–396 (2008). doi:10.1121/1.2817617

Alexander, J.M., Kluender, K.R.: Temporal properties of perceptual calibration to local and broad spectral characteristics of a listening context. J. Acoust. Soc. Am. **128**(6), 3597–3613 (2010). doi:10.1121/1.3500693

Assmann, P.F., Nearey, T.M.: Relationship between fundamental and formant frequencies in voice preference. J. Acoust. Soc. Am. **122**, 35–43 (2007). doi:10.1121/1.2719045

Assmann, P.F., Nearey, T.M.: Identification of frequency-shifted vowels. J. Acoust. Soc. Am. **124**, 3203–3212 (2008). doi:10.1121/1.2980456

Assmann, P.F., Summerfield, Q.: The perception of speech under adverse conditions. In: Greenberg, S., Ainsworth, W.A., Popper, A.N., Fay, R.R. (eds.) Speech Processing in the Auditory System, vol. 14, pp. 231–308. Springer, New York (2004). doi:10.1007/b97399

Attneave, F.: Some informational aspects of visual perception. Psychol. Rev. **61**, 183–193 (1954). doi:10.1037/h0054663

Attneave, F.: Applications of Information Theory to Psychology: A summary of Basic Concepts, Methods, and Results. Henry Holt and Company, Inc., New York (1959)

Backus, B.T.: Perceptual metamers in stereoscopic vision. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems, vol. 14. MIT Press, Cambridge (2002)

Barlow, H.B.: Sensory mechanisms, the reduction of redundancy, and intelligence. NPL Symp. Mech. Thought Process. **10**, 535–539 (1959)

Barlow, H.B.: Possible principles underlying the transformations of sensory messages. In: Rosenblith, W.A. (ed.) Sensory Communication, pp. 53–85. MIT Press, Wiley, Cambridge, New York (1961)

Barlow, H.B.: The knowledge used in vision and where it comes from. Philos. Trans. Roy. Soc. Lond. B, Biol. Sci. **352**(1358), 1141–1147 (1997). doi:10.1098/rstb.1997.0097

Barlow, H.B.: Redundancy reduction revisited. Netw: Comput. Neural Syst. **12**, 241–253 (2001). doi:10.1080/net.12.3.241.253

Barlow, H.B., Földiák, P.: Adaptation and decorrelation in the cortex. In: Durbin, R., Miall, C., Mitchison, G. (eds.) The Computing Neuron, pp. 54–72. Addison-Wesley, New York (1989)

Bladon, R.A.W., Lindblom, B.: Modeling the judgment of vowel quality differences. J. Acoust. Soc. Am. **69**, 1414–1422 (1981). doi:10.1121/1.385824

Blumstein, S.E.: The mapping from acoustic structure to the phonetic categories of speech: The invariance problem. Behav. Brain Sci. **21**, 260 (1998). doi:10.1017/S0140525X98221170

Broad, D.J.: Toward defining acoustic phonetic equivalence for vowels. Phonetica **33**, 401–424 (1976)

Bunton, K., Story, B.H.: Identification of synthetic vowels based on a time-varying model of the vocal tract area function. J. Acoust. Soc. Am. **127**, 146–152 (2010). doi:10.1121/1.3313921

Caclin, A., Brattico, E., Tervaniemi Näätänen, R., Morlet, D., Giard, M.-H., McAdams, S.: Separate neural processing of timbre dimensions in auditory sensory memory. J. Cogn. Neurosci. **18**, 1959–1972 (2006). doi:10.1162/jocn.2006.18.12.1959

Chechik, G., Anderson, M.J., Bar-Yosef, O., Young, E.D., Tishby, N., Nelken, I.: Reduction of information redundancy in the ascending auditory pathway. Neuron **51**, 359–368 (2006). doi:10.1016/j.neuron.2006.06.030

Chiba, T., Kajiyama, M.: The Vowel: Its Nature and Structure. Tokyo Publishing Co., Tokyo (1941)

Clements, G.N.: Does sonority have a phonetic basis? In: Raimy, E., Cairns, C. (eds.) Contemporary Views on Architecture and Representations in Phonological Theory, pp. 165–175. MIT Press, Cambridge (2009)

Clifford, C.W.G., et al.: Visual adaptation: neural, psychological and computational aspects. Vision. Res. **47**, 3125–3131 (2007). doi:10.1016/j.visres.2007.08.023

Cole, R., Yan, Y., Mak, B., Fanty, M., Bailey, T.: The contribution of consonants versus vowels to word recognition in fluent speech. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'96), pp. 853–856, Atlanta, GA, (1996)

Delattre, F.C., Liberman, A.M., Cooper, F.S.: Acoustic loci and transitional cues for consonants. J. Acoust. Soc. Am. **27**, 769–773 (1955). doi:10.1121/1.1908024

Diehl, R.L., Kluender, K.R.: On the objects of speech perception. Ecol. Psychol. **1**, 121–144 (1989). doi:10.1207/s15326969eco0102_2
Dresher, B.E., Rice, K.: Complexity in phonological representations. Toronto Working Papers in Linguistics, vol. 12, pp. i–iv (1994)
Fairhall, A.L., Lewen, G.D., Bialek, W., de Ruyter van Steveninck, R.R.: Efficiency and ambiguity in an adaptive neural code. Nature **412**, 787–792 (2001). doi:10.1038/35090500
Fant, C.G.M.: On the predictability of formant levels and spectrum envelopes from formant frequencies. In: Halle, M. (ed.) For Roman Jakobson: Essays on the Occasion of His Sixtieth Birthday, pp. 109–120. Mouton, The Hague (1956)
Fant, G.: Acoustic Theory of Speech Production with Calculations Based on X-Ray Studies of Russian Articulations. Mouton, The Hague (1970)
Fletcher, H.: Speech and Hearing in Communication. Krieger, New York, (1953/1995)
Fogerty, D., Kewley-Port, D.: Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. J. Acoust. Soc. Am. **126**, 847–857 (2009). doi:10.1121/1.3159302
Fowler, C.A.: An event approach to the study of speech perception from a direct-realist perspective. J. Phon. **14**, 3–28 (1986)
Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N.: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. National Institute of Standards and Technology, NTIS Order No. PB91–505065 (1990)
Geisler, W.S., Perry, J.S., Super, B.J., Gallogly, D.P.: Edge co-occurrence in natural images predicts contour grouping performance. Vision. Res. **41**, 711–724 (2001). doi:10.1016/S0042-6989(00)00277-7
Gerstman, L.: Classification of self-normalized vowels. IEEE Trans. Audio Electroacoust. **16**, 78–80 (1968). doi:10.1109/TAU.1968.1161953
Glasberg, B.R., Moore, B.C.J.: Deviation of auditory filter shapes from notched-noise data. Hear. Res. **47**, 103–138 (1990). doi:10.1016/0378-5955(90)90170-T
Gottfried, T.L., Miller, J.L., Payton, P.E.: Effect of speaking rate on the perception of vowels. Phonetica **47**, 155–172 (1990). doi:10.1159/000261860
Greenwood, D.D.: A cochlear frequency-position function for several species—29 years later. J. Acoust. Soc. Am. **87**, 2592–2605 (1990). doi:10.1121/1.399052
Hebb, D.O.: Organization of Behavior. Wiley, New York (1949)
Hedrick, M.S., Nábělek, A.K.: Effect of F2 intensity on identification of /u/ in degraded listening conditions. J. Speech Lang. Hear. Res. **47**, 1012–1021 (2004). doi:10.1044/1092-4388(2004/075
Hillenbrand, J.M., Nearey, T.M.: Identification of resynthesized /hVd/ utterances: effects of formant contour. J. Acoust. Soc. Am. **105**, 3509–3523 (1999). doi:10.1121/1.424676
Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**, 3099–3111 (1995). doi:10.1121/1.411872
Hillis, J.M., Ernst, M.O., Banks, M.S., Landy, M.S.: Combining sensory information: Mandatory fusion within, but not between, senses. Science **298**, 1627–1630 (2002). doi:10.1126/science.1075396
Houde, J.F., Jordan, M.I.: Sensorimotor adaptation of speech i: compensation and adaptation. J. Speech Lang. Hear. Res. **45**, 295–310 (2002). doi:10.1044/1092-4388(2002/023
Jakobson, R., Halle, M.: The Fundamentals of Language. Mouton, The Hague (1971)
Jenkins, J.J., Strange, W., Edman, T.R.: Identification of vowels in 'vowelless' syllables. Percept. Psychophys. **34**, 441–450 (1983). doi:10.3758/BF03203059
Katseff, S., Johnson, K., House, J.: Auditory feedback shifts in one formant cause multi-formant compensation (A). J. Acoust. Soc. Am. **127**, 1955 (2010). doi:10.1121/1.3384960
Katz, W.F., Assmann, P.F.: Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing. J. Phon. **29**, 23–51 (2001). doi:10.1006/jpho.2000.0135
Kent, R.D.: Isovowel lines for the evaluation of vowel formant structure in speech disorders. J. Speech Hear. Disord. **44**, 513–521 (1979)

Kent, R.D., Miolo, G.: Phonetic abilities in the first year of life. In: Fletcher, P., MacWhinney, B. (eds.) Handbook of Child Language, pp. 303–334. Blackwell, London (1995)
Kewley-Port, D., Burkle, T.Z., Lee, J.H.: Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. J. Acoust. Soc. Am. **122**, 2365–2375 (2007). doi:10.1121/1.2773986
Kiefte, M., Kluender, K.R.: The relative importance of spectral tilt in monophthongs and diphthongs. J. Acoust. Soc. Am. **117**, 1395–1404 (2005). doi:10.1121/1.1861158
Kiefte, M., Kluender, K.R.: Absorption of reliable spectral characteristics in auditory perception. J. Acoust. Soc. Am. **123**, 366–376 (2008). doi:10.1121/1.2804951
Kiefte, M.: The perception of spectrally and temporally distorted prevocalic stop consonants. unpublished doctoral dissertation, University of Alberta (2000)
Kiefte, M., Enright, T., Marshall, L.: The role of formant amplitude in the perception of /i/ and /u/. J. Acoust. Soc. Am. **127**, 2611–2621 (2010). doi:10.1121/1.3353124
Klatt, D.H.: Prediction of perceived phonetic distance from critical band spectra: a first step. In: Proceedings of ICASSP, pp. 1278–1281 (1982)
Kluender, K.R., Alexander, J.M.: Perception of speech sounds. In: Dallos, P., Oertel, D. (eds.) The Senses: A Comprehensive Reference, vol. 3, pp. 829–860. Academic, San Diego (2007)
Kluender, K.R., Kiefte, M.: Speech perception within a biologically-realistic information-theoretic framework. In: Gernsbacher, M.A., Traxler, M. (eds.) Handbook of Psycholinguistics, pp. 153–199. Elsevier, London (2006)
Kluender, K.R., Lotto, A.J.: Virtues and perils of empiricist approaches to speech perception. J. Acoust. Soc. Am. **105**, 503–511 (1999). doi:10.1121/1.424587
Kluender, K.R., Diehl, R.L., Killeen, P.R.: Japanese quail can learn phonetic categories. Science **237**, 1195–1197 (1987). doi:10.1126/science.3629235
Kluender, K.R., Coady, J.A., Kiefte, M.: Sensitivity to change in perception of speech. Speech Commun. **41**(1), 59–69 (2003). doi:10.1016/S0167-6393(02)00093-6
Ladefoged, P.: Three Areas of Experimental Phonetics. Oxford University Press, London (1967)
Ladefoged, P., Broadbent, D.: Information conveyed by vowels. J. Acoust. Soc. Am. **29**, 98–104 (1957). doi:10.1121/1.1908694
Lee, J.H., Kewley-Port, D.: Intelligibility of interrupted sentences at subsegmental levels in young normal-hearing and elderly hearing-impaired listeners. J. Acoust. Soc. Am. **125**, 1153–1163 (2009). doi:10.1121/1.3021304
Liberman, A.M., Mattingly, I.G.: The motor theory of speech perception revised. Cognition **21**, 1–36 (1985). doi:10.1016/0010-0277(85)90021-6
Liberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C.: The discrimination of speech sounds within and across phoneme boundaries. J. Exp. Psychol. **54**, 358–368 (1957). doi:10.1037/h0044417
Liljencrants, J., Lindblom, B.: Numerical simulation of vowel quality systems: the role of perceptual contrast. Language **48**(4), 839–862 (1972). doi:10.2307/411991
Lindblom, B.: Phonetic universals in vowel systems. In: Ohala, J.J., Jaeger, J.J. (eds.) Experimental Phonology, pp. 13–44. Academic, Orlando (1986)
Lindholm, J.M., Dorman, M., Taylor, B.E., Hannley, M.T.: Stimulus factors influencing the identification of voiced stop consonants by normal-hearing and hearing impaired adults. J. Acoust. Soc. Am. **83**, 1608–1614 (1988). doi:10.1121/1.395915
Lisker, L.: Rapid versus rabid: a catalogue of acoustical features that may cue the distinction. Haskins Laboratories Status Report on Speech Research, SR-54, pp. 127–132 (1978)
Lloyd, R.J.: Some Researches into the Nature of the Vowel-Sound. Turner and Dunnett, Liverpool (1890a)
Lloyd, R.J.: Speech sounds: their nature and causation (II-IV). Phonetische Studien **4**, 37–67, 183–214, 275–306 (1891)
Lloyd, R.J.: Speech sounds: their nature and causation (V-VII). Phonetische Studien **5**, 1–32, 129–141, 263–271 (1892b)
Lloyd, R.J.: Speech sounds: their nature and causation (I). Phonetische Studien **3**, 251–278 (1890b)

Lotto, A.J., Kluender, K.R., Holt, L.L.: Depolarizing the perceptual magnet effect. J. Acoust. Soc. Am. **103**, 3648–3655 (1998). doi:10.1121/1.423087

Miller, J.L.: Effects of speaking rate on segmental distinctions. In: Eimas, P.D., Miller, J.L. (eds.) Perspectives on the Study of Speech, pp. 39–74. Erlbaum Associates, New Jersey (1981)

Miller, J.D.: Auditory-perceptual interpretation of the vowel. J. Acoust. Soc. Am. **85**, 2114–2134 (1989). doi:10.1121/1.397862

Miller, J.L., Dexter, E.R.: Effects of speaking rate and lexical status on phonetic perception. J. Exp. Psychol. Hum. Percept. Perform. **14**, 369–378 (1988). doi:10.1037/0096-1523.14.3.369

Miller, J.L., Liberman, A.M.: Some effects of later-occurring information on the perception of stop-consonant and semivowel. Percept. Psychophys. **25**, 457–465 (1979). doi:10.3758/BF03213823

Miller, G.A., Nicely, P.E.: An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. **27**, 338–352 (1955). doi:10.1121/1.1907526. [Erratum: (1955) 27, 339. doi:10.1121/1.1907983]

Minifie, F.D.: Speech acoustics. In: Minifie, F.D., Hixon, T.J., Williams, F. (eds.) Normal Aspects of Speech, Hearing, and Language, pp. 235–284. Prentice-Hall, Englewood Cliffs (1973)

Moore, B.C.J., Glasberg, B.R.: Suggested formulas for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. **74**, 750–753 (1983). doi:10.1121/1.389861

Nearey, T.M.: Phonetic Feature Systems for Vowels. Indiana University Linguistics Club, Bloomington (1978)

Nearey, T.M.: Static, dynamic, and relational properties in vowel perception. J. Acoust. Soc. Am. **85**, 2088–2113 (1989). doi:10.1121/1.397861

Nearey, T.M.: Speech perception as pattern recognition. J. Acoust. Soc. Am. **101**, 3241–3254 (1997). doi:10.1121/1.418290

Nearey, T.M., Assmann, P.: Modeling the role of inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986). doi:10.1121/1.394433

Nearey, T.M.: Vowel inherent spectral change in the vowels of North American English. In: Morrison, G.S., Assmann, P.F. (Eds.) Vowel Inherent Spectral Change (ch. 4). Springer, Heidelberg (2012)

Ng, A.Y., Jordan, M.I: On discriminative vs. generative classifiers: a comparison of logistic regresión and naive Bayes. In: Dietterich, T., Becker, S., Ghahramani, Z. (Eds.) Advances in Neural Information Processing (NIPS), vol. 14, MIT Press, Cambridge (2002)

Nilsson, M., Soli, S., Sullivan, J.: Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Am. **95**, 1085–1099 (1994). doi:10.1121/1.408469

Nordström, P.-E., Lindblom, B.: A normalization procedure for vowel formant data. In: Proceedings of the 7th International Congress of Phonetic Sciences, Leeds, England (1975)

Nordström, P.-E.: Attempts to simulate female and infant vocal tracts from male area functions. Speech Transmission Laboratory Quarterly Progress and Status Report (KTH, Stockholm), pp. 2–3, 20–33, (1975)

Ohala, J.J.: There is no interface between phonology and phonetics: a personal view. J. Phon. **18**, 153–171 (1990)

Oja, E.: A simplified neuron model as a principal component analyzer. J. Math. Biol. **15**, 267–273 (1982). doi:10.1007/BF00275687

Okamura, M.: Shouni boin no nenrei teki henka ni kansuru kenkyuu: Sound Spectrograph niyoru formant kouzou to boin no bunka no kentou [Acoustical studies of Japanese vowels in children: The formant constructions and the developmental process]. Nippon Jibiinkoka Gakkai Kaiho [Japan. J. Otolaryngol.] **69**, 1198–1214 (1966). doi:10.3950/jibiinkoka.69.6_1198

Owren, M.J., Cardillo, G.C.: The relative roles of vowels and consonants in discriminating talker identity versus word meaning. J. Acoust. Soc. Am. **119**, 1727–1739 (2006). doi:10.1121/1.2161431

Patterson, R.D., Nimmo-Smith, I., Weber, D.L., Milroy, R.: The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold. J. Acoust. Soc. Am. **72**, 1788–1803 (1982). doi:10.1121/1.388652

Purcell, D.W., Munhall, K.G.: Weighting of auditory feedback across the English vowel space. In: Proceedings of the 8th International Seminar on Speech Production (2008)

Purcell, D.W., Munhall, K.G.: Adaptive control of vowel formant frequency: evidence from real-time formant manipulation. J. Acoust. Soc. Am. **120**, 966–977 (2006). doi:10.1121/1.2217714

Sanger, T.D.: Optimal unsupervised learning in a single-layer linear feedforward neural network. Neural Netw. **2**, 459–473 (1989). doi:10.1016/0893-6080(89)90044-0

Saberi, K., Perrott, D.R.: Cognitive restoration of reversed speech. Nature **398**, 760 (1999). doi:10.1038/19652

Schwartz, O., Simoncelli, E.P.: Natural signal statistics and sensory gain control. Nat. Neurosci. **4**, 819–825 (2001). doi:10.1038/90526

Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)

Simoncelli, E.P.: Vision and the statistics of the visual environment. Curr. Opinions Neurobiol. **13**, 144–149 (2003). doi:10.1016/S0959-4388(03)00047-3

Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. Annu. Rev. Neurosci. **24**, 1193–1215 (2001). doi:10.1146/annurev.neuro.24.1.1193

Stevens, K.N.: Acoustic Phonetics. MIT, Cambridge (1998)

Stevens, K.N., Blumstein, S.E.: The search for invariant acoustic correlates of phonetic features. In: Eimas, P.D., Miller, J.L. (eds.) Perspectives in the Study of Speech. Erlbaum, Hillsdale (1981)

Stilp, C.E., Kluender, K.R.: Cochlea-scaled spectral entropy, not consonants, vowels, or time, best predicts speech intelligibility. Proc. Natl. Acad. Sci. **107**(27), 12387–12392 (2010). doi:10.1073/pnas.0913625107

Stilp, C.E., Kluender, K.R.: Efficient coding and statistically optimal weighting of covariance among acoustic attributes in novel sounds. PLoS ONE **7**(1), e30845 (2012). doi:10.1371/journal.pone.0030845

Stilp, C.E., Alexander, J.M., Kiefte, M., Kluender, K.R.: Auditory color constancy: calibration to reliable spectral properties across nonspeech context and targets. Atten. Percept. Psychophys. **72**, 470–480 (2010a). doi:10.3758/APP.72.2.470

Stilp, C.E., Kiefte, M., Alexander, J.M., Kluender, K.R.: Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences. J. Acoust. Soc. Am. **128**, 2112–2126 (2010b). doi:10.1121/1.3483719

Stilp, C.E., Rogers, T.T., Kluender, K.R.: Rapid efficient coding of correlated complex auditory properties. Proc. Natl. Acad. Sci. **107**(50), 21914–21919 (2010c). doi:10.1073/pnas.1009020107

Story, B.H., Bunton, K.: Simulation and identification of vowels based on a time-varying model of the vocal tract area function. In: Morrison G.S., Assmann P.F. (Eds.) Vowel Inherent Spectral Change (ch. 7), Springer, Heidelberg (2012)

Sussman, H.M., McCaffrey, H.A., Matthews, S.A.: An investigation of locus equations as a source of relational invariance for stop place categorization. J. Acoust. Soc. Am. **90**, 1309–1325 (1991). doi:10.1121/1.401923

Sussman, H.M., Fruchter, D., Hilbert, J., Sirosh, J.: Linear correlates in the speech signal: the orderly output constraint. Behav. Brain Sci. **21**(2), 241–259 (1998). doi:10.1017/S0140525X98001174

Trubetzkoy, N.S.: Principles of Phonology (C. Baltaxe, Translator) University of California Press, Berkeley. (Original work published in 1939) (1969)

Vapnik, V.N.: Statistical Learning Theory. Wiley, New York (1998)

Vorperian, H.K., Kent, R.D., Lindstrom, M.J., Kalina, C.M., Gentry, L.R., Yandell, B.S.: Development of vocal tract length during early childhood: a magnetic resonance imaging study. J. Acoust. Soc. Am. **117**, 338–350 (2005). doi:10.1121/1.1835958

Vorperian, H.K., Kent, R.D., Gentry, L.R., Yandell, B.S.: Magnetic resonance imaging procedures to study the concurrent anatomic development of vocal tract structures: preliminary results. Int. J. Pediatr. Otorhinolaryngol. **49**, 197–206 (1999). doi:10.1016/S0165-5876(99)00208-6

Watkins, A.J.: Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. J. Acoust. Soc. Am. **90**, 2942–2955 (1991). doi:10.1121/1.401769

Watkins, A.J., Makin, S.J.: Perceptual compensation for speaker differences and for spectral-envelope distortion. J. Acoust. Soc. Am. **96**, 1263–1282 (1994). doi:10.1121/1.410275

Weiner, N.: Cybernetics. Wiley, New York (1948)

Part II
VISC Production

Simulation and Identification of Vowels Based on a Time-Varying Model of the Vocal Tract Area Function

Brad H. Story and Kate Bunton

Abstract In their purest form, vowels can be conceived as being produced with static configurations of the vocal tract shape. Laboratory measurements of both acoustic and articulatory characteristics of vowels are typically performed with this assumption. In the case of natural, connected speech, however, the vocal tract shape undergoes nearly continuous change thus a true "static" configuration is rarely produced. Listeners are able to identify vowels in this time-varying situation, often with greater accuracy than for a vowel deliberately produced without any vocal tract change. This chapter examines the time-varying changes of the vocal tract shape that produce vowel inherent spectral change. Specifically, a model of the vocal tract area function is used to investigate how time-dependent formant frequencies originate from movement of the vocal tract.

Abbreviations

$\Omega(x)$	Mean vocal tract diameter function
$\phi_1(x)$	Mode 1 (first principal component)
$\phi_2(x)$	Mode 2 (second principal component)
$A(x)$	Vocal tract area function
$A(x,t)$	Time-dependent vocal tract area function
q_1	Scaling coefficient of the first mode

Electronic Supplementary Material The online version of this article (doi:10.1007/978-3-642-14209-3_7) contains supplementary material which is available to authorized users.

B. H. Story (✉) · K. Bunton
Department of Speech, Language and Hearing Sciences, University of Arizona, Tucson, AZ, USA
e-mail: bstory@email.arizona.edu

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_7,

$q_1(t)$	Time dependent version of q_1
q_2	Scaling coefficient of the second mode
$q_2(t)$	Time dependent version of q_2
F0	Fundamental frequency
F1	First formant frequency
F2	Second formant frequency
F3	Third formant frequency
I0	Intensity
MRI	Magnetic resonance imaging
PCA	Principal component analysis
VISC	Vowel inherent spectral change
XRMB	X-ray microbeam

1 Introduction

The focus of this chapter is on the time-varying changes of the vocal tract shape that produce vowel inherent spectral change. Although a listening study will be discussed, the emphasis here is on how time-dependent formant frequencies originate from movement of the vocal tract and how they might be represented by a model of the vocal tract area function.

The vocal tract area function represents the effect of the articulatory configuration on the overall shape of the airspace extending from the vocal folds to the lips. In particular, it specifies the variation in cross-sectional area of the vocal tract as a function of the distance from the vocal folds (or more accurately, the glottis). A well known set of area functions for a range of vowels and consonants was reported by Fant (1960). These were derived from analysis of both x-ray images and plaster cast models of a talker's vocal tract. More recently, magnetic resonance imaging has been used to obtain three-dimensional representations of the vocal tract from which area functions can be measured directly (cf., Baer et al. 1991; Story et al. 1996). Accurate and realistic area functions are a primary component in the development of certain types of speech production models and speech synthesizers. The typical aim in using such models is to compute the acoustic characteristics of various structural and kinematic variations of the vocal tract, and compare them to similar measurements of natural speech. In this manner, the vocal tract shape-to-acoustic relation can be studied.

An equally important, but less common, aim is to assess the perceptual relevance of speech sounds produced by such models, allowing vocal tract shape, acoustic characteristics, *and* perception to be related. There are, however, few studies in which the stimuli for a perceptual experiment have been generated with

even a simple simulation[1] of the speech production process where, for example, the formant frequencies result directly from the resonant structure of the vocal tract tube rather than from idealizations of formant frequency patterns. An example of this approach was a vowel identification experiment reported by Bunton and Story (2009) in which simulated vowel samples were based on static vocal tract area functions of eight different speakers. A particular vowel was generated by specifying an area function that had been derived from previously-published magnetic-resonance-imaging (MRI)-based measurements (Story et al. 1996, 1998; Story 2005, 2008). Vowels were simulated with a one-dimensional wave-reflection type of vocal tract model coupled to a voice source. Results indicated that the vowels /i/, /æ/, and /u/ were identified most accurately (> 89 %) across all of the speakers. The other vowels /ɛ, ɔ, ɑ, ɪ, e, o, ʌ/ were accurately identified only about 50 % of the time across all of the speakers.

Although these results point to poor identification accuracy for the vowel samples generated from any of the speakers' vocal tract area functions, it was noted that most of the errors were between vowels adjacent to each other in the first-formant–second-formant [F1, F2] vowel space. This suggested that the area function shapes were fairly representative of the target vowels (as had already been shown based on formant frequency comparisons alone, e.g., Story et al. 1996) but some aspects of the simulation were causing listeners to confuse adjacent vowels. There is a fairly large body of research that has shown that vowel inherent spectral change (VISC), such as *time-varying* formant transitions and vowel duration, are important for identification accuracy (Nearey and Assmann 1986; Nearey 1989; Hillenbrand and Nearey 1999; Hillenbrand et al. 2000; Morrison and Nearey 2007; Nittrouer 2007; Hillenbrand 2012 (Chap. 2)), even for typically monophthongal vowels. Such time-dependent changes were absent in the simulated vowels of the Bunton and Story (2009) study because they were generated with a constant duration (300 ms) and the *static* vocal tract shapes produced flat formant frequency contours throughout the time course of each vowel. Presumably, the identification accuracy of the simulated vowels could be enhanced if the vocal tract shape defined by the area function was allowed to change slightly over the duration of each vowel, thereby generating time-varying formant frequencies. Bunton and Story (2010) performed such a study and reported significant improvements in identification of most vowels relative to those produced with the static vocal tract area functions. Interestingly, the identification of the corner vowels /i, æ, ɑ, u/ did not improve with the added time variation of the vocal tract shape.

The aims of this chapter are to (1) review a model of the time-varying vocal tract shape that can be used to generate simulated vowels, (2) review the findings and

[1] Although any artificially-generated speech is strictly *synthetic*, the term *simulated speech* is used in this article to denote that it is produced, to some degree, by simulating the physical processes of human sound production. These consist primarily of vocal fold vibration, acoustic wave propagation in the tracheal, nasal, and vocal tract systems, and the radiated acoustic output. In contrast, formant *synthesis* is an attempt to replicate the acoustic properties of the speech output signal, but not necessarily any of the physical processes that produce those properties.

discuss some limitations of a vowel identification experiment based on simulated vowels, and (3) describe a framework for future experiments using this type of model.

2 Model of the Time-Varying Vocal Tract Area Function

The vocal tract model originated from analysis of an inventory of area functions obtained with magnetic resonance imaging from a single speaker, but has since been generalized for similar data collected from several additional speakers. Thus, the model of the vocal tract will be described with reference to inventories of speaker-specific area functions that have been reported by Story and colleagues for eight speakers (4 females and 4 males) (Story et al. 1996, 1998; Story 2005). A second set of data obtained from the speaker presented in Story et al. (1996) has been published as well (Story 2008). These speakers will be identified in this chapter as they were previously in Bunton and Story (2009, 2010) as SF0, SF1, SF2, SF3, SM0, SM0-2, SM1, SM2, and SM3,[2] where "F" denotes female and "M" male. The SM0-2, SM1, SM3, SF1, SF2, and SF3 contained area functions for the eleven American English vowels (/i, ɪ, e, ɛ, æ, ʌ, ɑ, ɔ, o, ʊ, u/), whereas the SM0 and SF0 sets do not have an area function for the /e/ vowel, and the SM2 set does not have an /ɛ/. Hence, these latter sets contain only ten vowels each. Note that the SM0 and SM0-2 data sets were obtained from the same person but the data collection was separated by several years. These will be treated as if they are two separate speakers, thus the total number of speaker-specific data sets is nine.

Based on principal component analysis (PCA), a model of each speaker's vocal tract shape has been developed that consists of two canonical shaping patterns which can be scaled with amplitude coefficients and superimposed on a mean vocal tract shape to generate a desired area function. More specifically, for any given speaker-specific set of static vowel area functions, a particular vowel in the set can be represented as,

$$A(x) = \frac{\pi}{4}\left[\Omega(x) + q_1\phi_1(x) + q_2\phi_2(x)\right]^2 \tag{1}$$

where x is the distance from the glottis and $\Omega(x)$, $\phi_1(x)$, and $\phi_2(x)$ are the mean vocal tract diameter function and principal components (referred to as "modes"), respectively (Story and Titze 1998). The squaring operation and scaling factor of $\pi/4$ converts diameters to areas.[3] A unique pair of scaling coefficients, q_1 and q_2, corresponds to each vowel in the original area function set and can be used to reconstruct that vowel's area function with Eq. 1. The PCA can also include the

[2] The publications in which each area function set was reported are as follows: SF0 (Story et al. 1998); SF1, SF2, SF3, SM1, SM2, and SM3 (Story 2005); SM0 (Story et al. 1996; also again, but resampled, in Story and Titze 1998); SM0-2 (Story 2008).

[3] Following Story (2005, 2009), the PCA was performed on the equivalent diameters of the cross-sectional areas rather than on the areas themselves.

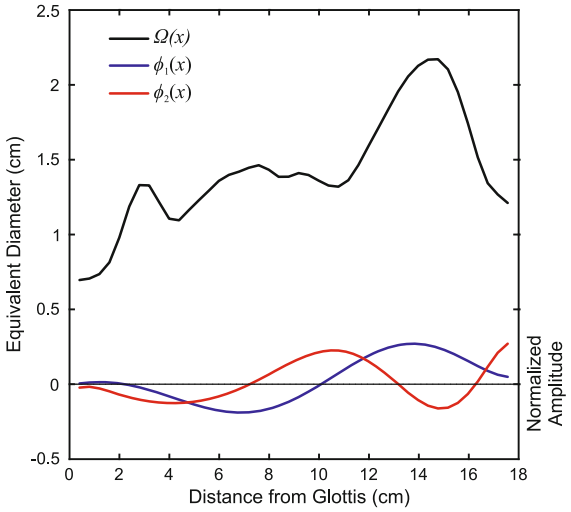


Fig. 1 Mean diameter function $\Omega(x)$ and two modes $\phi_1(x)$ and $\phi_2(x)$ derived from a PCA of area functions for speaker SM0-2 (see text)

effect of vocal tract length which is dependent on both speaker and vowel (Story 2009).

For any of the speaker-specific data sets described previously, the analysis results in a unique set of $\Omega(x)$, $\phi_1(x)$, and $\phi_2(x)$ functions. Shown in Fig. 1 is an example of these functions derived from the SM0-2 set of area functions. The mean diameter function and the modes are shown on the same plot but the axis labels are shown for each on the left and right margins respectively. The analysis also produces pairs of q_1 and q_2 for each of the vowels in the set. The /i/ vowel, for instance, is represented by coefficient values $[q_1, q_2] = [-5.47, 0.57]$ which means that the ϕ_1 mode would be increased in amplitude by more than a factor of five and its polarity would be reversed while the ϕ_2 mode would be reduced in amplitude by nearly one half. When combined with $\Omega(x)$ in Eq. 1, the resulting area function would be expanded in the pharyngeal portion (approximately between 3 and 10 cm from the glottis) and constricted in the oral portion. Interestingly, the shapes of ϕ_1 and ϕ_2 have been shown to be fairly consistent across speakers but the $\Omega(x)$ functions can be quite variable (Story 2005). It was further suggested in Story (2005) that the mode shapes may represent canonical muscle synergy patterns that can be activated to varying degrees to shape the vocal tract for vowel production, whereas, the $\Omega(x)$ function is the idiosyncratic vocal tract shape of a specific speaker on which the modes are superimposed. The modes were later shown to be shaped nearly identically to sum and difference combinations of acoustic sensitivity functions calculated for the $(\pi/4)\Omega^2(x)$ vocal tract shape (Story 2007). Thus, although each mode is derived from statistical analysis (i.e., PCA), they seem to represent vocal tract deformation patterns that are optimized for moving the F1 and F2 formant frequencies precisely and efficiently within a particular speaker's vowel space.

The $[q_1, q_2]$ coefficients for a particular speaker can be incremented along their respective ranges across all the vowels and combined to generate new coefficient

pairs, and hence new area functions that did not exist in the original data set (Story and Titze 1998). It has been shown that the [F1, F2] formant frequencies calculated for nearly all of these new area functions is mapped to a unique $[q_1, q_2]$ coefficient pair. That is, a nearly one-to-one mapping relates the $[q_1, q_2]$ coefficient space to the [F1, F2] vowel space (Story and Titze 1998; Story 2005, 2009). As an example, the mapping calculated for speaker SM0-2 is shown in Fig. 2 where the coefficient mesh in the left panel (Fig. 2a) is mapped to the [F1, F2] formant mesh in the right panel (Fig. 2b). The blue and red lines in Fig. 2a indicate a traversal of the full range of q_1 and q_2 values, respectively, while the other coefficient is set to zero. Their transformation into the [F1, F2] space is shown by the similarly colored lines in Fig. 2b. It can be observed that the q_1 traversal (negative to positive values) corresponds roughly to a continuum of [F1, F2] values extending from an /i/-like vowel to one that is /ɔ/-like, while the [F1, F2] values generated by traversal of the q_2 range extend from an /u/-like to an /æ/-like vowel. The white point located in each plot at the intersection of the red and blue lines signifies the mean vocal tract shape and associated [F1, F2] pair.

This mapping is perhaps most useful when used in the reverse direction to transform formant frequency trajectories extracted from recorded natural speech into coefficient trajectories. In Fig. 2d, the formant space is plotted again, but this time superimposed with [F1, F2] trajectories for 11 vowels measured from the natural speech,[4] of an adult male. The open and solid circles denote the onset and offset, respectively, of each trajectory and indicate the direction that time progresses over the course of the vowel. Transforming the [F1, F2] trajectories to the coefficient domain results in the $[q_1, q_2]$ trajectories superimposed on the mesh in Fig. 2c, where the open and solid markers again denote starting and ending points. With an appropriate time base, each coefficient trajectory can be represented by two time-dependent coefficient functions $q_1(t)$ and $q_2(t)$, as demonstrated in Fig. 3a for the /æ/ vowel. Finally, $q_1(t)$ and $q_2(t)$ can be used to generate a time-varying area function with,

$$A(x,t) = \frac{\pi}{4}\left[\Omega(x) + q_1(t)\phi_1(x) + q_2(t)\phi_2(x)\right]^2 \tag{2}$$

where the x-dependent terms are the same as in Eq. 1. The area function $A(x,t)$ that results from the coefficients for the /æ/ example, and the $\Omega(x)$, $\phi_1(x)$, and $\phi_2(x)$ terms based on speaker SM0-2 is shown in Fig. 3b. It can be seen that the

[4] To obtain spectro-temporal information for the vowel simulation, time-dependent formant frequencies were obtained from productions of eleven American English vowels (/i, ɪ, e, ɛ, æ, ʌ, ɑ, ɔ, o, ʊ, u/), spoken in citation form by an adult male speaker. The vowels were recorded in a sound-treated room with an AKG CS1000 microphone. The signal was acquired in digital form at a sampling frequency of 44.1 kHz with a Kay Elemetrics CSL4400 and saved to a file in "wav" format. Formant frequencies were then estimated over the time course of each vowel with the formant analysis module in Praat (Boersma and Weenink 2009). Formant analysis parameters were manually adjusted so that the formant contours of F1 and F2 were aligned with the centers of their respective formant bands in a simultaneously-displayed wide-band spectrogram. Fundamental frequency (F0) and intensity contours (I0) for each vowel were also extracted with the appropriate Praat modules. All formant, F0, and I0 contours were transferred to vector form in Matlab (Mathworks 2008) for further processing.

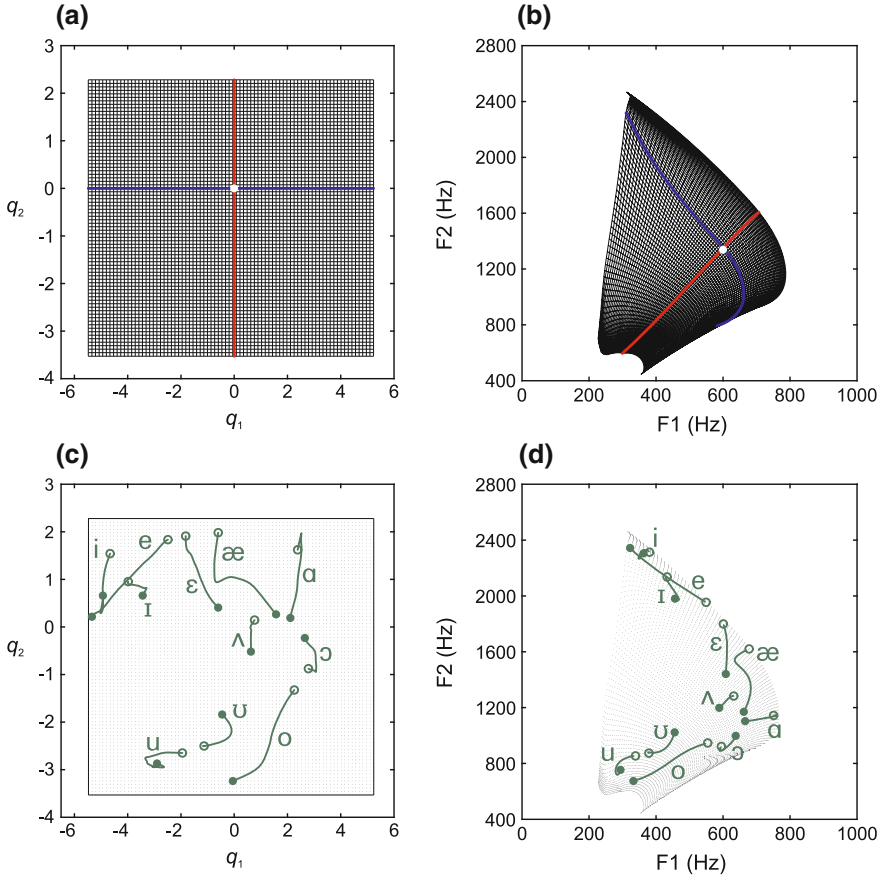


Fig. 2 Demonstration of the formant-to-coefficient mapping based on speaker SM0-2. **a** The mesh represents the mode coefficient space generated from the PCA of SM0-2's original eleven vowels. **b** The deformed mesh represents the [F1, F2] space generated from the coefficient mesh in (**a**). **c** The mesh in the background, bounded by the thin line, is the same as in (**a**) and the trajectories correspond to the formant trajectories in (**b**). **d** The deformed mesh in the background is the same as in (**b**), and the formant trajectories are those measured with formant analysis but slightly rescaled so that they fit entirely within the mesh. In both (**c**) and (**d**), the open and closed circles at the endpoints of each trajectory denote the onset and offset of the vowel, respectively

primary change in vocal tract shape is a slight expansion of the oral cavity (between 12 and 16 cm from the glottis).

3 Vowel Simulation

Simulations of vowels based on time-varying area functions were carried out with a voice source model acoustically and aerodynamically-coupled to a wave-reflection model of the trachea and vocal tract (Liljencrants 1985; Story 1995). The

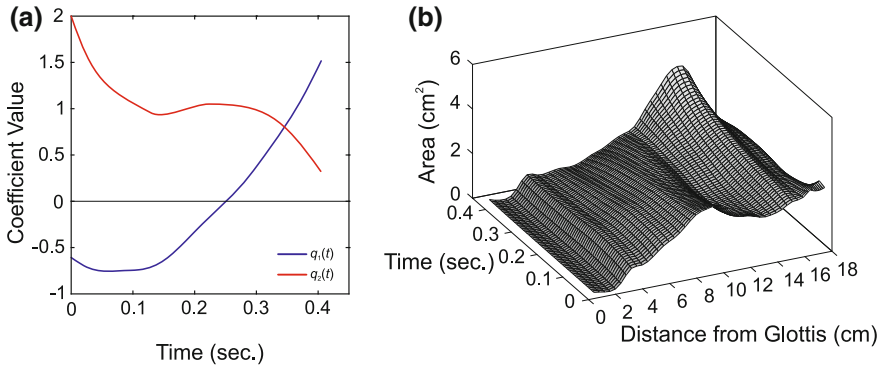


Fig. 3 Demonstration of generating a time-dependent area function for the /æ/ vowel based on speaker SM0-2's vocal tract. **a** Time-dependent mode coefficients $q_1(t)$ and $q_2(t)$ derived from the formant-to-coefficient mapping shown in Fig. 2. **b** Time-varying area function generated from the coefficients in (**a**) used as input to Eq. 2. The area function extends in space from the glottis to the lips and changes shape over approximately 0.4 s of time

vocal tract shape, which extends from glottis to lips, is dictated at every time sample by the given area function $A(x,t)$. The wave propagation algorithm includes energy losses due to yielding walls, viscosity, heat conduction, and radiation at the lips (Story 1995), and accommodates the different vocal tract lengths of each speaker and vowel.

The voice source model is based on a kinematic representation of the medial surface of the vocal folds (Titze 1984, 2006). Control parameters include fundamental frequency, degree of posterior adduction and respiratory pressure, all of which can be varied over the time course of a vowel (or longer utterance). The output of the vocal fold model is the glottal area function which is coupled to the pressures and air flows in the trachea and vocal tract through aerodynamic and acoustic considerations as described by Titze (2002). The resulting glottal flow is determined by the nonlinear interaction of the glottal area with the time-varying pressures present just inferior and superior to the glottis. In addition, a noise component was added to the glottal flow signal if the calculated Reynolds number within the glottis exceeded 1,200. Shown in Fig. 4 is an example of an F0 contour, glottal airflow signal, and output sound pressure generated for an /æ/ vowel based on speaker SM0-2's vocal tract. Time-dependent changes in the F1 and F2 formant frequencies can be seen in the wide-band spectrogram in the bottom panel of the figure. Note that F3 is also shown as a natural consequence of the simulation, but is nearly constant over the duration of the vowel. The corresponding audio sample is Audio 07-01.

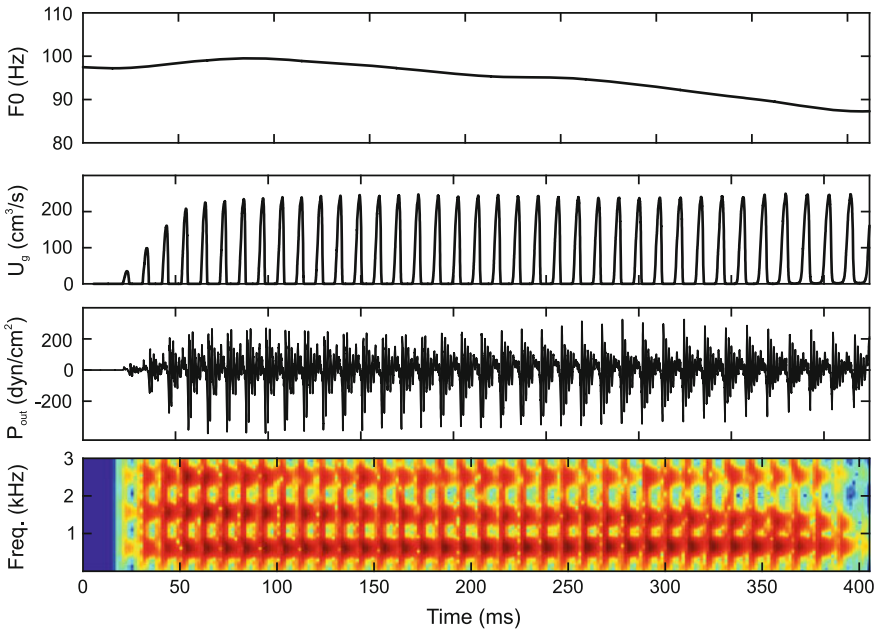


Fig. 4 Example waveforms for the æ vowel generated with the area function in Fig. 3. The *top panel* is the F0 contour used to drive the vibration frequency of the vocal folds. The *second* and *third panels* show the waveforms for glottal flow and radiated sound pressure, respectively, generated by the synthesis method. In the *bottom panel* is a wide-band spectrogram of the output pressure

4 Listener Experiment

Simulated vowels containing inherent spectral change were presented to listeners in a vowel identification experiment reported by Bunton and Story (2010). The purpose was to compare the results to those of a previous experiment (Bunton and Story 2009) in which simulated vowels based on *static* area functions were presented. Specifically, it was hypothesized that identification accuracy of simulated vowels based on vocal tract area functions would be enhanced if the shape defined by the area function was allowed to change over the duration of each vowel, and that duration was vowel dependent. Ideally, comparison of identification accuracy of time-varying versus static vowels should be based on a single experiment in which all listeners respond to both vowel types. The Bunton and Story (2009) study, however, was conducted as an independent experiment to perceptually assess the quality of vowels produced with measured area functions. The results of that experiment motivated the design of the more recent study (Bunton and Story 2010) with time-varying vowels. Thus, although the results of the two studies will be compared, the conclusions must be considered tentative until a more systematic experiment is performed, such as the one outlined in a later section of this chapter.

Time-varying area functions were generated with the formant-to-coefficient mapping technique for the eleven vowels of the speaker SM0-2 shown previously in Fig. 2c and d. For the other eight speakers described in an earlier section, the [F1, F2] trajectories from Fig. 2d were rescaled so that they fit entirely within a given speaker's formant mesh (i.e., each speaker's formant space is different). Each vowel trajectory was then transformed to that particular speaker's coefficient space so that a time-varying area function could be produced with Eq. 2. In total, 88 time-dependent area functions were generated across eight speakers and eleven vowels. Note that even though 3 of the original area function sets (for SF0, SM0, and SM2) contained only 10 vowels, 11 vowels could be generated with the formant-to-coefficient mapping technique.

A vowel was simulated based on each time-varying area function. The fundamental frequency (F0) for each male vowel sample was varied according to the F0 contour obtained by acoustic analysis[4]. For the female vowels, each measured F0 contour was multiplied by a factor of two. For example, the peak F0 in the contour for the male /i/ vowels was 112 Hz whereas for the female it was 224 Hz. The respiratory pressure for each sample, male or female, was ramped from 0 to 7,840 dyn/cm^2 in the initial 50 ms with a cosine function, and then maintained at a constant pressure for the remaining duration of the utterance. The posterior adduction of the vocal folds was varied slightly over the time course of each synthetic vowel according to the shape of the intensity contour measured in the acoustic analysis of the recorded vowels. That is, the adduction was greatest (vocal folds closest together) at the point where the intensity of a particular recorded vowel was highest. Because of the somewhat more breathy quality of female speakers (e.g., Klatt and Klatt 1990), the adduction was set to be 30 % greater for the vowels generated from the female area functions. Other model parameters were set to constant values throughout the time course of each utterance.

The durations of each simulated vowel were based on the measurements reported for male and female speakers by Hillenbrand et al. (1995, p. 3103). However, because they were measured for vowels embedded within "hVd" words, the durations were increased by 50 % so that the resulting isolated vowels would be similar to the length of an hVd word. The sound pressure signal produced by each vowel simulation was converted to an audio file for presentation in a listening experiment. The entire series of time-varying vowel samples based on speaker SM0-2 can be heard in Audio 07-02 and are arranged in this order: (/i, ɪ, e, ɛ, æ, ʌ, ɑ, ɔ, o, ʊ, u/), For purposes of comparison, the *static* versions of SM0-2's vowels used in Bunton and Story (2009) can be heard in Audio 07-03.

The simulated vowel samples generated for each speaker were presented to ten listeners, each of whom had either completed or were enrolled in a basic university-level phonetics course at the University of Arizona. Following presentation of the target vowel, listeners were asked to use a computer mouse to select one of the buttons displaying the eleven English vowels on a computer screen. Each button listed the phonetic symbol for the vowel (/i, ɪ, e, ɛ, æ, ʌ, ɑ, ɔ, o,

Table 1 Percentage of vowels identified correctly for each speaker across listeners

	Speaker									
Vowel	SM0	SM0-2	SM1	SM2	SM3	SF0	SF1	SF2	SF3	Mean
i	92	74	78	88	94	86	88	50	90	**82**
ɪ	42	68	88	78	80	80	84	82	80	**76**
e	100	98	100	100	98	98	94	94	98	**98**
ɛ	68	92	84	82	90	98	88	98	100	**89**
æ	94	62	84	92	92	92	58	92	74	**82**
ʌ	90	96	96	92	94	66	94	98	92	**91**
ɑ	52	40	48	68	52	22	52	32	50	**46**
ɔ	82	88	80	76	80	86	76	80	82	**81**
o	98	94	94	98	100	76	92	94	90	**93**
ʊ	64	80	92	76	90	88	86	88	76	**82**
u	82	86	86	92	86	70	66	68	78	**79**
Mean	**79**	**80**	**85**	**86**	**87**	**78**	**80**	**80**	**83**	

The bottom row indicates the mean identification accuracy *across all vowels for each speaker*, and the rightmost column indicates the mean identification accuracy *across all speakers for each vowel*. From Bunton and Story (2010)

ʊ, u/), and a corresponding hVd word. Each listener heard five repetitions of each vowel sample blocked by speaker sex in random order.

Percent correct identifications of each vowel based on each speaker are shown in Table 1. The mean identification accuracy across all vowels for individual speakers ranged from 79 to 87 % (see bottom row of table). For individual vowels, mean identification accuracy ranged from 46 % for /ɑ/ to 99 % for /e/. Mean accuracy was greater than 70 % for the remaining vowels. A composite confusion matrix including the identification data based on all speakers (across listeners) is shown in the upper half of Table 2. Correct identification of target vowels can be seen along the diagonal in boldface cells. Vowel confusions were typically between adjacent vowel categories in the [F1, F2] vowel space. For the vowel /i/, confusions occurred with both /ɪ/ and /ɛ/. Vowels /ɪ/ and /ɛ/ were frequently confused with each other. For all speakers, the central vowel /ʌ/ was confused with both /æ/ and /ɑ/. For the back vowels, /ɑ/ was most frequently confused with /ɔ/ which is not surprising since these two categories tend to be collapsed in the southwest United States (Labov et al. 2006) where the experiment was conducted. The vowels /u/ and /ʊ/ were also confused by listeners for all speakers.

To compare the results for the time-varying vowels in the present study to those for static vowels, a composite confusion matrix was calculated from the individual speaker confusion matrices reported in Bunton and Story (2009). This is shown in the lower half of Table 2. As mentioned previously, such a comparison is complicated by not presenting both the time-varying and static vowels to the same listeners. Coincidentally, however, eight of the ten listeners in the time-varying vowel experiment also participated in the previous static vowel experiment, and the two experiments were separated in time by more than one year so learning effects should not have influenced the results. It can be seen from the confusion

Table 2 Composite confusion matrix of the vowels identified across speakers

		Listeners' identification										
		i	ɪ	e	ɛ	æ	ʌ	ɑ	O	o	ʊ	u
Vowel intended by speaker (time-varying)	i	**82**	11	1	6	0	0	0	0	0	0	0
	ɪ	1	**76**	2	21	0	0	0	0	0	0	0
	e	0	0	**98**	0	1	0	0	0	0	0	0
	ɛ	0	0	0	**89**	9	1	2	0	0	0	0
	æ	0	0	2	10	**82**	2	3	1	0	0	0
	ʌ	0	0	0	3	2	**91**	4	1	0	0	0
	ɑ	0	0	0	0	16	1	**46**	37	0	0	0
	ɔ	0	0	0	0	0	1	17	**81**	0	0	0
	o	0	0	0	0	0	0	2	2	**93**	3	0
	ʊ	0	0	0	0	0	6	1	0	2	**82**	9
	u	0	0	0	0	0	1	0	0	2	18	**79**
Vowel intended by speaker (static)	i	**93**	6	0	1	0	0	0	0	0	0	0
	ɪ	2	**25**	39	22	1	9	0	0	0	0	0
	e	2	27	**34**	35	1	0	0	0	0	0	0
	ɛ	0	2	12	**36**	35	2	0	0	0	10	2
	æ	0	0	0	2	**97**	0	1	1	0	0	0
	ʌ	0	0	0	0	0	**34**	20	14	5	24	4
	ɑ	0	0	0	0	12	6	**50**	28	4	0	0
	ɔ	0	0	0	0	0	1	25	**61**	13	1	0
	o	0	0	0	0	0	9	7	7	**39**	28	11
	ʊ	0	0	0	0	0	2	7	8	12	**40**	32
	u	0	0	0	0	0	0	0	0	1	11	**88**

The values in each cell are shown as percent. The upper half of the table shows data from the present study and the lower half shows identification data based on the static vowel experiment reported in Bunton and Story (2009). From Bunton and Story (2010)

matrix in Table 2 that seven of the vowels (/ ɪ, e, ɛ, ʌ, ɔ, o, ʊ /) synthesized with a time-varying vocal tract shape were identified more accurately than the vowels based on static vocal tract shapes reported by Bunton and Story (2009). With the exception of /ɔ/, the increase in accuracy over the static cases was 50 % points or more. For /ɔ/ the increase was only 20 % points. Identification accuracy of the /ɑ/ vowel was similar in both studies. The three corner vowels /i, æ, u/ were less accurately identified when the vocal tract shape varied in time than when it was static. The decrease in identification accuracy for these vowels ranged from 9 to 16 % points.

Demonstrating improvement in identification of some vowels by incorporating additional time-varying cues was not unexpected. As has been explained in the other chapters of this volume, time-varying formant transitions and vowel duration are well known to be important cues for improved identification accuracy. It is curious, however, that the identification accuracy of the time-varying vowels in Bunton and Story (2010) was below those reported for similarly time-varying vowels generated with a formant synthesizer (e.g., Hillenbrand and Nearey 1999). One difference is that formant synthesis allows precise control of the formant

frequencies and bandwidths over the time course of a vowel, whereas simulation of vowels is based on generating movement of the vocal tract. Although formants extracted from natural speech were mapped onto movement information (i.e., $[q_1, q_2]$ coefficients) to drive the vocal tract model, they were based only on F1 and F2 (see Fig. 2). That is, when coupled to the voice source and trachea, the time-varying area functions produced sound samples that contained [F1, F2] formant trajectories based on the original recording, but there was no direct control of the formants higher than F2 even though higher formants existed in the signal due to the resonant structure of the vocal tract shape (cf. bottom panel of Fig. 4). For some of the vowels generated, the pattern of formants F3 and higher created information that likely conflicted with the target vowel. An example is the time-varying /i/ based on SM-02's vocal tract that is shown spectrographically in Fig. 5. The [F1, F2] trajectory for this synthetic vowel is precisely that shown in the upper left corner of Fig. 2d, and indicates little movement of either F1 or F2. There is, however, a downward glide of F3 such that the distance between F2 and F3 decreased over the duration of the vowel, and perhaps contributed to its confusion with /ɪ/. Interestingly, the length of the corresponding $[q_1, q_2]$ trajectory for this vowel in the coefficient space (Fig. 2c) indicates there was a change occurring in vocal tract shape, but in this case the change primarily affected F3.

It is not surprising that time-varying changes of the area function intended to move F1 and F2 in some specific pattern might also have unintended effects on the upper formant frequencies. Acoustic modeling of the vocal tract shape has shown that even subtle changes in cross-sectional area may have large effects on some formants (e.g., Stevens 1989; Story et al. 2001, Story 2006), especially when such changes occur in a part of the vocal tract that is already fairly constricted. Thus, it can be predicted that the vowels /i, ɑ, u/ would be particularly susceptible to these effects because, compared to other vowels, they typically are produced with the most constricted vocal tract shapes. Perhaps this is at least a partial explanation of why these vowels were not identified with greater accuracy that the static versions. A future step in this process is to build in more control of the upper formant frequencies via the area function model. For example, the addition of a third mode (from the PCA) to the model may allow for a systematic relation of the third

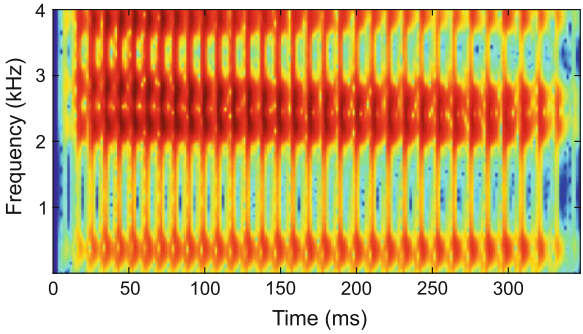


Fig. 5 Spectrogram of the simulated /i/ vowel based on speaker SM0-2. Note that F3 glides downward over the time course of the vowel

formant to vocal tract shape, although this would be expected to covary with the first two modes as well.

It is also noted that generation of the glottal flow signal based on nonlinear interaction of the vocal tract pressures and glottal aerodynamics is a realistic simulation of the process (Titze and Story 1997, Titze 2008) but is a radically different approach to producing artificial speech sounds than synthesizers based on linear source-filter theory. With nonlinear interaction, the amplitudes of harmonic components in the glottal flow signal are dependent on the vocal tract pressures that exist at each time instant, and thus dependent on the particular vocal tract shape that exists at a given point in time. The output pressure signal (i.e., the speech signal) is the result of this nonlinear interaction and the harmonic enhancement that occurs due to the resonances of the vocal tract. In a linear source-filter model, the amplitudes of the harmonic components in the glottal flow signal are entirely specified by the parameters of the glottal pulse shape, independent of any considerations of the vocal tract. These amplitudes are then enhanced or suppressed based on the characteristics of the vocal tract transfer function. In a nonlinearly interactive system, a time-varying area function may alter the harmonic amplitudes in the speech signal in ways that are not linearly related to the change in formant frequencies.

5 Future Directions

A limitation of the Bunton and Story (2010) study was that the [F1, F2] formant trajectories used to produce vowels for all speakers' vocal tracts were based on analysis[4] of a single recording of each vowel spoken by one male talker. This was done for consistency across all vowel samples but may have also contributed to some of the confusion between vowels. The effect of vowel inherent spectral change on vowel identification is often described in terms of three hypotheses, each of which ascribe high relevance to the initial formant frequencies but differ in regard to what is most relevant as the formants change over the vowel duration. As explained in Morrison (2013a), the first hypothesis considers the formant values at the end of the vowel to be most relevant, whereas in the second hypothesis the relevant cue is the rate of change of formant frequencies over time. In the third hypothesis, the direction of formant frequency change is considered most relevant. Because all three of these types of cues would have been embedded, more or less, in each vowel sample simulated for the Bunton and Story (2010) study, none of the three hypotheses were explicitly tested. Rather the hypothesis was simply that a time-varying vocal tract shape (area function) would produce more accurately identifiable vowels than static vocal tract shapes. The model described in this chapter, however, is configured such that a more systematic study of production and perception of time-varying vowels could be conducted. The purpose of this section is to describe a framework for future research with this type of model.

5.1 Vowel Perception Based on Static and Time-Varying Area Functions

It is proposed to conduct a series of experiments in which simulated vowels based on both static and time-varying area functions are presented to listeners in an identification paradigm. Unlike the previous studies that focused on vocal tract shapes or formant trajectories as specific vowel targets, the proposed study will first "map out" the vowel identification of stimuli generated from an equally-spaced sampling of $[q_1, q_2]$ pairs. This is demonstrated in Fig. 6 for a male (SM0 & SM0-2 combined as in Story 2008) and female (SF2 from Story 2005) speaker

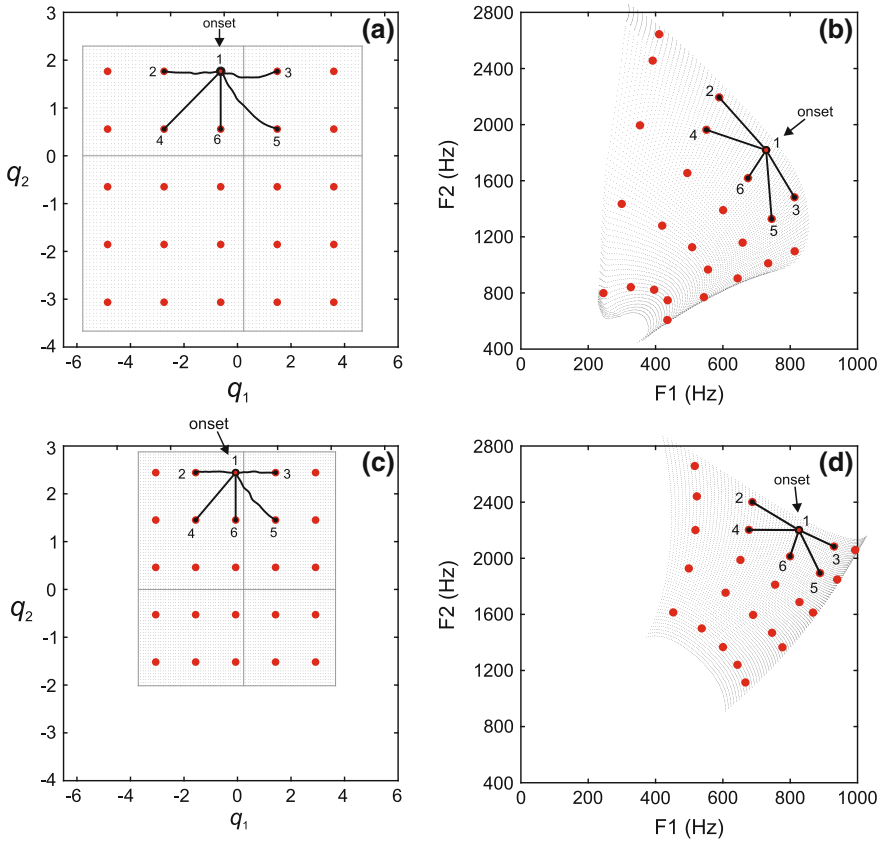


Fig. 6 Demonstration of how stimuli for a future experiment might be sampled from formant-to-coefficient mappings based on a male (**a** and **b**) and female (**c** and **d**) speaker. The *red dots* in each coefficient space represent possible sample points from which static area functions could be generated; the corresponding [F1, F2] pairs are similarly shown in the vowel spaces. The spider-like constructions indicate a possible technique for systematically-generating time-varying area functions

where the coefficient spaces in the left column (Fig. 6a and c) have been sampled in a five-by-five matrix along the q_1 and q_2 dimensions, as shown by the red dots. The corresponding [F1, F2] formant frequencies are shown superimposed on the formant space in Fig. 6b, also by red dots. The first aim would be to determine how each of these samples would be identified, for the purpose of understanding whether there are, in fact, static vocal tract shapes that produce easily identifiable vowels.

The next aim would be to systematically create a collection of time-varying area functions from which vowels with inherent spectral change can be generated. As a demonstration of the process, the black "spider-like" connections of one sample to another shown in Fig. 6b and d represent linear formant movement away from the onset point as indicated by the label. The movement, which could be set to be rapid, moderate, or slow, is directed toward each of the surrounding five sample points that serve as the [F1, F2] offset values. The mappings of these five formant shifts to their corresponding $[q_1, q_2]$ coefficient spaces are shown in Fig. 6a and c where the spider-like shape is retained but the nonlinearity of the mapping is apparent in the curvature of the coefficient trajectories. With Eq. 2, each trajectory can be used to generate a time-varying area function from which a vowel can be simulated and subsequently used in an identification experiment. Alternatively, the coefficient trajectories could be specified as linear (in the $[q_1, q_2]$ space) and then the resulting formant trajectories would take on a curvature due to the nonlinear coefficient-to-formant mapping. In either case, the coefficient trajectories can be thought of as a type of movement "signal," not representative of individual articulators, but rather a signal indicating a coordinated movement of the articulators that affects the shape of the entire vocal tract, for purposes of positioning and moving the resonance frequencies. The identification of stimuli produced in this manner would provide more insight into how the temporal variation of the vocal tract shape affects perception of vowels.

As an example, static and time-varying vowels were produced from the mapping of the black dots and spider-like connections, respectively, shown in Fig. 6a and b. These can be heard as a set of sound files described as follows. Audio 07-04–Audio 07-09 correspond to the static vowels indicated by the numbers in Fig. 6a and b. Audio 07-10–Audio 07-14 are sound samples of the time-varying vowels produced by movement from dot "1" to each of the other numbered dots. In all cases, the vowel duration was 330 ms and the F0 contour was falling.

5.2 Articulatory Fleshpoint Tracking

Story (2007) described an approach to investigating the relation of the time-varying vocal tract shape to the acoustic output that could perhaps be extended to facilitate understanding some aspects of vowel inherent spectral change. The approach involves measuring a cross-distance function of the oral portion of the

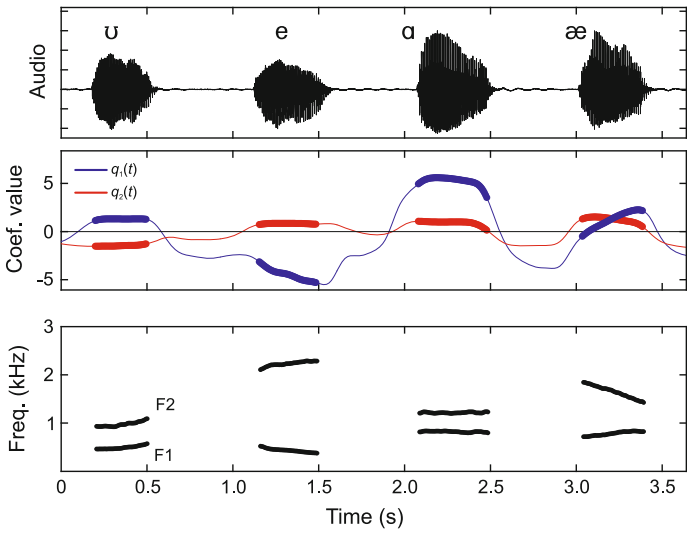


Fig. 7 Audio waveforms (*upper*), time-varying coefficients (*middle*), and formant contours (*bottom*) based on the production of four vowels by a male talker. Note the time-varying coefficients are continuous throughout the entire 3.6 s duration; the lines are thickened during the portions of time where sound is present

vocal tract based on fleshpoint pellets affixed to the tongue, lips and jaw in the midsagittal plane. Tracked over time with the X-ray microbeam (XRMB) system (Westbury 1994), these pellets can be used to produce a time-dependent representation of the oral part of the vocal tract shape that can be analyzed with a PCA in much the same way as was done with the MRI-based area functions. Shown in the top panel of Fig. 7 is the audio signal of a series of four vowels spoken in succession by a male talker while in the XRMB system. The middle panel shows the coefficients ($[q_1(t), q_2(t)]$) derived via PCA and subsequent processing (see Story (2007) for details) as they vary over the time course of the four vowels, similar to the coefficient traces shown previously in Fig. 3a for an /æ/ vowel.

An interesting by-product of this process is that the time-variation of the coefficients can be seen not only during production of the vowels, but also between them where no sound is produced. In Fig. 7, the segments of time during which sound is present, and hence formant frequencies expressed, are indicated by the thick lines, whereas the silent portions are shown as thin lines. This offers some insight into how a speaker configures the vocal tract prior to initiating voicing. For example, just before the ʊ vowel was excited acoustically, the speaker altered the shape of the vocal tract such that q_1 was increased from a negative to positive value. Over the duration of the vowel the coefficients change only slightly, but even so, there is a fairly large change in the corresponding [F1, F2] frequencies plotted in the bottom panel of the figure. Almost instantly following cessation of the /ʊ/ vowel (i.e., at about 0.5 s) the q_1 coefficient begins to shift downward toward the value needed to initiate the upcoming /e/ vowel. Unlike the previous

vowel, the /e/ is characterized by a continuous change in the q_1 coefficient during production of the vowel in order to generate the time-varying change in the formants (see bottom panel) needed to produce a diphthong. Again almost immediately after the vowel sound ends, the coefficients begin to shift toward the values needed to produce the next vowel /ɑ/. The /ɑ/ is produced with fairly constant coefficient values, as is reflected in the rather flat formant contour in the bottom panel of the figure. In contrast, q_1 rises continuously during production of the fourth vowel /æ/ while the q_2 coefficient decreases during this same period of time. These vocal tract shape changes have the effect of creating a nearly linear increase and decrease in F1 and F2, respectively.

An alternative view of these data is shown in Fig. 8 where the time-varying coefficients and formants have been plotted in the $[q_1, q_2]$ and [F1, F2] planes, respectively. In Fig. 8a the coefficient trajectory shown with the thin line represents the entire 3.6 s duration of the four vowels. The open and solid dots located in the lower left quadrant indicate the beginning and end of the trajectory, respectively, and the small arrows throughout the plot denote the direction of time. The portions of the trajectory shown with red lines indicate when sound was being produced by the talker; again the open dot on each indicates the starting point and the solid dot is the end point. The corresponding [F1, F2] trajectories are similarly plotted as red lines with direction arrows in Fig. 8b. These plots illustrate that each of the four vowels is produced as the $[q_1, q_2]$ trajectory passes through a loop in a particular region of the coefficient space. The talker initiates and terminates voicing at apparently precise points on each loop to excite the vocal tract resonances (hence producing formants in the speech signal) that are desired for a given vowel. As an example, the /ʊ/ vowel is produced only during a small portion of the upward excursion of the first trajectory loop (note that "small portion" refers to

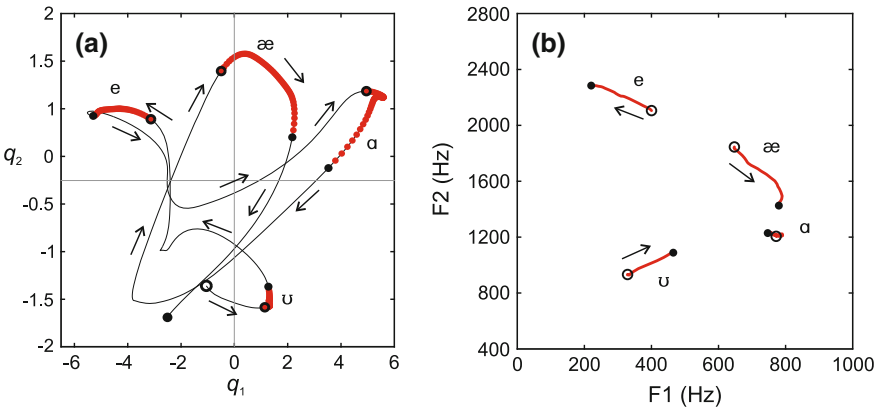


Fig. 8 $[q_1, q_2]$ and [F1, F2] trajectories of the four vowel sequence in Fig. 7. **a** The entire 3.6 s of the trajectory is shown with the *thin black line*; the *thick red lines* indicate portions of the trajectory where sound is present for the labeled vowels. **b** [F1, F2] trajectories that correspond to the the *red lines* in the (**a**)

the distance covered in the coefficient space and is not related to the duration of the vowel). This, however, generates a considerable upward shift in the corresponding [F1, F2] trajectory as seen in the lower left portion of Fig. 8b. The other three vowels are produced over larger portions of their respective loops, although for the /ɑ/ most of this movement occurs only for a short period near the end of the vowel. The resulting formant trajectories indicate considerable shift over the time course of both /e/ and /æ/, albeit in opposite directions, but nearly constant [F1, F2] values for /ɑ/.

These data suggest that at least some aspects of vowel inherent spectral change originate from the patterns of movement talkers use to produce speech sounds. While that may seem obvious, the point is that even for isolated vowels produced in succession (i.e., as read from a list) a talker's vocal tract is in nearly continuous motion. A question is how does a talker decide or know when to "turn on" and "turn off" the sound during these large movement patterns of the vocal tract in order to produce a target vowel? Do these initiation and termination points on the trajectory loops need to be precise or can they be somewhat sloppy? Do the loops provide a better sampling of the [F1, F2] space than is possible with a static vocal tract shape, thus facilitating perception? Or are such loops a necessity in generating movement between vowel targets, and the spectral change that occurs is simply a by-product of this process. Further research connecting vocal tract movement, sound production, and perception may allow for answers to some of these questions.

Acknowledgments This research was supported by NIH grant number R01-DC04789.

References

Baer, T., Gore, J.C., Gracco, L.C., Nye, P.W.: Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. J. Acoust. Soc. Am. **90**, 799–828 (1991). doi:10.1121/1.401949

Boersma, P., Weenink, D.: Praat, Version 5.1, www.praat.org, (2009) last viewed on February 2, 2009

Bunton, K., Story, B.H.: Identification of synthetic vowels based on selected vocal tract area functions. J. Acoust. Soc. Am. **125**, 19–22 (2009). doi:10.1121/1.3033740

Bunton, K., Story, B.H.: Identification of synthetic vowels based on a time-varying model of the vocal tract area function. J. Acoust. Soc. Am. **127**, EL146–EL152 (2010). doi:10.1121/1.3313921

Fant, G.: The Acoustic Theory of Speech Production. Mouton, The Hague (1960)

Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**, 3099–3111 (1995). doi:10.1121/1.409456

Hillenbrand, J., Nearey, T.: Identification of resynthesized /hVd/ utterances: Effects of formant contour. J. Acoust. Soc. Am. **105**, 3509–3523 (1999). doi:10.1121/1.411694

Hillenbrand, J., Clark, M., Houde, R.: Some effects of duration on vowel recognition. J. Acoust. Soc. Am. **108**, 3013–3022 (2000). doi:10.1121/1.1323463

Hillenbrand, J.M.: Static and dynamic approaches to understanding vowel perception. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 2). Springer , Heidelberg (2013)

Klatt, D.H., Klatt, L.C.: Analysis, synthesis, and perception of voice quality variations among male and female talkers. J. Acoust. Soc. Am. **87**, 820–857 (1990). doi:10.1121/1.398894

Labov, W., Ash, S., Boberg, C.: The Atlas of North American English: Phonetics. Mouton de Gruyter, Berlin (2006)

Liljencrants, J.: Speech synthesis with a reflection-type line analog. DS Dissertation, Department of Speech Communication and Music Acousstics, Royal Institute of Technology, Stockholm, Sweden (1985)

The Mathworks, Matlab, Version 7.6.0.324 (R2008a).

Morrison, G.S.: Theories of vowel inherent spectral change: A review. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 3). Springer, Heidelberg (2013a)

Morrison, G.S., Nearey, T.M.: Testing theories of vowel inherent spectral change. J. Acoust. Soc. Am. **122**, EL15–EL22 (2007). doi:10.1121/1.2739111

Nearey, T.M.: Static, dynamic, and relational properties in vowel perception. J. Acoust. Soc. Am. **85**, 2088–2113 (1989). doi:10.1121/1.397861

Nearey, T.M., Assmann, P.F.: Modeling the role of vowel inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986). doi:10.1121/1.394433

Nittrouer, S.: Dynamic spectral structure specifies vowels for children and adults. J. Acoust. Soc. Am. **122**, 2328–2339 (2007). doi:10.1121/1.2769624

Stevens, K.N.: On the quantal theory of speech. J. Phonetics **17**, 3–45 (1989)

Story, B.H.: Speech simulation with an enhanced wave-reflection model of the vocal tract. Dissertation, Ph. D, University of Iowa (1995)

Story, B.H., Titze, I.R., Hoffman, E.A.: Vocal tract area functions from magnetic resonance imaging. J. Acoust. Soc. Am **100**, 537–554 (1996). doi:10.1121/1.415960

Story, B.H., Titze, I.R., Hoffman, E.A.: Vocal tract area functions for an adult female speaker based on volumetric imaging. J. Acoust. Soc. Am **104**, 471–487 (1998). doi:10.1121/1.423298

Story, B.H., Titze, I.R.: Parameterization of vocal tract area functions by empirical orthogonal modes. J. Phonetics **26**, 223–260 (1998). doi:10.1006/jpho.1998.0076

Story, B.H., Titze, I.R., Hoffman, E.A.: The relationship of vocal tract shape to three voice qualities. J. Acoust. Soc. Am. **109**, 1651–1667 (2001). doi:10.1121/1.1352085

Story, B.H.: Synergistic modes of vocal tract articulation for American English vowels. J. Acoust. Soc. Am. **118**, 3834–3859 (2005). doi:10.1121/1.2118367

Story, B. H.: A technique for ''tuning'' vocal tract area functions based on acoustic sensitivity functions. J. Acoust. Soc. Am. **119**(2), 715–718 (2006). doi:10.1121/1.2151802

Story, B.H.: Time-dependence of vocal tract modes during production of vowels and vowel sequences. J. Acoust. Soc. Am. **121**, 3770–3789 (2007). doi:10.1121/1.2730621

Story, B.H.: Comparison of Magnetic Resonance Imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002. J. Acoust. Soc. Am. **123**, 327–335 (2008). doi:10.1121/1.2805683

Story, B.H.: Vocal tract modes based on multiple area function sets from one speaker. J. Acoust. Soc. Am. **125**, EL141–EL147 (2009). doi:10.1121/1.3082263

Titze, I.R.: Parameterization of the glottal area, glottal flow, and vocal fold contact area. J. Acoust. Soc. Am. **75**, 570–580 (1984). doi:10.1121/1.390530

Titze, I.R., Story, B.H.: Acoustic interactions of the voice source with the lower vocal tract. J. Acoust. Soc. Am. **101**(4), 2234–2243 (1997). doi:10.1121/1.418246

Titze, I.R.: Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. J. Acoust. Soc. Am. **111**, 367–376 (2002). doi:10.1121/1.1417526

Titze, I.R.: The myoelastic aerodynamic theory of phonation. National Cent. Voice Speech **1**, 197–214 (2006)

Titze, I.R.: Nonlinear source-filter coupling in phonation: theory. J. Acoust. Soc. Am. **123**(5), 2733–2749 (2008). doi:10.1121/1.2832337

Westbury, J. R.: X-ray microbeam speech production database user's handbook. (version 1.0) (UW-Madison), (1994)

Part III
VISC in Different Populations of Speakers

Cross-Dialectal Differences in Dynamic Formant Patterns in American English Vowels

Ewa Jacewicz and Robert Allen Fox

Abstract This chapter provides evidence that vowel inherent spectral change (VISC) can vary systematically across dialects of the same language. The nature and use of VISC in selected "monophthongs" is examined in three distinct dialect regions in the United States. In each dialect area, the dynamic formant pattern is analyzed for five different age groups in order to observe cross-generational change in relation to specific vowel shifts and other vowel changes currently active in each dialect. The dialect regions examined included central Ohio (representing the Midland dialect), southeastern Wisconsin (representing the Inland North whose vowel system is affected by the Northern Cities Shift) and western North Carolina (representing the South whose vowel system is affected by the Southern Vowel Shift). Following a description of these dialect areas, we first introduce principles of chain shifting and the transmission problem, originally developed in the fields of sound change and sociolinguistics. Selective acoustic data are then presented for each dialect region and cross-generational patterns of vowel change are discussed. The chapter concludes that variation in formant trajectories produced between vowel onset and offset (VISC) is central to what differentiates regional variants of American English in the United States. Furthermore, a systematic variation in VISC is found in cross-generational change in acoustic characteristics of vowels within each dialect. The perceptual relevance of this acoustic variation needs to be addressed in future research.

Abbreviations

A0	Child speakers (aged 8 to 12 years)
A1	Youngest adult speakers (aged 19 to 34 years)

E. Jacewicz · R. A. Fox (✉)
Department of Speech and Hearing Science, The Ohio State University, Columbus, USA
e-mail: fox.2@osu.edu

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*,
Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_8,

A2	Young adult speakers (aged 35 to 50 years)
A3	Older adult speakers (aged 51 to 65 years)
A4	Oldest adult speakers (aged 66 to 88 years)
F1	First formant
F2	Second formant
DARE	Dictionary of American Regional English
NCS	Northern Cities Shift
SVS	Southern Vowel Shift
TL	Trajectory length
VISC	Vowel inherent spectral change
VSL	Vowel section length

1 VISC in Regional Variation in Vowels

There is a long phonetic tradition which views a vowel as a static target, i.e. a linguistic category whose position in the acoustic space (defined as a two-dimensional F1 × F2 plane) can be adequately characterized by the formant values at the vowel's putative steady-state (see Chap. Static and Dynamic Approaches to Vowel Perception). This approach was used in the classic study of the American English vowel system by Peterson and Barney (1952) that brought to light considerable variation in the position of the static target within each vowel category. Years later, this type of acoustic variation became of particular interest to sociolinguists studying regional dialects and language change manifested in vowel shifts and mergers (e.g., Labov 1994; Labov et al. 1972; Thomas 2001). Using the steady-state approach, the regional differences in the overall positions of nominal monophthongs in the F1 by F2 plane became apparent. However, the question of how vowels may differ cross-dialectally in terms of the extent and direction of dynamic formant movements has not been addressed so far.

Vowel inherent spectral change or VISC (Nearey and Assmann 1986) has been found in American, Canadian, and Australian English vowels (e.g., Andruski and Nearey 1992; Hillenbrand et al. 1995; Watson and Harrington 1999). Although the primary focus of work on VISC has involved its perceptual relevance for vowel identification (see Chap. Static and Dynamic Approaches to Vowel Perception, for an overview), these studies brought to light the possibility that VISC may vary systematically across regional varieties of English. If so, VISC may potentially play an important role in synchronic variation in vowels and sound change over time. However, no previous studies have examined whether VISC varies systematically across dialects. This chapter aims to provide some insights into the nature and use of VISC in three distinct regional varieties of English spoken in the United States. In the *Atlas of North American English* (Labov et al. 2006), these regions are labeled Inland North (the Great Lakes region), Midland (south of

the Inland North) and the South encompassing several southern states. These three broad dialect regions are defined on the basis of lexical, phonological and grammatical differences (Labov et al. 2006; Wolfram and Schilling-Estes 2006).

A particular striking feature in these regional dialects is that their respective vowel systems are affected by a set of changes termed "vowel shifts" (see Labov 1994; Labov et al. 2006 for extensive discussions and classifications of vowel shifts). According to Labov et al. (2006), the Northern Cities Shift (NCS) is a set of vowel rotations which affects currently northern American English spreading from Buffalo, NY through northernmost Ohio, Michigan, northern Illinois to southeastern Wisconsin and includes the Inland North. A different and much more complex set of changes termed the Southern Vowel Shift (SVS) is found in the South (some parts of this broad region are also defined by another shift called the Back Upglide Shift). Features of the SVS are found in the southeastern states of Virginia, North and South Carolina, Georgia and extend through Alabama, Louisiana, Oklahoma and parts of Texas, including also Kentucky, Tennessee and southern Ohio. Finally, the Midland is regarded as a transitional region between the Inland North and the South, encompassing central Ohio, part of Indiana, Missouri and Kansas. The Midland's vowel system is not reported as undergoing any distinct pattern of shift but there are other ongoing vowel changes in parts of this region.

In this chapter, we present acoustic data sampled from three relatively homogeneous speech communities situated within each of these broad dialect regions. Southeastern Wisconsin (the area between Madison, Milwaukee and Green Bay) was selected as a testing ground for the NCS. Central Ohio (Columbus and adjacent areas) was selected as representative of the Midland dialect. Western North Carolina (including Jackson and Haywood counties in the Appalachian region, an area that was labeled in earlier dialect atlases by Kurath (1949) and Kurath and McDavid (1961) as "Mountain Southern," was selected as the core area of the Southern Vowel Shift. In Labov et al. (2006), this region is now defined as the Inland South, the area in which the Southern Vowel Shift is most developed. The map in Fig. 1 shows the respective locations of these speech communities in each state.

The widespread occurrence of vowel shifts across regions in the United States have stimulated research in dialectology and sociolinguistics, asking questions as to why such vowel changes take place, how they originate and how are they transmitted within a given region to create a dialect or "regional variety" of the language. In our present examination of acoustic characteristics of vowels in these three dialect regions, we aim to illuminate the cross-generational pattern of vowel change with reference to VISC. Although we do not take a particular position with respect to the causal factors in vowel shifts (which involve a complex set of phonetic and social variables), we provide evidence that the extent of VISC changes with each younger generation of speakers from the same speech community. Therefore, cross-generational vowel change within each dialect carries a corresponding change in VISC and this systematic variation may be associated with a particular stage of a given vowel shift. Before we turn to our results, the next section will present a few remarks on the notion of "chain shifting" and its application to American English.

Fig. 1 The three testing areas in southeastern Wisconsin (WI, representing Inland North), central Ohio (OH, representing Midland dialect) and western North Carolina (NC, representing Inland South)

2 Chain Shifting and the Cross-Generational Sound Transmission

Since the earliest reconstructible stages (Proto-Germanic) the rich vowel systems of Germanic languages (including English) have been undergoing continuous change over time. This process has been termed "chain shifting" in descriptive analyses of historical linguists (Stockwell 1978). Based on phonological analyses of historical scripts involving distinctive vowel features such as high/low and front/back, it was found that one or more vowels can change position within the phonological system (a local change) and this change then gradually affects the vowel system as a whole, as other vowels are "pushed" away from the moving vowel(s) or "pulled" into newly opened positions in the vowel system to maintain earlier sets of distinctions. It is in this sense that vowels are thought of as "moving" in the vowel system in chains so that their individual relative positions change over time (i.e., across different generations of speakers) in a specific direction. In the subsequent linguistic phonetic studies, these relative positions have been measured and defined in the two-dimensional F1 by F2 plane, showing that a particular vowel can "rise" in the vowel space (such as from /æ/ to /ɛ/ which involves a decrease of F1 frequency) or "fall" (such as from /u/ to /o/ which involves an increase of F1 frequency).

In American English, the sociolinguistic work of Labov and his colleagues (notably Labov et al. 1972, 2006; Labov 1994) has pointed to the operation of three general principles of chain shifting across regions of the United States which are thought to motivate vowel shifts such as the NCS or SVS. These principles date back to Sievers (1876/1881) and have been treated since then as a central type of language change. Accordingly, in chain shifts, Principle I predicts that long

vowels rise, Principle II states that short vowels fall and Principle III predicts that back vowels will move to the front. Labov restates these principles in more modern terms in his *Peripherality Hypothesis*, which proposes that in chain shifts, tense vowels (which have longer durations and more extreme articulatory positions) move upward along a peripheral track and lax vowels move downward along a non-peripheral track (see Labov et al. 2006, pp. 15–20, for further details, including an effort to formulate a single principle of chain shifting).

In our present cross-dialect work, we expect to find at least some indication of the operation of these general principles of chain shifting. In particular, we examine positional changes of the three short vowels /ɪ, ɛ, æ/ in the production of five generations of speakers from each dialect area in order to evaluate the operation of Principle II. However, it needs to be emphasized that the principles of vowel change were formulated on the basis that vowels are "static targets" and the only change involves their positional movement in the vowel system. No previous research indicated that vowel change may also involve "intrinsic" spectral properties such as VISC. Therefore, the second aim of our research is to explore how changes in VISC may be manifested across generations of individuals who grew up in the same dialect area as both speakers and listeners of the local dialect.

The logical question arises as to how the vowel changes are transmitted through multiple generations always shifting in the same direction. The transmission problem cannot be confined to a single generation of speakers. Indeed, Stockwell (1978) has identified the "perseverance problem" (how successive generations keep moving vowels in the same directions) as a central issue in sound change. Studying synchronic variation in vowels, sociolinguists overwhelmingly point to the role of social factors in sound transmission such as social class, age, gender, ethnicity, neighborhood and social networks (see Labov 2001 for extensive discussion). In terms of purely phonetic effects, consonantal contexts (usually consonants in the syllable coda) are identified as a primary aspect in segmental conditioning of positional vowel change. In our work, we admit the possibility that cross-generational vowel change (both in terms of positional change and change in VISC) may be, in part, a direct effect of prosodic organization of English language and its specific use of linguistic stress (see Jacewicz et al. 2006, for extensive discussion). Because vowels in stressed syllables are longer and exhibit a greater spectral change compared to reduced vowels in unstressed syllables, they may trigger a shift-like change in that younger generations acquire as their norm those variants which were produced with greater emphasis a generation earlier. To explore these effects, our present acoustic analysis includes both emphatic and nonemphatic vowel variants produced by each successive generation of speakers.

What follows is a report on a few results of a large-scale acoustic study undertaken to examine cross-generational changes in vowels in the three selected dialect regions. The study is conducted in "apparent time" (as opposed to "real time"). From a strictly technical point of view, an apparent-time study is a specific type of cross-sectional study. The apparent-time methodology considers the speech samples of different generations of speakers collected during a single time period to be representative of different "stages" of the dialect and not simply

cross-sectional differences as a function of age (see Bailey et al. 1991). Apparent-time studies are a common practice in sociolinguistic research because of obvious problems in conducting a study in a "real time." Real-time studies require collection of samples from the population of interest over an extended stretch of time (usually, decades). These studies would include longitudinal studies (data collected from the same individuals across their life span) and historical studies (e.g., comparing older speech recordings, e.g., from the *DARE* project (*Dictionary of American Regional English*, Cassidy and Hall 1985), to recent speech samples. There are obvious challenges to the "real time" approach, either in terms of the time commitment and/or poor quality of older recordings. However, the results of apparent-time studies prove to be generally reliable if speech communities remain largely the same over the course of time (Chambers 2003).

In each testing location, five generations of speakers were recorded who were born, raised and spent most of their lives in the respective regions. The participants fell into five age groups (A0–A4) whose ages (in years) were: children 8–12 (A0), youngest adults 19–34 (A1), young adults 35–50 (A2), older adults 51–65 (A3) and oldest adults aged 66 and up to the late 80s (A4). There were both males and females in each age group for a grand total of about 400 speakers. The number of speakers varied from 9 to 16 per gender/age/dialect subgroup, depending on availability of subjects within the time frame of the project. Each participant produced a set of single words in citation form, a set of sentences with a variable main sentence stress and a spontaneous talk. The speech material was recorded in the years 2006–2008.

In this chapter, we will report mainly on the results for three vowels /ɪ, ɛ, æ/ which were produced by 198 female speakers (66 per dialect) in a sentential context in stressed and unstressed syllables in a/b_dz/ environment (the words were *bids*, *beds*, *bads*). The vowels in stressed syllables are referred to here as emphatic and those in unstressed syllables as nonemphatic to underscore the fact that such differences are not always related to linguistic stress per se but the emphasis can be produced by a variety of phonetic and paralinguistic factors. Female speakers were chosen to simplify the presentation only and the data from male speakers produce similar patterns. Each speaker read two randomly presented repetitions of a sentence containing the word of interest in each emphasis position for a grand total of 2376 vowel tokens (66 speakers × 3 vowels × 2 emphasis positions × 2 repetitions × 3 dialects). Our long-term working hypothesis is that the emphatic vowels (rather than nonemphatic) will lead the sound change across generations in a specific direction.

Emphatic vowels are expected to occupy a more peripheral position relative to nonemphatic vowels due to the expansion of the vowel space as a function of emphasis (e.g., Moon and Lindblom 1994). Also, emphatically spoken words "tend to be articulated more forcefully, resulting in longer and more extensive vowel […] gestures" (Agwuele et al. 2008, p. 207) which suggests a greater formant movement in emphatic vowels. In this chapter, we will observe whether and how formant dynamics change cross-generationally in both emphatic and nonemphatic variants in each dialect region. Our present interest focuses on two

cross-generational changes: (1) differences in the relative positions of the vowels in the acoustic space and (2) variation in the extent of formant movement (or the amount of VISC). For each dialect, we expect that vowels will change position in the vowel space as a function of speaker generation. What we cannot predict is the extent of the corresponding change in formant dynamics or whether dialects differ in the use of VISC cross-generationally.

3 The Midland Vowel Changes in Central Ohio

We begin with a presentation of the data for the Midland dialect in central Ohio, which is not thought to be participating in any known chain shift at present. The panels in Fig. 2 show cross-generational plots across all age groups (A4–A0) for a total of 66 female speakers. The data points in the plots indicate mean F1 and F2 values measured at five equidistant temporal locations corresponding to the 20–35–50–65–80 %-point in the vowel. These multiple measurement points allow us to estimate the extent of formant movement spanning over the central 60 %-section of the vowel to the relative exclusion of immediate effects of surrounding consonants on vowel transitions. Although this approach uses five points (rather than the commonly used onset and offset measurements), it still only provides an estimate of the shape of the actual trajectory. Yet, as we will see, the five-point measurement technique is sufficient enough to capture the basic variation in VISC across dialects and ages (along with the variation due to vowel emphasis) and is relatively easy to implement in analyzing a larger corpus. More details pertaining to this analysis can be found in Fox and Jacewicz (2009). In the current plots, we follow the sociophonetic tradition in displaying the vowels in the F1 by F2 plane in which the axes show values in descending order. Direction of formant movement is indicated by arrows.

Turning to the plots, we see substantial variation in formant dynamics across all three individual vowels each of which is classified as a nominal monophthong in American English. Formant trajectories for the emphatic and nonemphatic versions of all three vowels are relatively close and parallel when produced by oldest speakers (A4). With each successively younger generation, the emphatic and nonemphatic variants of the vowels become progressively more separated one from the other. This represents a change in the basic position of the vowels in the acoustic space. A second difference can be seen in terms of a change in vowel dynamics. As the relative position of the vowel changes across generations, so does the amount of VISC. With each younger generation, each vowel exhibits less formant movement (especially in F2) becoming progressively less diphthongal. The differences between A4 and A0 groups are rather drastic and include also the direction of formant movement. Although emphatic vowels are consistently more peripheral than nonemphatic across all generations, the most salient cross-generational differences seem to lie in the amount of VISC.

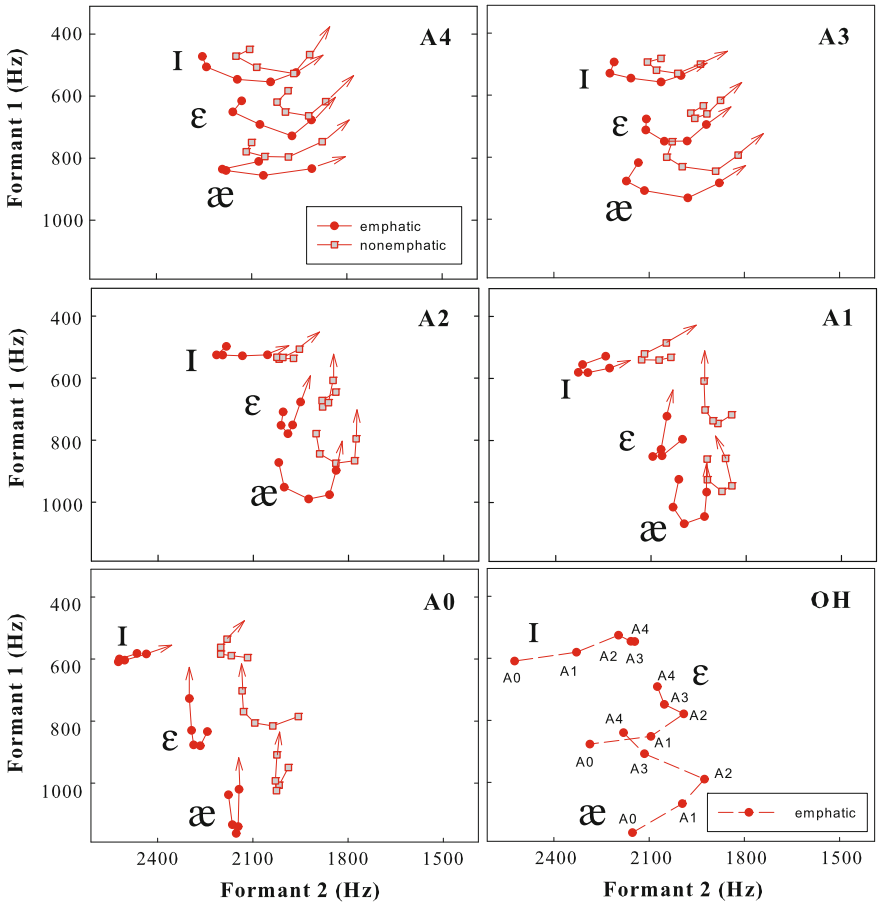


Fig. 2 Means of F1 and F2 measured at five equidistant timepoints in a vowel across five age groups (A4–A0) for the central Ohio dialect. Positional changes of vowel midpoints are shown in the last panel (*bottom right*)

In an initial attempt to quantify variation in the size of the VISC across generations and vowels, we calculated a trajectory length (TL) in the F1 by F2 acoustic space for each vowel. The TL represents a sum of the lengths of the four separate vowel sections between the 20 and 80 %-point, where the length of one vowel section (VSL) is:

$$VSL_n = \sqrt{(F1_n - F1_{n+1})^2 + (F2_n - F2_{n+1})^2}$$

The assumption is that a longer TL reflects a greater amount of VISC. Shown in Table 1 are the mean TLs for the Ohio speakers. Although there is some variability in the means, there is an overall tendency for the TLs for younger speakers to be shorter than those for older speakers, suggesting that the amount of VISC changes

Table 1 Mean trajectory length (in Hz) for Ohio speakers broken down by vowel, emphasis condition and age group; standard deviations are in parentheses

Vowel	Emphasis condition	Age group				
		A4	A3	A2	A1	A0
/ɪ/	Emphatic	440 (116)	394 (70)	353 (101)	344 (75)	320 (110)
	Nonemphatic	390 (106)	306 (91)	253 (117)	251 (86)	223 (90)
	Combined	415 (111)	350 (95)	303 (119)	303 (94)	272 (110)
/ɛ/	Emphatic	456 (104)	423 (95)	346 (101)	378 (67)	353 (116)
	Nonemphatic	313 (62)	287 (107)	244 (80)	270 (79)	308 (147)
	Combined	38 (111)	355 (121)	295 (104)	324 (91)	330 (131)
/æ/	Emphatic	593 (105)	538 (123)	435 (137)	469 (181)	430 (149)
	Nonemphatic	419 (114)	380 (121)	331 (110)	339 (155)	303 (95)
	Combined	506 (139)	459 (145)	383 (131)	404 (177)	367 (139)

across generations. Also, the mean TLs vary with vowel category: they tend to be longer for the vowel /æ/ and shorter for either /ɪ/ or /ɛ/, which do not seem to differ much from one another. Across all vowels and groups, the mean TLs of the emphatic vowels are longer than of the nonemphatic.

In the final panel of Fig. 2, the formant values at the 50 %-point (often considered as representing the *vowel nucleus* in the sociophonetic literature) of the emphatic vowels are replotted for an immediate display of positional changes across generations. We find a progressive fronting of /ɪ/ with each younger generation. The pattern for /ɛ/ is clearly different in that the vowel first lowers, is then fronted (A1) and shows further fronting when produced by children. The greatest positional change is in /æ/ which not only consistently descends across generations but displays considerable backing and then fronting in young adults and children. Given the ages of the speakers in the A0 group (8–12 year-old girls; mean age for the A0 group was around 10 in all three dialect groups) we would expect that their generally shorter vocal tracts would produce some elevation of both F1 and F2 compared with the adult female speakers. However, these differences may not be as substantial as those between boys and adult males (see Lee et al. 1999; Assmann et al. 2009) and, given the overall direction of the positional differences between the A0 and A1 groups, children seem to continue the path of vowel change seen in the adult cross-generational data.

How do these cross-generational changes relate to the general principles of chain shifting? As pointed out, there are no attested vowel chain shifts in central Ohio. We thus do not expect the operation of Labovian Principle II according to which the non-peripheral (lax) vowels move downward along a non-peripheral track. However, we see specific changes which seem to parallel those in accordance with Principle II. For example, the vowel /æ/ seems to have descended to the bottom of the non-peripheral track and is being pressured to move further. Will it enter the lower peripheral track and, being fronted, will it start rising and breaking to introduce the Northern Cities Shift to this geographic area? Only time can answer this and new data from new generations of Ohioans. Similarly, the vowel

/ɛ/ follows the general direction of /æ/. Does its fronting in children's production signal a reorganization of vowel subsystem in this corner of the vowel space or will it return to its position as produced by children's grandparents? If so, will it develop a greater formant movement as generations ago? Although not related to Principle II, does the fronting of /ɪ/ indicate a more general process of fronting in American English (such as in response to the back vowel fronting) or is this cross-generational change specific to central Ohio? Further work is needed to explain these interesting patterns.

4 Northern Cities Shift in Southeastern Wisconsin

While vowel changes in central Ohio do not appear to be part of a chain shift, the NCS reflects an attested chain shift, first reported in late 1960s and widely recognized since Labov et al. (1972). Before presenting our Wisconsin data, we summarize briefly the main characteristics of NCS. The triggering event for the NCS is the raising and fronting of /æ/. This event initiates a series of the following vowel rotations: fronting of /ɔ/, lowering of /ɑ/, lowering and backing of /ɛ/, backing of /ʌ/, and some centralization and backing of /ɪ/. In this way, the vowels exchange their positions like in a chain, operating along the peripheral track (/æ, ɔ, ɑ/) and non-peripheral track (/ɛ, ʌ, ɪ/) until the entire cycle is complete. The exact order of the stages in the chain in the NCS is subject to some inter-speaker variation (see Labov et al. 2006, for further details) but the general rotation is assumed to be clock-wise as shown, schematically, in Fig. 3. Some variation in directionality also occurs within this pattern, as shown by Gordon (2001). Most notable for present purposes is that /ɛ/ can move down rather than back.

Our results for three vowels /ɪ, ɛ, æ/ come from 66 Wisconsin female speakers who produced exactly the same speech material as Ohio speakers and were tested under the same experimental conditions. The recordings were made at the University of Wisconsin-Madison. Figure 4 displays plots across all age groups (A4–A0). The oldest speakers (A4) show the characteristic mark of NCS, i.e., the raising and fronting of /æ/, particularly in its early portion before vowel midpoint (known as Northern breaking). This vowel exhibits an extended formant movement, which is also true for the other two vowels in A4 productions, /ɪ/ and /ɛ/. In the next generation (A3), while /æ/ does not change much in general, we see a

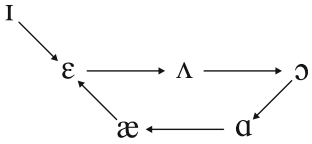


Fig. 3 Schematic rotation of vowels in the Northern Cities Shift indicating a change in their pronunciation. "bid" → "bed" → "bud" → "bawd" → "bod" → "bad" → "bed"/"bid"

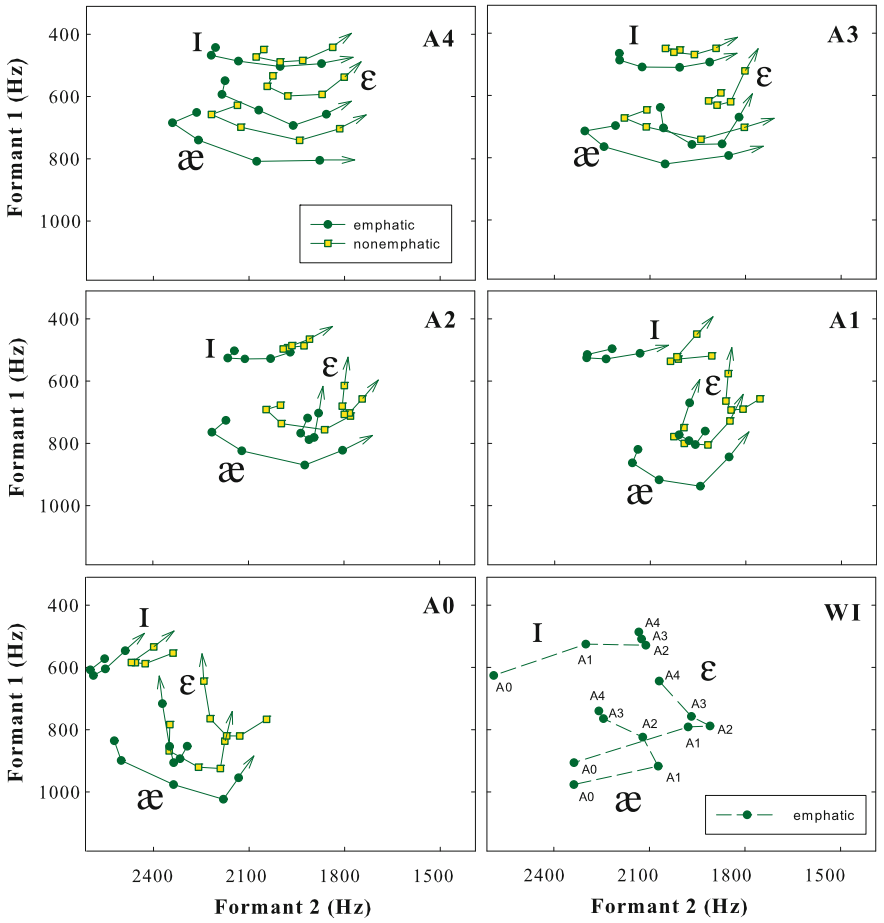


Fig. 4 Means of F1 and F2 for the southeastern Wisconsin dialect along with positional changes of vowel midpoints (*bottom right*)

substantial lowering and backing of /ɛ/, particularly in the stressed position so that the emphatic /ɛ/ tends to overlap with the nonemphatic /æ/. The next generation (A2) introduces a significant change in that /ɛ/ not only moves further downward but it changes its acoustic characteristics. As can be seen, the extensive formant movement found in earlier generations is greatly reduced, especially in emphatic positions. The "monophthongization" of /ɛ/ may signal a later stage of the NCS in this geographic area to be followed only by backing of /ʌ/ and shifting of /ɪ/ in response to the movement of /ɛ/. In the next generation (A1), we see a further decrease in formant movement, particularly in /ɪ/ (which also undergoes fronting) and to some extent in /æ/, which also lowers. Children's vowels (A0) generally maintain the direction shown in young adults (A1) although their vowels have a rather high F2 which can be in part due to their unnormalized formant values.

Table 2 Mean trajectory length (in Hz) for Wisconsin speakers broken down by vowel, emphasis level and age group; standard deviations are in parentheses

Vowel	Emphasis condition	Age group				
		A4	A3	A2	A1	A0
/ɪ/	Emphatic	478 (94)	419 (100)	377 (127)	385 (93)	374 (119)
	Nonemphatic	311 (129)	280 (125)	265 (66)	308 (85)	300 (122)
	Combined	395 (140)	348 (131)	321 (115)	347 (95)	337 (124)
/ɛ/	Emphatic	503 (136)	467 (146)	350 (121)	364 (113)	399 (148)
	Nonemphatic	369 (118)	308 (108)	252 (55)	270 (60)	416 (176)
	Combined	436 (141)	387 (150)	301 (105)	317 (100)	407 (160)
/æ/	Emphatic	694 (111)	658 (156)	611 (121)	516 (124)	607 (216)
	Nonemphatic	562 (196)	518 (159)	423 (171)	362 (104)	421 (155)
	Combined	606 (161)	588 (171)	517 (174)	439 (136)	514 (208)

As was done for the Ohio data, mean TLs were calculated for each vowel across all Wisconsin speakers. These means are shown in Table 2. While pattern of TLs of both /ɪ/ and /ɛ/ is not much different from that seen in the Ohio dialect, the TLs for /æ/ are longer for Wisconsin than Ohio speakers which suggests a greater extent of VISC in Wisconsin /æ/.

In terms of cross-generational changes, unlike in the Ohio data, each Wisconsin vowel showed a somewhat different pattern of variation in TL. For /ɪ/, mean TLs did not differ substantially across the age groups. For /ɛ/, TLs were longest in A4, dropped sharply in A2 but then increased progressively in A1 and A0. Finally, the vowel /æ/ revealed yet another pattern. Mean TLs for the emphatic variant remained approximately the same across all generations, with some drop in A1. However, there were more changes in the nonemphatic variant: the longest mean TLs were in A4 and then decreased progressively across the younger age groups until reaching the shortest values in A1. They increased again in A0.

How do these cross-generational vowel changes relate to the NCS? If we track the movement of vowel nuclei only (the 50 % point) as displayed in the last panel of Fig. 4, we see the raised /æ/ in A4 and its progressive lowering and backing across next generations along with an unexpected fronting in children. While the raised /æ/ is clearly a mark of the NCS, its subsequent downward movement is not. However, in terms of the *Peripherality Hypothesis*, the vowel could have reached its highest possible position in the peripheral track and entered a non-peripheral track which would explain its lowering and some backing in accord with Principle II. This, of course, is not the expected progression of the vowel in the chain, which should continue rising to approximate the positions of /ɛ/ and /ɪ/ with each younger generation. It could be the case that such vowel rotations can be observed in other consonantal contexts since contextual variation is a possible conditioning factor. The cross-generational data from the present speakers in this particular consonantal context do not indicate a continuous rising of /æ/ in this part of Wisconsin. The lowering and backing of /ɛ/ and /ɪ/ across A4, A3 and A2 groups is entirely in agreement with the direction of the NCS (and the Principle II). However, younger

generations (A1 and A0) introduce a directional change to this pattern (i.e., fronting). It is unclear how this change relates to the NCS and this issue can only be solved by data from future generations.

Operation of the NCS has not been definitively linked to changes in formant dynamics aside from observations about the ongliding of /æ/. Yet, based on the present results, the cross-generational changes in VISC are apparent. Some changes are more abrupt than others such as those in the emphatic /ɛ/ and to a lesser extent in /ɪ/ introduced by A2 speakers. Namely, in the process of lowering and backing, the extent of formant movement of /ɛ/ undergoes a drastic reduction in comparison to A4 along with a change in directionally of movement. A relatively smaller reduction in the movement of /ɪ/ prepares the stage for a greater change in F2 the next generation (A1) along with its progressive fronting. Is such variation in VISC linked to the particular stages of the shift? Certainly, formant values measured traditionally at the vowel's midpoint cannot reveal such potentially complex relationships and we do not know whether sudden changes in formant dynamics precondition stages in vowel shifting. The general principles of chain shifting were formulated on the basis of the assumption that the nominal monophthongs do not exhibit formant movement.

5 Southern Vowel Shift in Western North Carolina (Inland South)

The third set of cross-generational data included in this chapter pertains to the operation of the chain shift called the SVS which defines the dialects of the South. As indicated earlier, our participants come from the Appalachian area in western North Carolina which is identified as a center of the most advanced features of the southern vowel system (Labov et al. 2006). The present data come from 66 female speakers who were born and raised in the area, produced the same speech material and were tested under the same experimental conditions. The recordings took place at Western Carolina University in Cullowhee, NC. In this section, we first present cross-generational data for the vowels /ɪ, ɛ, æ/ in order to be consistent with the previous reports for Ohio and Wisconsin. We then discuss the operation of the SVS in a greater detail in a set of vowels produced in citation form.

Figure 5 displays plots across all age groups (A4–A0) for 66 female speakers. The trajectory shapes of vowels produced by A4, A3 and A2 speakers indicate a type of formant movement typical of the "Southern drawl" or breaking, where the vowel "breaks" into two parts (Sledd 1966). That is, the vowel in *bit* may be pronounced as [bi:jɪt] giving an auditory impression of two vowels in a sequence connected by the glide [j]. Note the proximity of /ɪ/ and /ɛ/ in the production of A4 and A3 speakers and a continuous separation of all vowels (including their emphatic and nonemphatic variants) with each younger generation. The last panel (bottom right) tracks the positional changes of vowel midpoints, indicating that the

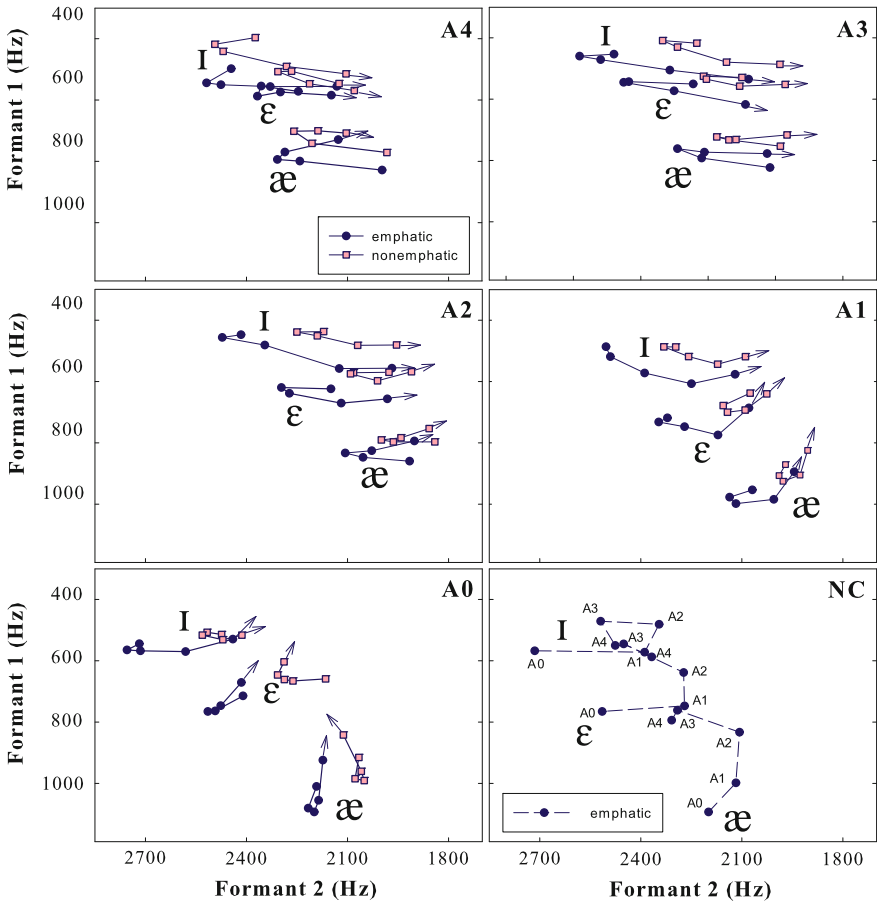


Fig. 5 Means of F1 and F2 for the western North Carolina (Inland South) dialect along with positional changes of vowel midpoints (*bottom right*)

vowels /ɛ/ and /æ/ begin to descend in the acoustic space in the production of A2 speakers while /ɪ/ shows a less decisive pattern. Table 3 shows a summary of TL values across age groups. Notable differences from both Ohio and Wisconsin data include generally longer TLs for North Carolina speakers, indicating greater formant movement. Also, it is the A3 group (and not A4) that shows the greatest extent of VISC.

A trend common in all three dialects (including the Inland South) was that TLs for emphatic variants were significantly longer than for nonemphatic; also, TLs were shorter for the younger than for the older speakers. However, of greatest interest in examining the North Carolina vowels is not only the extent to which these three front vowels have changed cross-generationally; the operation of the SVS involves two additional essential vowels, /i/ and /aɪ/. Unfortunately, the North Carolina speakers were not asked to produce the vowel /i/ in sentences as variation

Table 3 Mean trajectory length (in Hz) for North Carolina speakers broken down by vowel, emphasis condition and age group; standard deviations are in parentheses

Vowel	Emphasis condition	Age group				
		A4	A3	A2	A1	A0
/ɪ/	Emphatic	623 (175)	717 (166)	694 (135)	587 (124)	541 (210)
	Nonemphatic	607 (169)	520 (138)	437 (156)	372 (209)	375 (232)
	Combined	616 (156)	619 (180)	565 (194)	480 (200)	458 (234)
/ɛ/	Emphatic	557 (178)	730 (232)	614 (179)	504 (152)	449 (190)
	Nonemphatic	474 (126)	441 (122)	422 (145)	337 (128)	293 (102)
	Combined	515 (155)	586 (234)	518 (188)	421 (161)	371 (170)
/æ/	Emphatic	662 (213)	713 (255)	552 (186)	443 (176)	460 (166)
	Nonemphatic	489 (167)	495 (180)	392 (155)	323 (71)	388 (188)
	Combined	576 (206)	604 (244)	472 (187)	383 (144)	424 (178)

in this vowel was not of interest to the study as a whole (i.e., the vowel is not a part of NCS-rotations nor are there any changes to /i/ in central Ohio). For this reason, we now turn to a different set of acoustic measurements taken from citation form words in the *hVd* context, which were produced by all participants of the study.

The SVS is a more complex chain shift than the NCS and is manifested somewhat differently in different parts of the South. Generally, the triggering event of the SVS (Stage 1) is the deletion of the offglide in /aɪ/ and its monophthongization so the vowel is pronounced basically as an /a/. Stage 2 is the centralization and lowering of /e/ and fronting and raising of /ɛ/, the so called reversal of the front/back locations of /e/ and /ɛ/. Stage 3 is a parallel lowering of /i/ and fronting and raising of /ɪ/ which results in another reversal of the front/back locations of /i/ and /ɪ/. The upper left panel in Fig. 6 captures Stage 3, the most advanced stage of the SVS, in the production of A4 speakers. For clarity of presentation, vowel /e/ is not included in the plots in Fig. 6 as it has considerable overlap with /ɪ, ɛ, æ/.

We will focus our discussion on the subsequent cross-generational reorganization of the vowel subsystem affected by the SVS. First, we find the oldest speakers producing a monophthongal version of the diphthong /aɪ/. Their vowels /i/ and /ɪ/ are in close proximity which may be a reflection of their earlier reversal. A4 speakers also produce the raised and fronted /ɛ/ and /æ/. Beginning with the A3 generation, there is progressive re-diphthongization of the monophthongal /aɪ/, progressive fronting of /i/ and lowering and backing of /ɪ, ɛ, æ/. Finally, the young adults (A1) have a very fronted /i/, a clear diphthong /aɪ/, and no breaking in /ɛ/ and /æ/ although both vowels still show a considerable amount of formant movement. The vowel /ɪ/, on the other hand, does remain diphthongized and is undergoing fronting as if it were "pulled" by /i/. The children's data show a continuation of this trend: both /ɛ/ and /æ/ are slightly lowered and less diphthongized and /ɪ/ still shows a substantial amount of spectral change. However, both /i/ and /ɪ/ are separated one from another. Will /ɪ/ undergo a further reduction in VISC in the next generation from this area and will the existing subsystem be a subject to further reorganization? Only time will allow us to answer these questions.

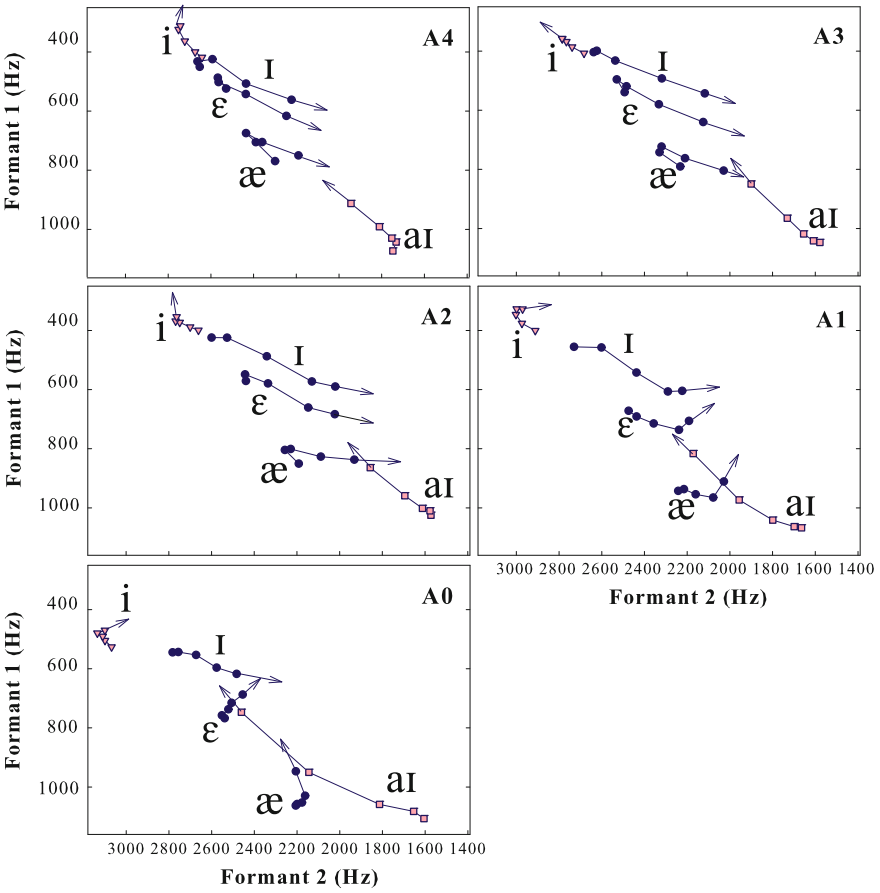


Fig. 6 Means of F1 and F2 for subset of vowels (produced in citation form *hVd* words) in the Southern Vowel Shift in western North Carolina

The complexity of changes to the vowel subsystem affected by the SVS can only be appreciated when we examine the cross-generational changes in VISC. Clearly, formant dynamics change across generations of North Carolina speakers and the vowel system of the youngest ones does not look like the system of speakers from older generations. These data provide new support for the view that the systemic reorganization occurs gradually, generation by generation, and particular vowels change their acoustic characteristics in terms of both formant frequencies and the extent of spectral change.

6 Cross-Dialectal Variation in Formant Dynamics of /æ/

The final set of observations about the dialect-specific use of formant dynamics pertains to vowel /æ/. The case of /æ/ is of interest to us for the following reason. We found a progressive lowering and backing of the Ohio variant with each younger generation which also corresponds to a reduction in its formant movement. Lowering and backing was also found in Wisconsin /æ/ whose formant movement (at least in the emphatic variant) did not generally change across generations. How do these patterns compare to the third variant of /æ/, that spoken in the South, which is generally considered to demonstrate "Southern raising?" In our data, southern /æ/ also lowers with each younger generation following the path of the two other dialects. If it is the case that /æ/ lowers cross-dialectally in accord with Principle II, what would be the corresponding change in the extent of VISC for each of the three regional variants?

The plots in Fig. 7 show the three emphatic variants of /æ/, one for each dialect, across all five age groups redrawn from Figs. 2, 4, and 5. As can be seen, there are positional differences among the three variants of /æ/ which also show three distinct patterns of formant movement. Wisconsin /æ/ is the most raised and the most fronted, Ohio /æ/ is the lowest, and North Carolina /æ/ is positioned in between these two. All three dialect variants are raised in the production of the oldest generation and descend in the acoustic space in a steady fashion with each younger generation. Of particular interest are the cross-generational changes in the nature and amount of formant movement.

All three variants are heavily diphthongized in A4 speakers and we can see three distinct patterns of /æ/-production: the Wisconsin variant has its nucleus raised due to the NCS, the North Carolina variant indicates the Southern breaking (notice the difference in the directionality of formant movement between the northern and the southern variant) and the Ohio /æ/ does not show any raising of its nucleus (and no influence of the NCS) although its direction of movement is as in the Wisconsin variant. This general pattern is maintained in the productions of the three older generations and then changes rather abruptly in young adults (A1). In particular, Wisconsin /æ/ has a smaller extent of formant movement and its nucleus is not raised as much as in the previous generations, North Carolina /æ/ changed its directionality so that the Southern breaking is no longer present, and Ohio /æ/ shows a spectral change in F1 rather than F2, which is opposite to what we see in A4, A3 and to some extent in A2 speakers. Finally, while Wisconsin children's /æ/ still shows the NCS influence, the North Carolina and Ohio variants changed greatly so that the only spectral change is in F1.

Clearly, these three dialect variants differ in the nature and the extent of VISC although the common trend for /æ/ is to lower in the acoustic space with each younger generation in accord with Principle II. This also includes the southern variant which is assumed to be raised in the South. It is unclear why emphatic variants in both Ohio and North Carolina change their dynamics across generations while the Wisconsin vowel does not. The principles of chain shifting seem to have

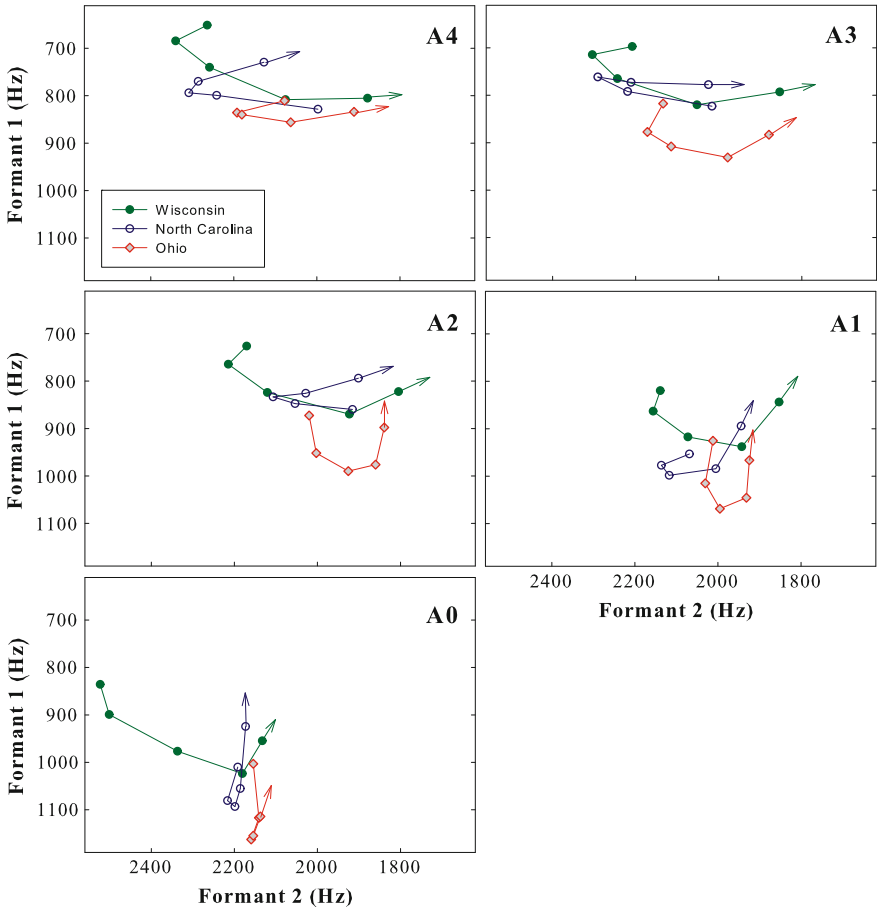


Fig. 7 Means of F1 and F2 for the emphatic variants of /æ/ across dialects and age groups

reached their explanatory limit in this case and a deeper study of cross-generational variation in VISC in each of the three vowel systems is most likely to increase our understanding of such patterns.

7 Summary of Regional Variation in Formant Dynamics

The cross-dialectal and cross-generational data presented in this chapter underscore the richness of acoustic information between the onset and offset of a vowel. Changes in the shapes of the formant trajectories during the production of a vowel are precisely what differentiates dialects and generations of speakers across regions in the United States. Vowels in American English change their acoustic

characteristics over time as new generations of speakers are born and acquire vowel systems from available linguistic input. While sociolinguists have focused on the directionality of vowel movement in the acoustic space (as determined by the formant values at the syllabic nucleus), the variation in dynamic vowel characteristics has not been addressed so far as a dialect feature.

The use of multi-point measurement of the whole trajectory, as in the present approach, helps us realize that even the most monophthongal vowel is rarely static, i.e., its formant shapes do not resemble straight lines which would be plotted in a F1 by F2 plane as a single data point. Rather, spectral change includes at least minimal variation in either F1 or F2 or a combination of both. Some additional variation in VISC comes from phonetic sources such as prosodic effects which may affect the shape of the whole trajectory as shown in the present data. Obviously, consonantal environment introduces the most predictable type of changes. However, there is also another powerful source of variation in VISC, that of the dialect-specific use of the dynamic information in a vowel. A good example here is the cross-dialectal variation in the trajectories of /æ/ whose shape and direction of movement may signal Southern breaking, reflect the first link in the NCS, or follow yet other patterns, as in central Ohio. What we also find across the dialects is that the extent of VISC changes cross-generationally. This change is most likely related to dialect-specific chain shifts or other vowel changes which we find in regional vowel systems. Although the principles of chain shifting have been formulated on the basis of positional changes of vowel nucleus, there seems to be a correspondence between a change in the extent of VISC and the positional change of a vowel related to the reorganization of a given subsystem.

As hypothesized earlier (Jacewicz et al. 2006), phonetic stress may play a role in chain shifting. Emphatic vowels have a potential to lead the vowel change due to their enhanced acoustic characteristics on which children may focus in the process of cross-generational sound transmission. The present data suggest that emphatic variants indeed "pull" vowels in specific directions over time, although this possibility must be explored further in future studies. The present data clearly show a substantial difference in TL values between the emphatic and nonemphatic variants. For all vowels and all age groups, the emphatic vowels have longer TLs than nonemphatic vowels and the only exception is the vowel /ɛ/ in Wisconsin children. Although longer TLs for the emphatic variants were expected, the cross-generational changes in the extent of VISC in these emphatic variants were not. As already pointed out, the present data suggest a type of coordinated systemic changes associated with reorganization of a vowel subsystem which include both positional changes and changes in the extent of VISC, which are most readily observed in the emphatic variants. However, more vowels (including back vowels) need to be studied and more work needs to be done to understand this complex relationship.

8 Dialectal Spectro-Temporal Variation and Vowel Perception

We used the TL measure in this chapter to compare the total trajectory change (restricted to the 5-point measurement) across dialects and ages. Although, indisputably, a greater number of measurement points can produce a more accurate, veridical representation of the formant trajectory, a 5-point measurement system can still produce a good estimate of the actual trajectory length and a reasonable characterization of the trajectory shape. However, the TL measure fails to account for the change in the directionally of movement. For example, while Wisconsin and North Carolina /æ/ in A4 group have comparable TL values (606 and 576 Hz, respectively), their formant trajectories move in opposite directions. A similar problem occurs when the angle of formant movement changes in the course of vowel duration.

Also, while the TL measure provides information about the cumulative size of the formant changes, it fails to account for the speed of these changes as it has no true temporal component. However, there may be important dynamic differences across dialects and speaker age that relate to how quickly (or slowly) these spectral changes are made. The spectral rate of change measure, also included in Fox and Jacewicz (2009), has been shown to be quite effective in addressing the cross-dialectal variation in VISC when restricted to a single age group. However, caution and further modeling is needed when applying the spectral rate of change measure to cross-generational data because vowel inherent duration may be confounded by several factors including articulation rate, aging effects and dialect differences.

As it turns out, vowel duration is also subject to regional variation and systematic differences in vowel duration have been found for the three dialect regions studied here (Jacewicz et al. 2007). Moreover, the dialect differences in articulation rate also proved to be significant. In particular, articulation rate (excluding pauses) in the Wisconsin speech is faster than in North Carolina speech for both young and older adults apart from the aging effects (Jacewicz et al. 2009a). Given that vowels are significantly longer in North Carolina speech and articulation rate is slower, one would assume a straightforward relationship between these two. That this is not the case and the slower articulation rate does not imply a change "across the board" in temporal properties of segments has been shown in a study of stop closure voicing for these two dialects (Jacewicz et al. 2009b). Namely, the stop closure duration was found to be longer in Wisconsin and not in North Carolina speech. This would suggest that vowel duration is dialect-specific as is the nature of formant movement along with positional relations among the vowels. Further modeling is needed to relate vowel-inherent duration to temporal variation as a function of dialectal and cross-generational changes in vowels.

The set of data presented in this chapter lends support to the conclusion reached earlier by Cox (1999) who studied vowel change in Australian English. She points to the

> Changeable nature of language and the fact that specifications of formant structures are only valid for a particular dialect at a particular time in history. The systemic nature of vowel change is clearly documented as well as the close interdependent relationships between the monophthongs and diphthongs. Change in one class can be seen to affect the other in a parallel fashion in this dialect of English (p. 20).

In our view, the mechanism of language change utilizes the variation in VISC in ways which are not yet well understood. However, many questions arise as to the perceptual relevance of this type of synchronic variation in formant dynamics to vowel identification. Although the present apparent time data show cross-generational changes in vowel characteristics, we need to bear in mind that the speakers used in this study have been born and lived in the same speech community most of their lives. This suggests that both adults and children, despite differences in their respective productions, have been exposed as listeners to diverse dynamic cues in vowels including a variety of shapes of their formant trajectories. How will this experience with dialect-specific features affect their ability to make vowel identification decisions? Will spectral information between vowel onset and offset be largely ignored, will it be helpful or will it be critical in identifying specific vowels? How sensitive are these listeners to cross-generational changes in VISC and do they perceive such variation at all?

A set of related questions can be asked about listeners growing up in one dialect area who have never been in contact with features of another dialect. Will formant dynamics of vowels from another dialect influence their perceptual response in a listening task or will the attunement to their own dialect guide their identification choices? What will be the confusion pattern and what will it tell us about listeners' use of spectral dynamics in vowels from their non-native dialect? Will they manifest sensitivity to cross-generational variation in VISC in their non-native dialect? These and other questions await answers in future research.

Acknowledgments This work was supported by the grant R01 DC006871 from the National Institute of Deafness and Other Communication Disorders, National Institutes of Health. We thank Joseph Salmons for his contributions to this research and his comments on this chapter. Special thanks go to the personnel of the Speech Perception and Acoustics Labs at Ohio State as well as our collaborators at the University of Wisconsin-Madison (Dilara Tepeli) and Western Carolina University (Janaye Houghton) for their help with recordings, data collection and analysis. We would also like to thank Peter Assmann, Geoffrey Stewart Morrison and Michael Kiefte for their comments on earlier versions of this chapter.

References

Agwuele, A., Sussman, H.M., Lindblom, B.: The effects of speaking rate on consonant vowel coarticulation. Phonetica **65**, 194–209 (2008). doi:10.1159/000192792

Andruski, J.E., Nearey, T.M.: On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables. J. Acoust. Soc. Am. **91**, 390–410 (1992). doi:10.1121/1.402781

Assmann, P.F., Nearey, T.M., Bharadwaj, S.V.: Developmental study of vowel-inherent spectral change. J. Acoust. Soc. Am. **125**, 2696 (2009)

Bailey, G., Wikle, T., Tillery, J., Sand, L.: The apparent time construct. Lang. Var. Change **3**, 241–264 (1991)
Cassidy, F.G., Hall, J.H.: The Dictionary of American Regional English. Harvard University Press, Cambridge (1985)
Chambers, J.K.: Sociolinguistic Theory, 2nd edn. Blackwell, Oxford (2003)
Cox, F.: Vowel change in Australian English. Phonetica **56**, 1–27 (1999). doi:10.1159/000028438
Fox, R.A., Jacewicz, E.: Cross-dialectal variation in formant dynamics of American English vowels. J. Acoust. Soc. Am. **126**, 2603–2618 (2009). doi:10.1121/1.3212921
Gordon, M.: Small-town Values and Big-city Vowels: A Study of the Northern Cities Shift in Michigan. Duke University Press, Durham (2001)
Hillenbrand, J.M., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**, 3099–3111 (1995). doi:10.1121/1.411872
Jacewicz, E., Fox, R.A., Salmons, J.: Prosodic prominence effects on vowels in chain shifts. Lang. Var. Change **18**, 285–316 (2006)
Jacewicz, E., Fox, R.A., Salmons, J.: Vowel duration in three American English dialects. Am. Speech **82**, 367–385 (2007). doi:10.1215/00031283-2007-024
Jacewicz, E., Fox, R.A., O'Neill, C., Salmons, J.: Articulation rate across dialect, age, and gender. Lang. Var. Change **21**, 233–256 (2009a). doi:10.1017/S0954394509990093
Jacewicz, E., Fox, R.A., Lyle, S.: Variation in stop consonant voicing in two regional varieties of American English. J. Int. Phonetic Assoc **39**, 313–334 (2009b). doi:10.1017/S0025100309990156
Kurath, H.: A Word Geography of the Eastern United States. University of Michigan Press, Ann Arbor (1949)
Kurath, H., McDavid, R.I.: The Pronunciation of English in the Atlantic States. University of Michigan Press, Ann Arbor (1961)
Labov, W.: Principles of Linguistic Change. 1: Internal Factors. Blackwell, Oxford (1994)
Labov, W: Principles of Linguistic Change. 2: Social Factors. Blackwell, Oxford (2001)
Labov, W., Ash, S., Boberg, C.: Atlas of North American English: Phonetics, Phonology, and Sound Change. Mouton de Gruyter, Berlin (2006)
Labov, W., Jaeger, M., Steiner, R.: A Quantitative Study of Sound Change in Progress. U.S. Regional Survey, Philadelphia (1972)
Lee, S., Potamianos, A., Narayanan, S.: Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. **105**, 1455–1468 (1999). doi:10.1121/1.426686
Moon, S.-J., Lindblom, B.: Interaction between duration, context, and speaking style in English stressed vowels. J. Acoust. Soc. Am. **96**, 40–55 (1994). doi:10.1121/1.410492
Nearey, T.M., Assmann, P.F.: Modeling the role of vowel inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986). doi:10.1121/1.394433
Peterson, G., Barney, H.L.: Control methods used in a study of the vowels. J. Acoust. Soc. Am. **24**, 175–184 (1952). doi:10.1121/1.1906875
Sievers, E.: Grundzüge der Phonetik. Breitkopf and Härtel, Leipzig (1st Edn., 1876) (1881)
Sledd, J.: Breaking, umlaut, and the southern drawl. Language **42**, 18–41 (1966)
Stockwell, R.: Perseverance in the English vowel shift. In: Fisiak, J. (ed.) Recent Developments in Historical Phonology, pp. 337–348. Mouton, The Hague (1978)
Thomas, E.R.: An Acoustic Analysis of Vowel Variation in New World English. Publication of the American Dialect Society, p. 85 (2001)
Watson, C., Harrington, J.: Acoustic evidence of dynamic formant trajectories in Australian English vowels. J. Acoust. Soc. Am. **106**, 458–468 (1999). doi:10.1121/1.427069
Wolfram, W., Schilling-Estes, N.: American English. Blackwell, Malden (2006)

Developmental Patterns in Children's Speech: Patterns of Spectral Change in Vowels

Peter F. Assmann, Terrance M. Nearey and Sneha V. Bharadwaj

Abstract The aim of this chapter is to compare the patterns of spectral change in American English vowels spoken by children and adults from the North Texas region. Children's speech differs from adult speech in several important ways. First, children have smaller larynges and supra-laryngeal vocal tracts than adults, with the result that their formants and fundamental frequencies are higher. Second, the temporal and spectral properties of children's speech are inherently more variable, a consequence of developmental changes in motor control. Both of these sources of variability raise interesting questions for the representation of vowel inherent spectral change (VISC) and theories of vowel specification. Acoustic analyses of children's vowels indicate reliable VISC properties as early as age five, the youngest group studied here. Consistent with developmental changes in vocal tract anatomy, the frequencies of vowel formants show an overall systematic decrease with age, and these changes are larger in males than females. The effects of age on formant frequencies vary somewhat from vowel to vowel, but these discrepancies do not appear to interact systematically with VISC. Pattern classification tests indicate that (1) vowels are more accurately recognized when two analysis frames, sampled around 20 and 70 % of the vowel duration, are presented to the classifier, compared to any single frame; (2) adding a third analysis frame does not yield substantially higher recognition scores; and (3) the optimum

P. F. Assmann (✉)
School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, TX, USA
e-mail: assmann@utdallas.edu

T. M. Nearey
Department of Linguistics, University of Alberta, Edmonton, AB, Canada

S. V. Bharadwaj
Department of Communication Sciences and Disorders, Texas Woman's University, Denton, TX, USA

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_9,

locations for sampling the formant trajectory are consistent across different age groups of children.

Abbreviations

ASR	Automatic speech recognition
ANOVA	Analysis of variance
CVC	Consonant-vowel-consonant
dB	Decibel
F0	Fundamental frequency
F1	First formant
F2	Second formant
F3	Third formant
F4	Fourth formant
LPC	Linear predictive coding
LDA	Linear discriminant analysis
VISC	Vowel inherent spectral change

1 Introduction

Compared to other classes of speech sounds, vowels are associated with relatively slow changes in their spectral properties. However, the traditional description of English vowels as consisting of simple steady-state targets has been challenged for some time. It has long been recognized that diphthongs such as /aɪ/, /aʊ/ and /ɔɪ/ in American English must be treated differently, either as a sequence of two targets or in terms of a more complex description based on the detailed aspects of spectral change (Lehiste and Peterson 1961; Holbrook and Fairbanks 1962; Gay 1968). Subsequent research has shown that several vowels of North American English, traditionally classed as monophthongs, show systematic patterns of vowel inherent spectral change (VISC) which are apparently not contextually conditioned, but seem instead to reflect inherent properties of the vowels themselves. Furthermore, these time-varying changes in the formants contribute in important ways to vowel perception (Nearey and Assmann 1986; Andruski and Nearey 1992; Hillenbrand 2013 Chap. 2). Differences in vowel duration also help to distinguish pairs of vowels with similar spectral qualities (Tiffany 1953; Peterson and Lehiste 1960). In this chapter, we present an acoustic analysis of VISC and vowel duration in American English vowels spoken by children in North Texas.

One reason why the study of VISC in children's voices is of interest is that most previous studies of VISC have been restricted to adult voices. Children's speech is inherently more variable than adult speech, raising the question of the impact of this variability on vowel perception. Several factors that may contribute to this increased variability have been investigated in the literature, including anatomical

differences in the size and shape of the vocal structures related to the talker's age and sex (Fitch and Giedd 1999; Vorperian and Kent 2007), developmental changes in motor control (Green et al. 2002), and the coarticulatory effects of adjacent consonants (Sussman et al. 1992). Previous research has indicated that formant frequency trajectories in children's vowels are more variable than those of adults, and there may be developmental differences in the production of VISC (Lee et al. 2004) and in the extent of coarticulation with adjacent consonants (Gibson and Ohde 2007). These studies suggest there might be age-related differences in vowel formant trajectories that include both context-dependent and phoneme-specific aspects, warranting further investigation in a large-scale, controlled sample.

A second reason why the study of VISC in children's voices is of interest is that children's speech presents significant problems for automatic speech recognition (ASR) and speech synthesis. Using acoustic models trained on adult speech, error rates are on average 2–5 times higher for children's speech compared to adults' (Potamianos and Narayanan 2003). Human speech recognition accuracy is also somewhat lower for children's voices compared to those of adults (e.g., Assmann and Katz 2000, 2005), but these differences are small compared to the substantial degradation in ASR performance. The reasons for the human–machine discrepancy are not well understood and have not yet been studied extensively. Children's voices are also generally harder to synthesize than those of adults (Klatt and Klatt 1990). Comparisons of the acoustic properties of children's and adults' speech (including the analysis and modeling of VISC) could lead to more effective strategies for synthesizing realistic children's voices.

In the following sections, we describe a database of vowels recorded in /hVd/ syllables by adults and children ranging from 5 to 18 years. This work represents the first stages of a planned program of research to investigate a wider range of phonetic contexts in children and adults. Vowel formant frequencies are plotted as trajectories in formant space (F1 × F2 frequency; Potter and Peterson 1948). Following a graphical portrayal of formant trajectories, we report the results of a series of statistical pattern classification tests to determine how well different classes of model can distinguish the vowels using information related to VISC. Measurements from the children's vowel database are sampled in different ways to determine the temporal distribution of vowel information in the formant trajectories. Guided by these results and previous studies in the literature, these parameters are used to provide a statistical summary of the patterns of VISC for different vowels and demonstrate how they vary as a function of age.

1.1 Background

Most studies of the acoustic properties of vowels in children have analyzed a single sample from the "steady-state" portion; very little information is available on the role of VISC. In their review of the development of the acoustic vowel space, Vorperian and Kent (2007) point out that children show a gradual reduction

in the formant frequencies with age, accompanied by a decrease in the size of the F1–F2 vowel space. In addition, they show a reduction in formant frequency variability, and some evidence suggests that F1 stabilizes earlier than F2 (Nittrouer 1993). Male–female differences in vowel formant frequencies begin to emerge as early as 4 years of age (Perry et al. 2001) and reach adult-like proportions around age 16, with jumps in formant frequency occurring at ages when anatomical growth spurts are noted.

Lee et al. (1999) reported a large-scale acoustic study of children's vowels, including measures of formant frequencies, fundamental frequency (F0) and duration in children ranging in age from 5 to 18 years. They recorded vowels in CVC syllables and reported formant frequency measurements from the "steady-state" portion of the vowel, excluding the diphthongs /aɪ/, /aʊ/ and /ɔɪ/ as well as /e/ and /o/. Each vowel was recorded in a single CVC context, and different contexts were used for different subsets of the vowels (five vowels used the /bVt/ context; two used /bVd/; one used /bVl/; and two used /pVt/). Lee et al. reported a progressive lowering of vowel formant frequencies with age, approaching adult values at approximately 15 years in males and 14 years in females, and little further decrease in older children. Males and females begin to diverge around age 11, with a steeper and more nearly linear decrease with age in males, consistent with sex differences in vocal tract anatomy. Lee et al. did not include explicit measures of VISC in their analyses. However, they did quantify spectral variability over the time course of the vowel, using the Euclidean distance between the first and second half of the vowel in a mel cepstrum representation. This measure showed a significant reduction in variability with age up to around 15 years. They also reported a decrease in vowel duration with age, as well as a progressive reduction in within-subject duration variability.

2 Vowel Database

We investigated the pattern of VISC as a function of age using a database of vowel recordings obtained from 208 children ranging in age from 5 through 18 years from the Dallas, Texas region over a two-year period between 2006 and 2008. At least 10 children were included at each age level, with roughly equal numbers of boys and girls. Recordings were also obtained from 36 adults (20 males, 17 females) who were undergraduate students at the University of Texas at Dallas, ranging in age from 19 to 45 years. All speakers had normal speech and hearing, spoke English as their native language and were long-term residents of the North Texas region.

2.1 Vowel Subclasses

The vowel set comprised the 12 nominal monophthongs of North Texas English: /i/, heed; /ɪ/, hid; /e/, hayed; /ɛ/, head; /æ/, had; /ʌ/, hud; /ɚ/, herd; /ɑ/, hod; /ɔ/, hawed; /o/, hoed; /ʊ/, hood; /u/, who'd; and the three diphthongs: /aɪ/, hide; /aʊ/, how'd; and /ɔɪ/, hoyed. For descriptive purposes (and to facilitate the graphical portrayal of vowel formant frequencies) the vowels are grouped into four sets in the analyses and illustrations below, based on the following considerations.

1. The phonemic diphthongs /aɪ/, /aʊ/, and /ɔɪ / are regarded as single phonemes based on their distributional properties but exhibit phonetic properties that suggest a sequence of two vowels or a vowel plus glide sequence, and are characterized by a substantial amount of VISC.
2. The second group consists of /i/, /e/, /o/, and /u/. In North American English, /e/ and /o/ show formant frequency movement toward the positions in the vowel space occupied by /i/ and /u/, respectively. These vowels are described as having diphthongal phonetic manifestations [eɪ] and [oʊ] respectively in phonetics texts, and are commonly excluded in acoustic studies of North American English monophthongs (e.g., Peterson and Barney 1952). Some transcription systems (Trager and Smith 1951; Prator and Robinnette 1985) also treat /i/ and /u/ as diphthongal, reflected in their transcriptions as [ij] and [uw], or [i^{i}] and [u^{u}]. Ladefoged and Johnson (Ladefoged and Johnson 2010, p. 91) note that the vowel /i/ may be produced as a diphthong by some speakers in the Northeastern United States (see Labov et al. 2006). We group /i/ with /e/ and /u/ with /o/ because they occupy adjacent regions of the vowel space and are predicted (at least implicitly in some transcription systems) to show similar patterns of VISC.
3. The vowels /ɪ/, /ɛ/, /æ/, and /ʊ/ may exhibit formant movement toward the low central region of the vowel space (Nearey and Assmann 1986; Andruski and Nearey 1992) which might be rendered as [$ɪ^{a}$, $ɛ^{a}$, $æ^{a}$, $ʊ^{a}$], with the [a] superscript indicating movement in the direction of a low central vowel (with high F1 and neutral F2 values), a pattern dubbed by Nearey (2013 Chap. 4) as 'alpha-VISC'.
4. The remaining vowels /ʌ/, /ɚ/, /ɑ/ and /ɔ/ have not been associated with a distinctive pattern of VISC in dialects of North American English. These include the central vowels /ʌ/ and /ɚ/ and the back vowels /ɑ/ and /ɔ/.

2.2 Recording Methods

Recording sessions lasted 45 min to one hour. The children were paid $10 for their participation; the adults received course credit for research participation. Vowels

were recorded in /hVd/ syllables, both in isolation and in a carrier sentence, "Please say the word _____ again." The analyses reported here include only the isolated /hVd/ syllables; other CVC contexts will be investigated in a future study.

The recordings were elicited following a screen prompt that displayed the orthographic representation of the /hVd/ syllable along with an audio example spoken by an adult female from the Dallas area.[1] Five repetitions of each of the 12 vowels were elicited from each of the children, and 10 repetitions from each adult. Recordings were made in a sound-treated room using a Shure SM-94 microphone, Symetrix SX202 dual-microphone pre-amplifier and Tucker-Davis Technologies data acquisition hardware (MA1, RP2.1). The digital waveforms were stored on computer disk at a sample rate of 48 kHz and 16-bit resolution. Following data collection, each recording was screened by phonetically trained listeners and tokens judged to be misarticulated or of low recording quality were omitted from subsequent analysis.

2.3 Acoustic Analysis

A semi-automated procedure was used to mark the onset and offset of the vowel in the isolated /hVd/ words. Vowel onset was defined as the beginning of the first pitch period of the voiced segment of the syllable. Vowel offset was defined as the end of the last pitch period before the silent interval created by the final /d/ or by a substantial drop (>15 dB) in the levels of the higher formants (F2–F5) in cases where voicing filled the closure. Formant frequencies were measured every 2 ms using an automatic formant tracking program (Nearey et al. 2002).[2] Formant tracks were overlaid on the spectrogram of each syllable and tracking errors were corrected using a graphical track editor (Assmann et al. 1994). Fundamental frequency was estimated at 1-ms intervals using the procedure developed by Kawahara et al. (2005). Figure 1 displays the analysis of a recording of the syllable /hæd/ from a 13-year old female speaker.

[1] The audio + visual prompt provides a effective protocol for eliciting vowels from younger children who have not yet learned to read, and limits the occurrence of spelling confusions in older children (Assmann and Katz 2000).

[2] Formant candidates (F1, F2 and F3) were estimated using a 9th-order autocorrelation LPC analysis followed by root extraction. The analysis was performed at a sample rate of 12 kHz, using a 25-ms raised cosine window updated at 2 ms intervals. The F3–F4 cutoff frequency (using selective LPC) was varied to accommodate the formant ranges for speakers of different age and sex classes using a "goodness score" that incorporates several measures to reflect the plausibility and stability of the estimates. Formant tracks were smoothed using a five-point running median filter.

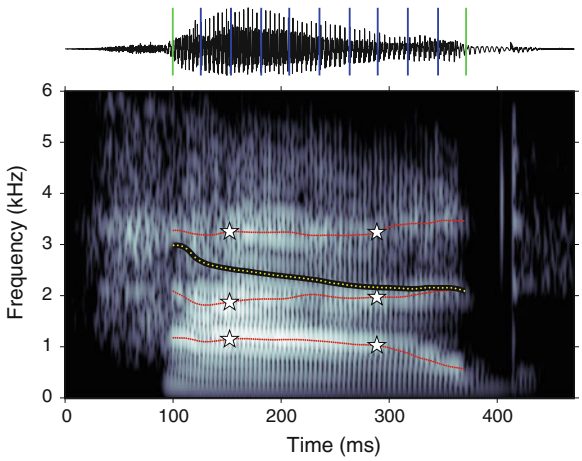


Fig. 1 Waveform and spectrogram with superimposed formant tracks for the syllable /hæd/ recorded by a 13-year old female speaker. *Green vertical lines* on the waveform indicate vowel onset and offset as defined by the gating procedure described above; *blue lines* mark 10 % intervals. *White stars* on the spectrogram indicate the 20 and 70 % points in the vowel. F1, F2 and F3 frequency estimates are indicated by *red lines*. The *dotted yellow line* shows F0 measurements, multiplied by a factor of 10 for comparison purposes

2.4 Formant and Age Groupings

To analyze the pattern of formant frequency movement in each vowel, we divided the formant trajectory across the vowel (with onsets and offsets as defined above) into 10 equal segments, producing 11 analysis frames. This provides a time-normalized representation that simplifies the comparison of the same vowel produced at different speaking rates and with different durations. To facilitate the visual interpretation of the interaction of formant movement with talker age, we grouped the 5–18 year range into four sub-ranges: (1) 5–8 year olds (26 males, 36 females); (2) 9–12 year olds (31 males, 34 females); (3) 13–16 year olds (26 males, 30 females); and (4) 17–18 year olds (11 males, 13 females).[3]

3 Graphical Portrayal of Formant Trajectories

Figures 2–8 present the formant frequency trajectories (F1 vs. F2) for the four vowel classes described above. Formant frequencies were taken from the vocalic portion of /hVd/ syllables, sampled at 0, 10, 20 … 100 % of the duration from

[3] These age groupings and cutoffs are adopted as a descriptive convenience; they do not imply discontinuities at the group boundaries.

vowel onset to offset as shown in Fig. 1. Each sample point is the mean F1 and F2 across talkers, based on the medians of all of each talker's productions of the vowel. To facilitate visual comparison, the axis limits are adjusted for each vowel to match the frequency range occupied by that vowel across the age and sex classes. Four points along the trajectory are highlighted: small circles indicate the vowel onset at 0 %. The squares indicate 20 and 70 % points, which provide effective sample points for statistical pattern classifiers for vowel classification (e.g., Hillenbrand et al. 2001); and the arrows indicate the vowel offset at 100 %.

3.1 Formant Trajectories for /aɪ/, /aʊ/, /ɔɪ/

Figure 2 displays the formant frequency trajectories (F1 vs. F2) for /aɪ/(upper panel), /aʊ/ (middle panel), and /ɔɪ/ (lower panel) separately for males (left panels) and females (right panels). As expected, the diphthongs /aɪ/, /aʊ/ and /ɔɪ/ showed extensive and distinct patterns of VISC, with fairly similar trajectories across the different age/sex groups. In all three diphthongs there is little change between the 0 and 20 % point in the vowel. Between 20 and 70 % the diphthong /aɪ/ shows the expected pattern of falling F1 and rising F2. Considerable formant movement is present, as indicated by the distance between the 20 and 70 % points. Between 70 and 100 % some further upward movement is evident, probably reflecting the influence of the final /d/. The 100 % point (indicated by the arrowhead) represents a sample point just before the /d/ closure, which has an expected F2 locus around 1,800 Hz for adult males and around 2,000 Hz for adult females (Sussman et al. 1997).

The North Texas region is on the western edge of the Southern dialect region in the sociolinguistic analysis provided by Labov et al. (2006). The diphthong /aɪ/ is an interesting case in this region because it does not consistently follow the shift to monophthongization found in other parts of the South [e.g., in the North Carolina dialect documented by Jacewicz and Fox (2013) Chap. 8] especially among younger/urban speakers (Thomas 1997, 2003).

In the case of /aʊ/, both F1 and F2 decline between the 20 and 70 % points, but around the 70 % point there is an abrupt increase in F2, reflecting the contextual influence of the final /d/ consonant. We refer to this abrupt change in the trajectory of the second formant as a "switchback" pattern, and it illustrates the combined forces of the vowel and consonant on the F2 trajectory. The diphthong /ɔɪ/ in the bottom panel shows formant movement patterns between the 20 and 70 % points that deviate rather markedly from a linear path in the log F1 × log F2 space. An initial increase in F1 is followed by a decrease, which extends beyond the 70 % point up to the 100 % terminus. F2 shows a steady increase between 20 and 70 % points and eventually levels off. For this diphthong, the vowel-related movement is in the same direction as the influence of the final consonant. Overall,

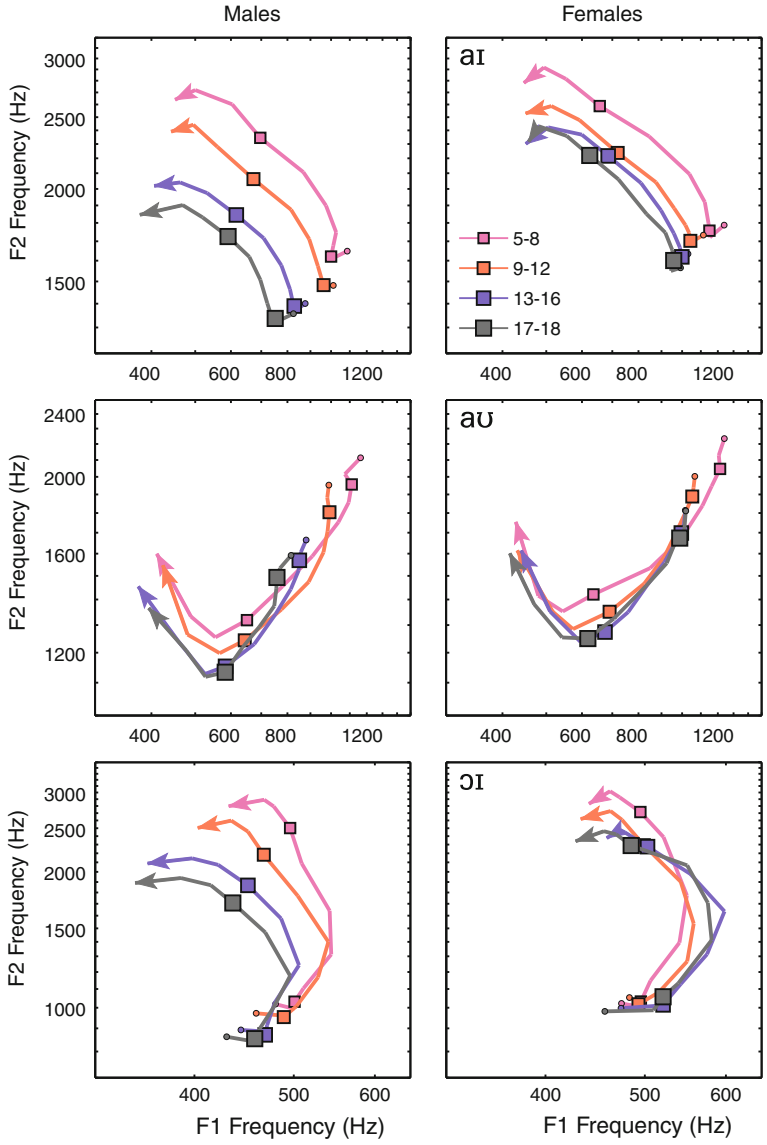


Fig. 2 Formant frequency trajectories (in 10 % steps from vowel onset to offset) for the 3 diphthongs /aɪ/ (*upper panels*), /aʊ/ (*middle panels*), and /ɔɪ/ (*lower panels*) as a function of age. *Small circles* indicate the onset of the F1 and F2 trajectory, *arrows* indicate the terminal value, and *squares* indicate the 20 and 70 % points in the vowel. Speaker age ranges (in years) are indicated by *symbol color*

Fig. 2 illustrates that the 20 and 70 % sample points can be used to provide a useful representation of VISC, because vowel-related formant movement is largely completed before the switchback pattern is initiated. This is most evident for /aʊ/, where the F2 trajectory takes an abrupt turn. In the case of /aɪ/ and /ɔɪ/ the switchback is less pronounced, presumably because the termination of the diphthong and the formant movement toward the final /d/ are in the same direction; both involve an increase in F2 frequency and a decrease in F1.

For male speakers, there was a systematic downward shift in F1 × F2 trajectories as a function of age, reflecting the anatomical changes in vocal tract size (Vorperian and Kent 2007). For female speakers, a downward shift of smaller magnitude was observed, but with overlap across the age groups, especially for the two older groups, 13–16 and 17–18 years, and a somewhat inconsistent pattern between onset and offset frequencies for some vowels. However, where age differences between the groups are present, these differences are generally in the expected direction of lowered formant patterns for older speakers.

The age-related lowering of formant frequencies is most clearly seen in /aɪ/, while there is greater overlap across the age groups for /aʊ/ and /ɔɪ/. All four age groups have similar F1 and F2 values near the onset of /ɔɪ/ for the female speakers but diverge somewhat at the offset, a pattern that appears to be associated with higher F2 frequencies. Some aspects of the data, such as the greater spreading of the formant trajectories for males across age, may be a consequence of the larger changes in vocal tract length associated with pubescence for males. Age-related differences between males and females may also reflect non-uniform growth patterns in the vocal tract which lead to nonlinear scaling of formant frequencies (Fant 1966), or even sociophonetic variation in articulation patterns (Johnson 2006). Each of these factors requires further investigation.

3.2 Formant Trajectories for /e/, /o/, /i/, /u/

Figure 3 displays the formant frequency trajectories for /e/ and /o/. The vowel /e/ is characterized by a decrease in F1 coupled with an increase in F2 between the 20 and 70 % points, while for /o/ both F1 and F2 are decreasing. In both vowels there is an abrupt change in F1 and F2 between the 70 and 100 % points, producing a *bowed* trajectory. These cases provide further evidence of the switchback pattern in formant frequencies resulting from the combined influence of VISC and the contextual influence of the final consonant. The formant trajectories for the male speakers show fairly consistent age-related changes, while the female speakers again show greater overlap across age groups.

The formant trajectories for /i/ and /u/ are shown in Fig. 4. The vowel /i/ (upper panels) shows an abrupt increase in F1 at the onset and a decrease at the offset, but relatively little movement in either F1 or F2 between the 20 and 70 % points.

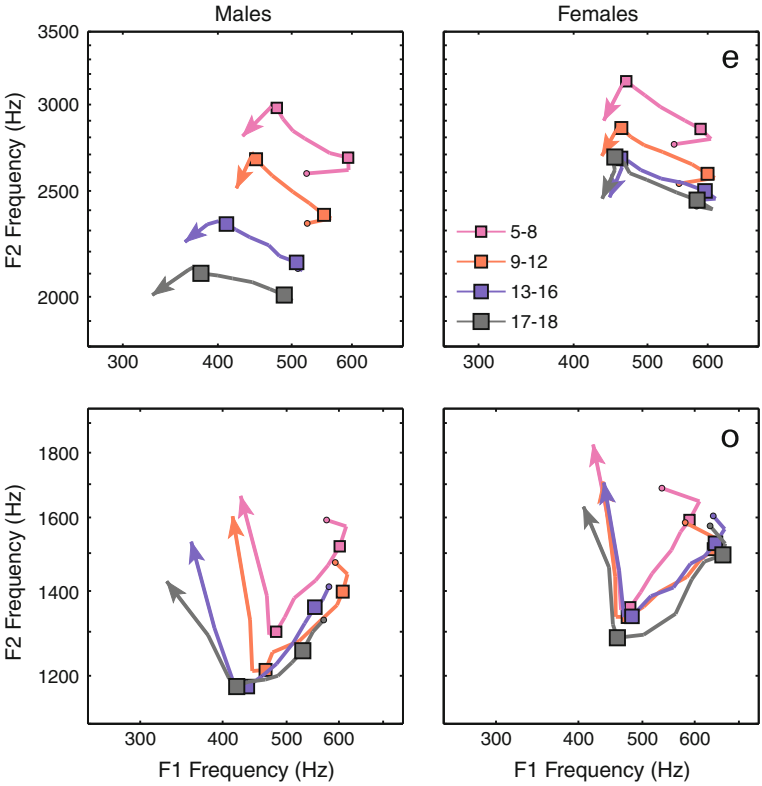


Fig. 3 Formant trajectories for /e/ (*upper panels*) and /o/ (*lower panels*) as a function of age

There is a clear progression toward lower F1 and F2 frequencies as a function of age in males, but this pattern is less evident for female speakers. Compared to /i/, the vowel /u/ (lower panels) shows relatively greater formant movement, with a small decrease in F1 and a fairly steep decrease in F2 between the 20 and 70 % points. There are also fairly abrupt changes in F2 at onset and offset, with a decrease in F2 from 0 to 20 % and an increase from 70 to 100 % in most cases.

3.3 Formant Trajectories for /ɪ/, /ɛ/, /æ/, /ʊ/

Figure 5 displays the formant trajectories for /ɪ/ and /ɛ/. Similar to /i/, the vowel /ɪ/ (upper panels) increases in F1 at the onset and decreases at the offset, with a relatively small distance between the 20 and 70 % points indicating a small degree of formant movement. Males show a fairly linear decrease in F1and F2 with age, while females show a more complex pattern; although where age differences occur

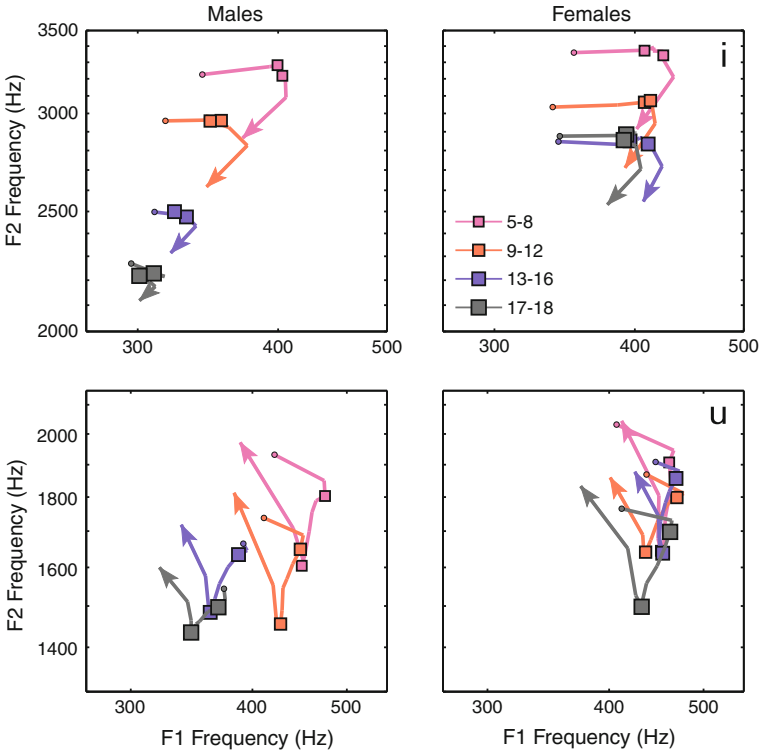


Fig. 4 Formant trajectories for /i/ (*upper panels*) and /u/ (*lower panels*) as a function of age

they are in the predicted direction. For /ɛ/ the formant movement is restricted to a decrease in F1 throughout the vowel, with the largest change occurring between 70 and 100 %, accompanied by a small decrease in F2.

The vowel /æ/, shown in the upper panels of Fig. 6, has a pattern similar to /ɛ/, with little movement in the vowel up to the 70 % point, followed by an abrupt decrease in F1 and a slight increase in F2. For the male speakers, there is a fairly linear decrease in formant frequencies as a function of age in both F1 and F2 and a clear separation of F2, but converging F1 frequencies around the 100 % point. Female speakers also show a clear F2 separation as a function of age, although the two oldest groups of speakers show a reduced separation. F1 frequencies are not well separated by age and converge for several age groups.

The lower panels of Fig. 6 show that the vowel /ʊ/ has a rising F2 over the course of the vowel, with greater formant movement between the 20 and 70 % points than in the other three lax vowels. There are small changes in F1 over the central portion of the vowel, but a substantial decrease between the 70 and 100 % points.

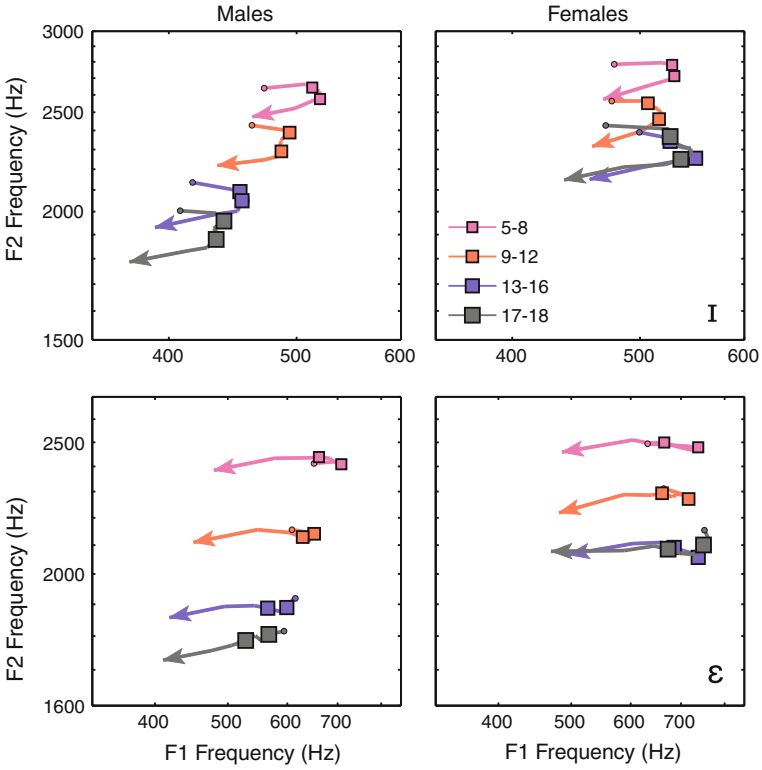


Fig. 5 Formant trajectories for /ɪ/ (*upper panels*) and /ɛ/ (*lower panels*) as a function of age

3.4 Formant Trajectories for /ʌ/, /ɚ/, /ɑ/, /ɔ/

The central vowel /ʌ/ (Fig. 7 upper panels) is characterized by a decrease in F1 and increase in F2 from 20 to 70 % points followed by a further drop in F1 and increase in F2 from 70 to 100 %. A similar pattern of formant movement is seen for the vowel /ɚ/ (Fig. 7 lower panels), which also shows a decrease in F3 over the central portion of the vowel (not shown here). Similar to other vowels, there is a clear separation by age for the male speakers but greater overlap across age groups for the female speakers.

The vowels /ɑ/ and /ɔ/ (Fig. 8) show very similar patterns of formant movement. This similarity was anticipated, as the distinction between these two vowel categories is not consistently maintained in this dialect (Labov et al. 2006; Thomas 2003; Assmann and Katz 2000; Katz and Assmann 2001). Both vowels show an increase in F2 but little change in F1 from the 20 to 70 % points. From 70 to 100 % there is a substantial further decrease in F1 and an increase in F2. Differences across age groups are similar to the pattern for /ʌ/ and /ɚ/.

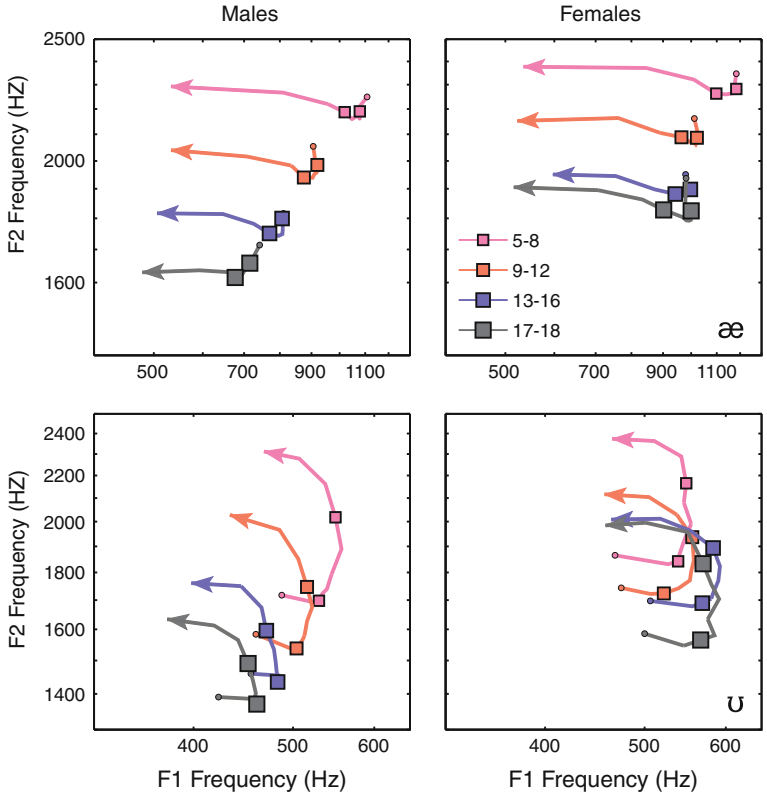


Fig. 6 Formant trajectories for /æ/ (*upper panels*) and /ʊ/ (*lower panels*) as a function of age

3.5 Summary of Formant Trajectories

Consistent with earlier findings, the formant trajectories in Figs. 2–8 suggest the presence of VISC in several vowels traditionally classified as monophthongs. The trajectories of several vowels, most notably /e/ and /o/, show evidence of switchback patterns, where formant movement ostensibly associated with the vowel shows an abrupt change in direction near the final portion of the syllable in anticipation of the final consonant. In other cases, such as /ʌ/, there is nearly continuous formant movement from the 20 to 100 % points with little change in direction. In such cases, it is difficult to separate VISC from consonant context effects. In cases where VISC and consonant context exert opposing influences on the formant trajectories, such as F2 in /o/ and /aʊ/, the formant movement associated with VISC appears to be largely completed by the 70 % point in the vowel, and the switchback occurs between the 70 and 100 % points. However, the degree to which formant trajectories can be partitioned into vowel- and consonant-related

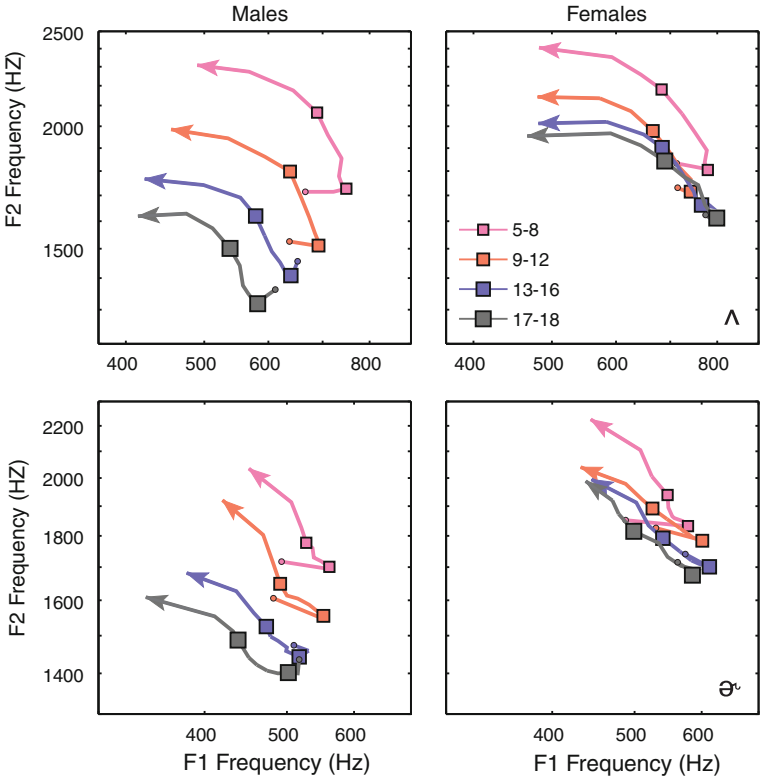


Fig. 7 Formant trajectories for /ʌ/ (*upper panels*) and /ɚ/ (*lower panels*) as a function of age

components has not been formally established, and requires a set of comparisons across a range of vowels and consonant contexts (Nearey 2013, Chap. 4).

With minor exceptions, the distinctive shapes of the formant trajectories for each vowel are well preserved across age and sex classes. For the male speakers, every vowel shows a progressive lowering of the formant frequencies with age, and as a result the formant trajectories undergo a downward and leftward shift in log F1 × log F2 space as a function of increasing age. A similar trend appears in the data for female speakers, but in some vowels the trajectories reach an earlier asymptote, producing an overlap of trajectories for the 13–16 and 17–18 year-olds. The reduced range of variation for female speakers compared to males is qualitatively consistent with anatomical changes in vocal tract size (Vorperian and Kent 2007).

For some vowels, the formant trajectories spoken by females show reduced overlap across age groups. As noted above, this is partly a consequence of sex-specific anatomical changes in vocal tract size, which may include non-uniform scaling associated with male–female differences in the relative growth of the

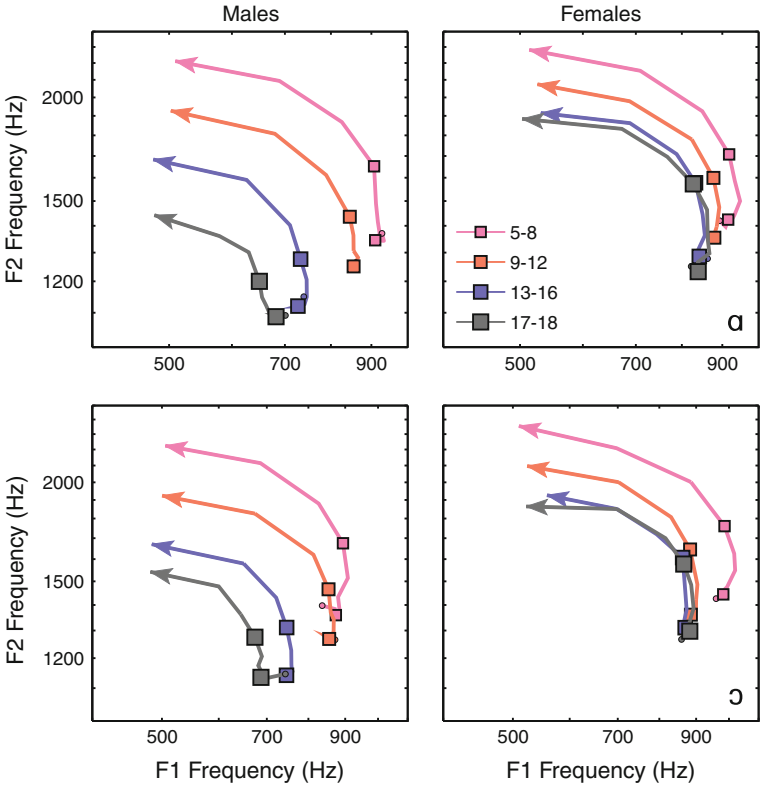


Fig. 8 Formant trajectories for /ɑ/ (*upper panels*) and /ɔ/ (*lower panels*) as a function of age

anterior and posterior cavities of the vocal tract (Fant 1966); it may also reflect social and dialect factors (Johnson 2005). In general, age-related differences are most evident in vowels which undergo the largest frequency excursions (note the different axis limits for individual vowels). Some caution is needed in the interpretation of age-related changes, however, because formant frequency measurements are more susceptible to error when F0 is high. Some younger speakers have breathy voices and show glottalization, particularly near the endpoint of the syllable where the formants may change rapidly, reducing the accuracy of formant measurement.

An additional point is that the formant trajectories in Figs. 2–8 represent averages across a number of speakers and repetitions of each vowel. The question of statistical generalization (whether there is reliable formant movement in the vowels across the set of speakers) is addressed in subsequent sections. Before addressing this question, we return to the question of where and when to sample the formant trajectory so as to provide maximum information about the phonemic identity of the vowel. We examine whether there is more information if the

formant trajectories are sampled at two or three locations rather than a single location using a statistical pattern classifier, and what the optimum sampling location(s) are from the perspective of vowel categorization. This information will help to characterize formant trajectories for the purpose of predicting both the identity of the vowel and adjacent consonants, and will provide a basis for refining models of the perceptual specification of vowel inherent spectral change.[4]

4 Statistical Pattern Classification

Nearey and Assmann (1986) used a two-sample characterization of VISC (with the first sample representing the onset or "nucleus" early in the vowel, and the second sample representing the offset or "tail"). They estimated the formant frequencies (F1, F2, and F3) at the 24 and 64 % points in the vowel, along with mean F0 and duration. The measurements were used to train a pattern classifier (linear discriminant analysis). The classifier successfully predicted key aspects of the perceptual responses of listeners, including the pattern of confusion errors in several conditions with natural and modified vowels. Subsequent studies of vowels with "flattened formants" have provided further validation of this approach (Hillenbrand and Nearey 1999; Assmann and Katz 2000, 2005). Both listeners and pattern classifiers exhibited a drop in overall accuracy in formant-synthesized vowels when the natural time variation was removed by flattening the formant contours over the time course of the vowel.

In the present study, we examine whether a two-sample representation is adequate for the characterization of VISC in children's vowels [for additional evidence and discussion see Morrison (2013a) Chap. 3]. The analyses reported here consider the overall accuracy of vowel classification as a function of age and sex. Evaluations are not based on the correspondence between automatic classification results and identification accuracy by listeners, nor do we take into account the pattern of confusion errors.

A series of statistical pattern classification tests examined VISC properties as a function of age. The approach used was similar to previous studies (e.g., Nearey and Assmann 1986; Hillenbrand and Nearey 1999) in which the statistical variation in acoustic measurements was analyzed using pattern recognition models to

[4] Although there are many other ways a vowel formant trajectory might be summarized e.g., cosine basis functions (Zahorian and Jagharghi 1993; Watson and Harrington 1999) or smoothing splines (Enzinger 2010) (see Morrison 2013b, Chap. 3), we limit our exploration here to multiple samples from the vowel trajectory. In the one direct comparison of vowel classification of which we are aware, Hillenbrand et al. (2001) found nearly identical performance between cosine basis functions and two-slice representations similar to those explored here. The present analysis has the advantage of suggesting specific temporal regions that may be more stable across speakers in certain environments than others. Such an analysis is not possible with a whole-trajectory approach. Whether such temporal 'hot spots' have wider implications remains to be seen.

determine the effectiveness of putative VISC cues. This approach has provided useful insights into the relationship between acoustic properties and the perceptual responses of listeners in vowel identification experiments. In the present study, we report only the statistical classification results; the relationship of these predictions to listeners' responses remains a topic for future investigation.

4.1 Measurement Variables

Following the general strategy outlined by Hillenbrand and Nearey (1999), we incorporated the following measures: mean F0, estimated at 1-ms intervals and averaged over the duration of the vowel; formant frequencies F1, F2 and F3, estimated at 2-ms intervals from vowel onset (0 %) to offset (100 %), as illustrated in Fig. 1, and sampled as described below; and vowel duration in milliseconds. All frequency variables (F0 and formant frequencies) were log-transformed. Measurements of F0, F1, F2, F3 and duration were taken from 10287 /hVd/ tokens of 11 monophthongal vowels (the phonemic diphthongs were excluded, along with /ɔ/ which is not consistently distinguished from /ɑ/ by many speakers of the north Texas dialect) from 208 children ranging in age from 5 through 18 years. This included up to 5 repetitions of each vowel from each child. A series of linear discriminant analyses (LDAs) was carried out, using leave-one-token-out cross validation to classify the test stimuli. These analyses differed in the temporal location(s) at which the formant frequencies were measured and the number (1, 2, or 3) of sample points. All of the LDAs included measurements of vowel duration and mean F0, which are summarized in the next section.

4.2 Duration and Fundamental Frequency

Vowel duration measurements are shown in Fig. 9. Consistent with earlier studies (e.g. Lee et al. 1999) there was a progressive decline in vowel durations with age, $F(3, 66) = 18.10$; $p < 0.01$. There were significant differences in duration among the vowels, $F(11, 726) = 215.04$; $p < 0.01$ but no differences between males and females, and no significant 2-way or 3-way interactions of vowel, age or sex.

Figure 10 shows F0 measurements, taken at 1-ms intervals and averaged across the vocalic portion of the syllable. The pattern is similar to that reported by Lee et al. (1999) with a progressive decline in F0 as a function of age, $F(3, 66) = 72.33$; $p < 0.01$. There was also a significant main effect of speaker sex, $F(1, 66) = 148.27$; $p < 0.01$ and a significant interaction of age by sex, $F(3, 66) = 46.10$; $p < 0.01$, reflecting the substantially larger and more abrupt drop in F0 for males, compared to the more uniform and gradual decline for

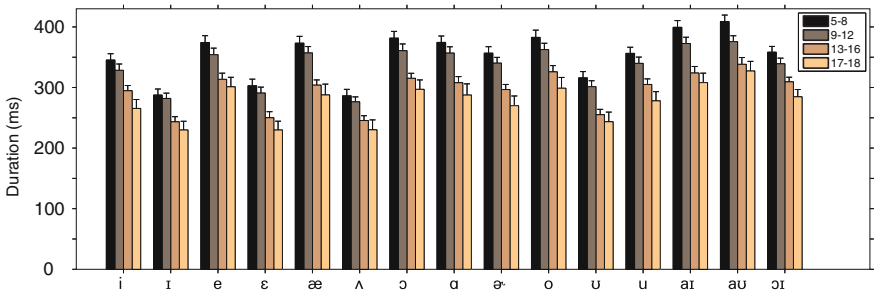


Fig. 9 Vowel duration (in ms) for the 15 vowels as a function of age. *Bars* indicate the means across the talkers in each age group, summarizing each talker's data by the median duration across all repetitions of the vowel by that talker. *Error bars* indicate the standard error across talkers. [Data pooled across males and females. Age ranges 5–8 years, 62 speakers; 9–12 years, 66 speakers; 13–16 years, 56 speakers; 17–18 years, 24 speakers]

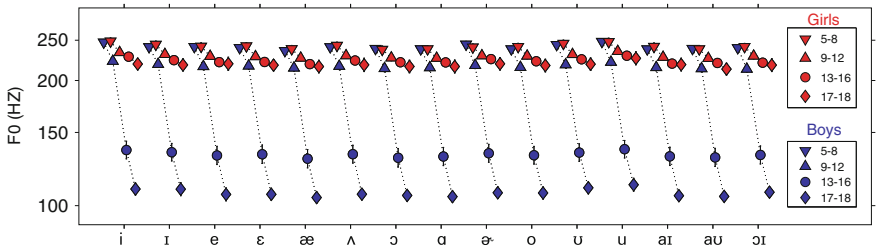


Fig. 10 Mean fundamental frequency (in Hz) for the 15 vowels as a function of age and sex. *Error bars* indicate the standard error across talkers

females. No other interactions were significant. The error bars indicate an increase in variability for male speakers in the 13–16 year range that can be attributed to the variable onset of puberty in this group (Hollien et al. 1994). There was a significant main effect of vowel, $F(11, 726) = 24.07$; $p < 0.01$ reflecting intrinsic F0 differences among the vowels (see Katz and Assmann 2001 for related findings in this dialect). These findings suggest that information provided by F0 and duration is effectively maintained across age and sex classes in children.

4.3 Discriminant Analysis

Linear discriminant analyses were conducted to determine whether an improvement in vowel classification accuracy can be obtained when the formant frequencies are sampled at two locations rather than one, and whether an additional

improvement in overall accuracy is provided by a three-sample representation. We also examined the effects of sampling at different temporal locations in an attempt to determine the statistically optimal sample points for predicting the identity of the vowel and we investigated whether these optimum sample points vary as a function of age and sex.

4.4 Selection of Sample Points

In the first analysis, formant measurements were taken from a single fixed time point in the vowel. Since the vowels were recorded in /hVd/ context, it is reasonable to assume that the earliest and latest time points in the vowel would show the greatest influence of the flanking consonants (Hillenbrand and Nearey 1999; Assmann and Katz 2000, 2005) and that the middle of the vowel would be least susceptible to these influences. As in Assmann and Katz (2000), the formant frequencies F1, F2, F3 were estimated as the median of five successive measurements spaced 2 ms apart, with the middle sample centered on the nominal sample point. For the 0 and 100 % samples, the median was calculated over the initial and final five measurements to accommodate the endpoints.

Figure 11 shows correct-classification rates as a function of frame location within the vowel. The results indicate that the 50 % point in the vowel provides a reasonable sample point if a single frame is used for classification, with nearly identical results for the 20 through 60 % frames. Classification scores are progressively worse for samples taken later in the vowel, dropping by more than 30 % for the final sample point. In comparison, scores are only 5–10 % lower for samples taken at the earliest point in the vowel, indicating that more reliable information about vowel identity is provided in the first half of the vowel. Differences between adjacent frames are small, but are highly consistent across age and sex groups. Overall, classification accuracy was higher for the two older groups of speakers compared to two younger groups (by 7.8 %, on average) with a small difference between males and females (2.0 % higher for males, on average).

Figure 12 shows single-frame classification results across age and sex classes, with correct classification rates for the "best" two frames and "best" three frames. Classification performance improves overall as a function of age, with slightly higher accuracy for males than females, in some cases, for younger talkers. A similar pattern of age-related differences in overall accuracy has been reported in vowel identification by listeners (e.g. Assmann and Katz 2000), but whether predicted classification accuracy closely mirrors identification accuracy by human listeners remains to be investigated. For all age and sex classes, the use of two frames provides consistently more accurate classification than a single frame, with a mean increase of 10.8 %. The results for three frames were mixed: adding a third

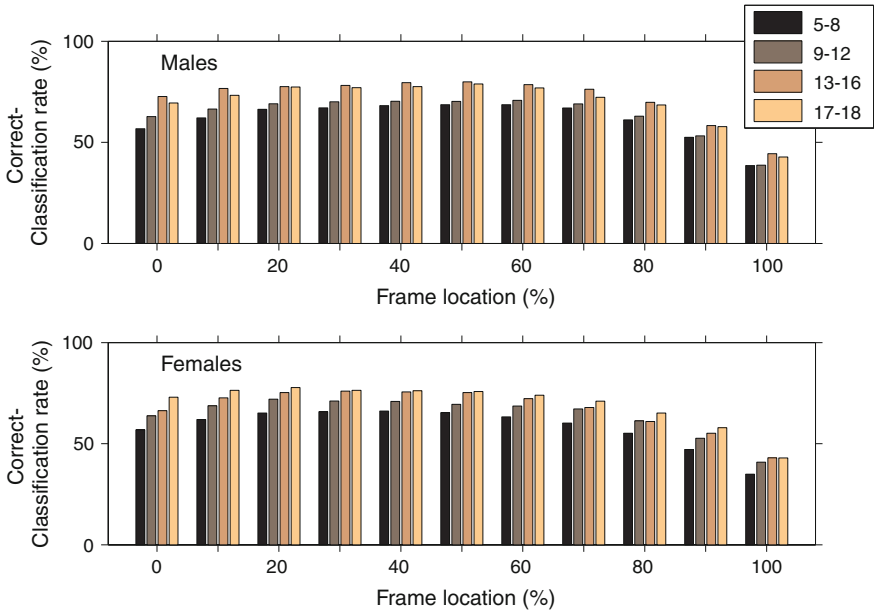


Fig. 11 Single frame classification results as a function of frame location (expressed as percentage of vowel duration). *Bars* indicate mean correct classification rates across vowels and talkers

frame leads to improved classification in some cases, but lowered accuracy in others with less than 1 % change in mean classification overall.

Figure 13 shows contour plots of classification accuracy for two-frame analyses as a function of the frame location in percent, with each sex and age group shown in a separate subplot. All possible pairwise combinations of two frame positions are considered, including same-frame pairings along the diagonal, which are equivalent to a single frame analysis.

A comparison of the eight subplots shows that in each case, the highest classification results are obtained when the initial sample is chosen relatively early in the vowel. For 7 of the 8 age/sex categories, maximum performance is found when the initial sample corresponds to the 20 % frame; the only exception is the 13–16 year old female group for whom the 10 % sample is marginally better. The optimal choice for the second sample occurs relatively early in the second half of the vowel. In 5 of the 8 cases this corresponds to the 70 % frame; the 80 % frame was slightly better for 5–8 year old girls, and the 60 % frame was better for boys in the two oldest age groups. We conclude from these findings that the optimum sampling points lie close to the 20 and 70 % points suggested in earlier research (Hillenbrand et al. 2001).

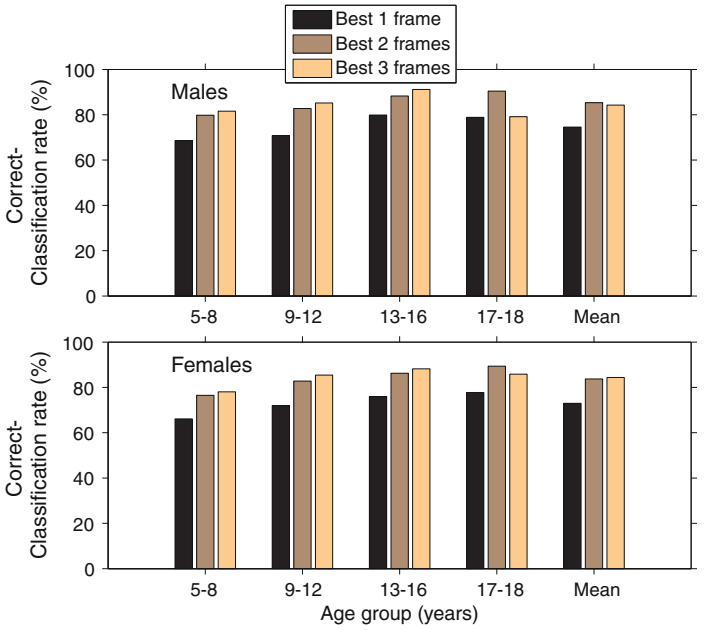


Fig. 12 Vowel classification, as a function of age and sex, for "best" single frame, along with "best" two frame and three frame analyses

5 Statistical Tests for VISC in Adults and Children

Given that the 20 and 70 % points provided the highest LDA classification scores, we analyzed the formant frequency measurements at these two sample points. The aim was to determine which vowels in this dialect show consistent and reliable VISC, and to examine the interaction of VISC with age and sex in children's vowels. Since the presence of VISC in diphthongs is well established, the three diphthongs /aɪ/, /aʊ/, and /ɔɪ/ were excluded from these analyses. We conducted separate analyses of variance (ANOVAs) for adults and children and for males and females, and we analyzed each formant separately since VISC leads to different patterns of F1 and F2 movement. For the adults, we conducted fully repeated measures ANOVAs with two within-subjects factors: Vowel (with 12 levels: /i/, /ɪ/, /e/, /ɛ/, /æ/, /ʌ/, /ɚ/, /ɑ/, /ɔ/, /o/, /ʊ/, /u/); and Frame (onset = 20 % vs. offset = 70 %). For children, we carried out partially repeated measures ANOVAs with the factors Vowel, Frame and Age (a between-subjects factor with four levels: 5–8, 9–12, 13–16, and 17–18 years). Log-transformed F1 or F2 frequency served as the dependent variable. Formant frequencies were summarized by taking the median across all repetitions of the same vowel by the same speaker (up to 10 repetitions of each vowel for adults and up to 5 for children, prior to screening which eliminated a subset of the repetitions).

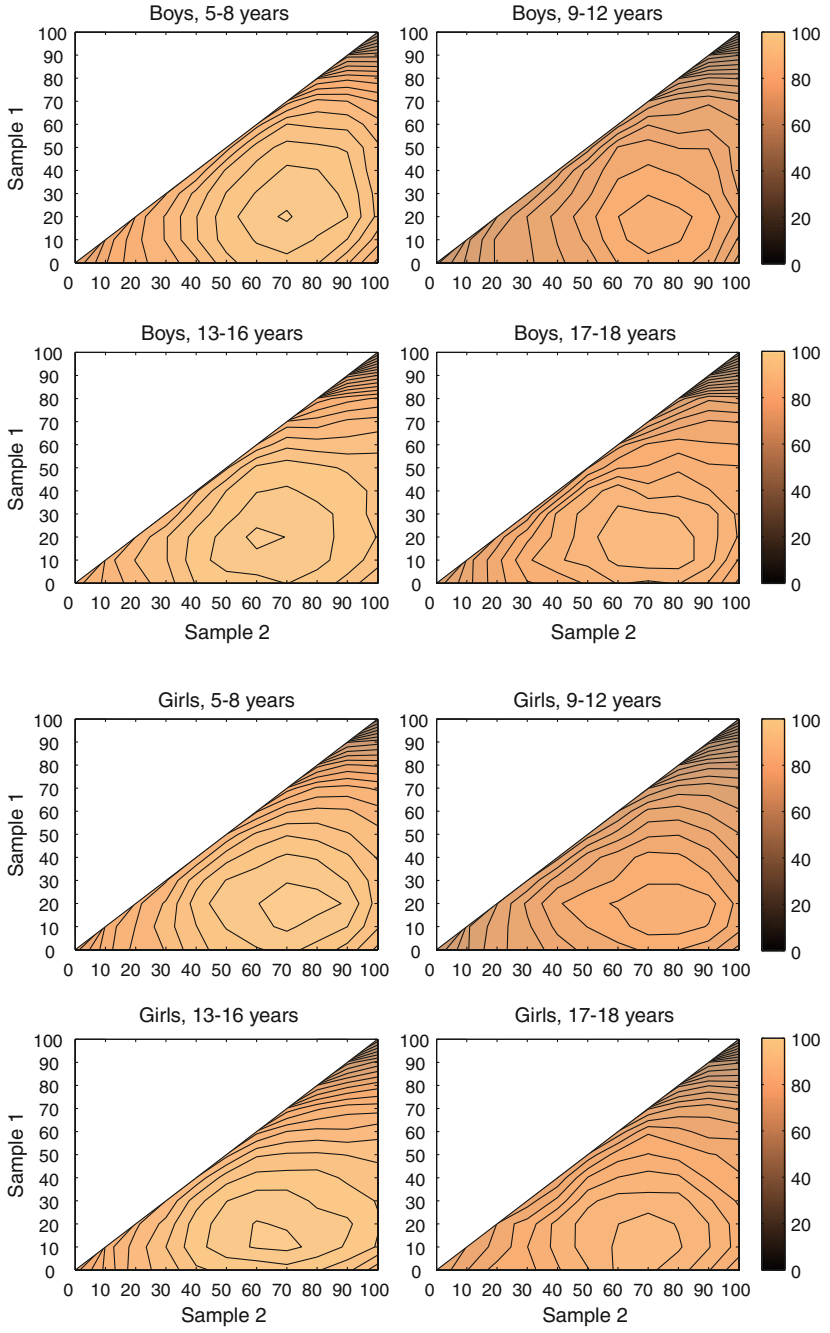


Fig. 13 Contour plots for two-sample classification as a function of time frame position (expressed as a percentage of the vowel duration). Results for each sex and age group are shown in separate panels

5.1 Adults

The analysis of F1 frequency for adult males (20 talkers) resulted in a significant effect of Vowel $F(11, 19) = 334.35$; $p < 0.01$; and Frame, $F(1, 19) = 86.23$; $p < 0.01$, as well as a significant Vowel by Frame interaction, $F(11, 209) = 52.92$; $p < 0.01$. The analysis of F2 frequency resulted in a significant effect of Vowel $F(11, 19) = 224.38$; $p < 0.01$, but no significant effect of Frame, $F(1, 19) = 0.80$; $p = 0.38$, and a significant Vowel by Frame interaction, $F(11, 209) = 53.50$; $p < 0.01$.

A similar pattern was found for F1 frequency in adult females (17 talkers), with significant effects of Vowel $F(11, 19) = 326.19$; $p < 0.01$; and Frame, $F(1, 19) = 155.10$; $p < 0.01$, and a significant Vowel by Frame interaction, $F(11, 209) = 52.26$; $p < 0.01$. F2 frequency produced a significant effect of Vowel $F(11, 19) = 234.28$; $p < 0.01$, Frame, $F(1, 19) = 13.08$; $p < 0.01$, and a significant Vowel by Frame interaction, $F(11, 209) = 33.80$; $p < 0.01$.

To study the pattern of VISC, the Vowel by Frame interaction was analyzed using a set of *post hoc* *t*-tests. Tables 1 and 2 summarize the pattern of F1 and F2 movement in different vowels from the 20 to the 70 % points. For adult males, 8 of the 12 vowels showed significant movement in F1 frequency and 11 of 12 vowels showed significant F2 movement. Adult females showed significant F1 movement in 7 of the 12 vowels and 9 of 12 vowels showed significant F2 movement.

Adopting the analysis of vowel types used in Figs. 3–8, we see that the tense vowels /e/, /o/, /i/, /u/ exhibit formant movement largely consistent with the pattern of VISC implied by the Prator and Robinette (1985) transcription system (Sect. 2.1). F1 frequency showed a downward trend, significant in all four tense vowels

Table 1 Initial (20 %) F1 and F2 measurements along with means and standard deviations in ΔF1 and ΔF2 (70–20 % points)

Vowel	F1	Mean ΔF1	STD ΔF1	*t*	F2	Mean ΔF2	STD ΔF2	*t*
/i/	322	−14.0	20.7	−3.03[a]	2251	20.8	49.5	1.88
/ɪ/	416	21.3	35.6	2.67[a]	1937	−110.9	76.3	−6.50[b]
/e/	502	−100.2	28.8	−15.55[b]	1913	177.0	81.6	9.70[b]
/ɛ/	554	−9.9	33.8	−1.31	1758	−35.4	45.8	−3.46[b]
/æ/	710	−17.1	37.6	−2.03	1695	−58.0	59.3	−4.38[b]
/ʌ/	600	−27.0	21.5	−5.61[b]	1308	102.8	52.7	8.72[b]
/ɔ/	664	−3.8	31.5	−0.54	1050	67.5	59.9	5.03[b]
/ɑ/	698	−10.9	27.4	−1.78	1118	93.6	53.9	7.77[b]
/ɚ/	499	−52.7	25.9	−9.11[b]	1330	49.0	45.0	4.86[b]
/o/	531	−105.7	30.9	−15.30[b]	1171	−118.5	48.0	−11.04[b]
/ʊ/	446	13.2	14.9	3.96[b]	1249	119.1	68.5	7.78[b]
/u/	371	−24.7	17.3	−6.38[b]	1373	−80.4	79.0	−4.55[b]

Adult males ($N = 20$)
[a] $p < 0.01$; [b] $p < 0.05$

Table 2 Initial (20 %) F1 and F2 measurements, along with means and standard deviations in ΔF1 and ΔF2 (70–20 % points)

Vowel	F1	Mean ΔF1	STD ΔF1	*t*	F2	Mean ΔF2	STD ΔF2	*t*
/i/	418	−5.4	23.4	−0.95	2850	13.8	47.5	1.20
/ɪ/	503	9.9	42.4	0.96	2454	−126.9	88.5	−5.91[b]
/e/	629	−158.7	43.5	−15.03[b]	2469	202.3	95.9	8.70[b]
/ɛ/	748	−62.8	31.9	−8.12[b]	2198	−30.4	76.3	−1.64
/æ/	951	−34.7	38.5	−3.71[b]	1981	−33.9	81.9	−1.71
/ʌ/	788	−87.3	31.5	−11.44[b]	1660	184.4	67.0	11.34[b]
/ɔ/	844	−17.7	44.7	−1.63	1240	142.3	141.4	4.15[b]
/ɑ/	878	−18.2	35.9	−2.09	1300	208.7	135.5	6.35[b]
/ɚ/	621	−96.6	24.2	−16.46[b]	1672	71.7	52.7	5.61[b]
/o/	659	−164.7	39.6	−17.14[b]	1473	−162.3	74.6	−8.97[b]
/ʊ/	536	2.7	30.6	0.36	1561	209.6	120.5	7.17[b]
/u/	458	−21.4	31.8	−2.77[a]	1664	−121.7	70.7	−7.10[b]

Adult females ($N = 17$)
[a] $p < 0.01$; [b] $p < 0.05$

for both males and females except for /i/ in females, which was not significant. F2 frequency increased for /i/ and /e/ and decreased for /o/ and /u/, but the difference was not significant for /i/ for either men or women.

Nearey (2013 Chap. 4) reviews evidence from several dialects that some or all of the lax vowels /ɪ/, /ɛ/, /æ/, /ʊ/ show alpha-VISC, which is characterized by an increase in F1 and a movement of F2 toward central values (a decrease for front vowels and in increase for back vowels). The current data provides some support for this pattern in the case of the vowels /ɪ/ and /ʊ/ in both formants. The adult males showed significant upward movement in F1 and toward central values for F2. The females showed similar weak (non-significant) trends in F1 and significant F2 movement. The F1 effects for males seems almost certainly due to vowel-inherent properties, since the effects of the final voiced stop should be toward lowering the F1 in the latter stages of the vowel. Presumably, only a vowel-directed gesture associated with an offglide would be sufficient to resist this trend. Indeed, all the vowels of both males and females except /ɪ/ and /ʊ/ showed downward movement of F1.

Consonantal context effects associated with the final /d/ locus render simple interpretation of F2 movement as VISC all but impossible, except in the cases of /i/, /e/, /o/, and /u/ discussed above, where VISC toward extra high or extra low F2 overcomes the contextual influences of a mid-frequency F2 locus (1,600–1,800 Hz for males and females). The alpha-VISC candidates /ɪ/, /ɛ/, /æ/,/ʊ/ and vowels not implicated in previous studies (/ʌ/, /ɚ/, /ɑ/, /ɔ/) show F2 change directions that are compatible with early effects of movement toward the /d/ locus.

5.2 Children

The analysis of F1 frequency for boys (12 talkers per age group) led to a significant effect of Age, $F(3, 31) = 27.89$; $p < 0.01$, Vowel, $F(11, 341) = 561.51$; $p < 0.01$; and Frame, $F(1, 31) = 294.03$; $p < 0.01$; a significant interaction of Age by Vowel, $F(33, 341) = 2.01$; $p < 0.01$; Vowel by Frame, $F(11, 341) = 52.37$; $p < 0.01$; but no Age by Frame interaction, $F(3, 31) = 1.35$; $p = 0.28$. The 3-way interaction of Age by Vowel by Frame was not significant, $F(33, 341) = 1.01$; $p = 0.45$.

For F2 frequency in boys there was a significant effect of Age, $F(3, 31) = 58.22$; $p < 0.01$, Vowel, $F(11, 341) = 453.77$; $p < 0.01$; and Frame, $F(1, 31) = 80.77$; $p < 0.01$; a significant interaction of Age by Vowel, $F(33, 341) = 2.89$; $p < 0.01$; Vowel by Frame, $F(11, 341) = 181.86$; $p < 0.01$; but no Age by Frame interaction, $F(3, 31) = 1.50$; $p = 0.23$. The 3-way interaction of Age by Vowel by Frame was significant, $F(33, 341) = 3.76$; $p < 0.01$.

For girls the pattern was similar: for F1, there was a significant effect of Age, $F(3, 35) = 3.47$; $p < 0.01$; Vowel, $F(11, 385) = 636.06$; $p < 0.01$; and Frame, $F(1, 35) = 291.32$; $p < 0.01$; a significant interaction of Age by Vowel, $F(33, 385) = 3.22$; $p < 0.01$; Vowel by Frame, $F(11, 385) = 74.23$; $p < 0.01$; no significant Age by Frame interaction, $F(3, 35) = 1.57$; $p = 0.21$. The 3-way interaction of Age by Vowel by Frame was not significant, $F(33, 385) = 1.32$; $p = 0.11$.

For girls, F2 showed a significant effect of Age, $F(3, 35) = 17.35$; $p < 0.01$; Vowel, $F(11, 385) = 911.34$; $p < 0.01$; and Frame, $F(1, 35) = 39.90$; $p < 0.01$; a significant interaction of Age by Vowel, $F(33, 385) = 3.39$; $p < 0.01$; Vowel by Frame, $F(11, 385) = 79.75$; $p < 0.01$; but no Age by Frame interaction, $F(3, 35) = 1.88$; $p = 0.15$. The 3-way interaction of Age by Vowel by Frame was not significant, $F(33, 385) = 1.22$; $p = 0.19$.

The key finding that emerges from these analyses is that all groups of children showed a significant interaction of Vowel by Frame for both F1 and F2, consistent with the presence of VISC. Three of the four 3-way interactions with Age were not significant, providing no evidence for different patterns of VISC as a function of age. The exception was a significant 3-way interaction for F2 frequency in boys. Figure 14 shows that younger boys (ages 5–8 years) showed steeper upward F2 transitions for some vowels (e.g., /ɑ/, /ɔ/, and /ʊ/) compared to older boys. Note that these age-related differences are not restricted to vowels that show significant VISC. However, they are most pronounced for vowels in which F2 undergoes a large excursion. The increase in slope was not associated with reversals in F2 direction for any individual vowel.

As presented for the adults in Tables 1 and 2, the detailed pattern of formant movement from the 20 to 70 % points in different vowels (pooled across the four age groups of children) was analyzed using *post hoc t*-tests in Tables 3 and 4. Similar to adults, boys and girls showed reliable shifts in F1 and F2 for the tense vowels /e/ and /o/. The pattern for /u/ was similar to that for adults, with a

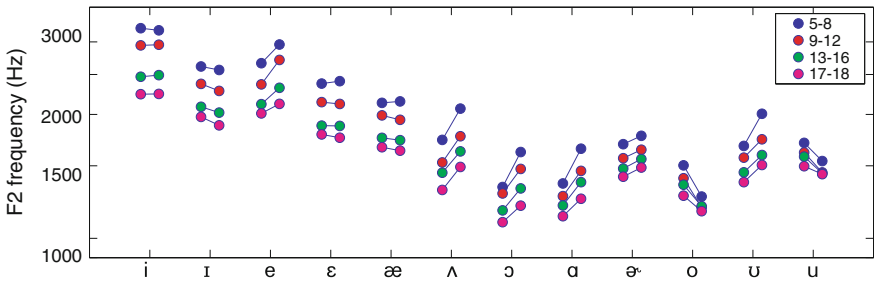


Fig. 14 Three-way interaction of age by slice by vowel for F2 for boys. *Circles joined by lines* indicate F2 frequencies sampled at the 20 and 70 % points. Each point is based on the median F2 frequency across each speaker's productions of the vowel, averaged across all speakers within the designated age group

Table 3 Initial (20 %) F1 and F2 measurements, along with means and standard deviations in ΔF1 and ΔF2 (70–20 % points)

Vowel	F1	Mean ΔF1	STD ΔF1	*t*	F2	Mean ΔF2	STD ΔF2	*t*
/i/	360	−4.1	26.7	−1.48	2835	−7.1	66.1	−1.04
/ɪ/	481	3.3	39.1	0.81	2336	−72.9	64.1	−11.02[b]
/e/	551	−117.4	63.5	−17.91[b]	2355	256.1	126.2	19.67[b]
/ɛ/	646	−39.9	55.8	−6.93[b]	2108	−0.9	52.5	−0.17
/æ/	903	−46.2	94.7	−4.73[b]	1945	−25.7	84.6	−2.94[b]
/ʌ/	687	−66.4	83.7	−7.69[b]	1535	255.2	150.8	16.41[b]
/ɔ/	803	2.2	81.5	0.26	1274	199.1	146.6	13.17[b]
/ɑ/	804	−2.3	90.3	−0.25	1258	213.0	136.0	15.19[b]
/ɚ/	539	−48.0	42.4	−10.98[b]	1566	80.9	68.3	11.48[b]
/o/	578	−126.0	57.7	−21.17[b]	1408	−191.3	110.7	−16.76[b]
/ʊ/	504	6.8	52.7	1.25	1542	215.1	143.7	14.51[b]
/u/	426	−24.9	31.1	−7.78[b]	1649	−165.8	143.6	−11.19[b]

(a) Boys (all age groups combined) ($N = 94$)
[a] $p < 0.01$; [b] $p < 0.05$

significant decrease in F1 and F2 for both boys and girls. The pattern for /i/ was also similar to adults for boys, however, girls showed a significant downward shift in F2 which runs counter to the VISC patterns predicted for this vowel, but this could be attributed to the influence of the final /d/.

For the lax vowels, the alpha VISC candidates /ɪ/ and /ʊ/ showed an F1 increase for both boys and girls, though the difference was not significant for boys. F2 showed a decrease for the front vowels /ɪ/ and /æ/, though /ɪ/ was significant only for boys. The vowel /ɛ/ showed a non-significant F2 increase for boys, but a significant and unexpected increase for girls. The source of this small but unexpected increase is unclear since the F2 at the /d/ closure is actually a bit lower on average than at the 70 % point (see Fig. 5). For /ʊ/, F2 showed an increase for both boys and girls, though it was significant only for the girls. Other vowels showed

Table 4 Initial (20 %) F1 and F2 measurements, along with means and standard deviations in ΔF1 and ΔF2 (70–20 % points)

Vowel	F1	Mean ΔF1	STD ΔF1	*t*	F2	Mean ΔF2	STD ΔF2	*t*
/i/	401	5.3	32.7	1.73	3096	−11.7	97.0	−1.29
/ɪ/	525	9.9	42.5	2.47[a]	2550	−78.1	57.9	−14.34[b]
/e/	587	−123.0	67.0	−19.51[b]	2640	265.3	123.3	22.87[b]
/ɛ/	731	−62.7	62.3	−10.70[b]	2281	16.5	60.3	2.91[a]
/æ/	1052	−64.7	79.8	−8.62[b]	2060	−2.0	79.7	−0.26
/ʌ/	764	−83.7	71.4	−12.47[b]	1745	272.7	125.7	23.05[b]
/ɔ/	865	−4.1	105.3	−0.41	1360	287.0	275.5	11.07[b]
/ɑ/	894	−10.5	93.9	−1.19	1387	290.9	278.5	11.11[b]
/ɚ/	595	−59.5	53.9	−11.62[b]	1776	100.5	94.4	11.21[b]
/o/	626	−142.1	68.0	−22.22[b]	1554	−213.7	98.1	−23.14[b]
/ʊ/	553	9.3	48.3	2.04[a]	1737	248.1	131.8	20.01[b]
/u/	466	−18.8	32.4	−6.16[b]	1825	−216.0	171.5	−13.39[b]

(b) Girls (all age groups combined) ($N = 114$)
[a] $p < 0.01$; [b] $p < 0.05$

formant movement similar to the patterns displayed by adults in the direction expected for the contextual influence of the final /d/.

As in the case of adults, with the exception of /ɛ/, F2 movements for the lax vowels are largely consistent with alpha-VISC patterns and with anticipation of the final /d/ locus. However, only /ɪ/ and /ʊ/ showed the F1 increase expected with alpha-VISC. Of the 48 ΔF1 values reported in Tables 1–4, only 9 showed positive movement. Eight of these represent the vowels /ɪ/ and /ʊ/, which showed positive movement for all speaker groups. The only other case of a positive F1 movement was observed in /ɔ/ for boys. The other groups all showed small negative movements for this vowel.

Overall, the pattern of formant movement is broadly comparable for adults and children, with the tense vowels generally following the VISC patterns predicted by earlier studies. For the lax vowels the pattern was more complex, both in adults and children. However, where significant discrepancies occurred, the direction could often be predicted by the contextual influence of the final /d/. Inconsistencies across sex and age groups may indicate that not all speakers employ the same formant movement patterns, or that they vary in the relative "pull" exerted by the final /d/ transitions. To unravel these factors it will be necessary to study the acoustic properties of vowels across a range of consonant contexts, and we are currently exploring these possibilities using a nonlinear regression model developed by Nearey (2013 Chap. 4).

6 Summary and Conclusions

The results indicate consistent evidence for VISC patterns in both adults and children from the North Texas region. Vowels displaying VISC include, as expected, the phonemic diphthongs /aɪ/, /aʊ/, and /ɔɪ/ as well as /e/ and /o/.

The vowel /u/ showed evidence of some vowel inherent lowering of F2 not attributable to consonantal context, while /i/ showed no real evidence of F1 movement and a small downward F2 movement that could be attributed to the influence of the final /d/ context (a pattern which is not consistent with the [ij]-like characterization of /i/ in some dialects). The lax vowels /ɪ/ and /ʊ/ showed some evidence of rising F1 while all other vowels showed consistently (or predominantly in the case of /ɔ/) falling F1 patterns. The consistent F2 movements in these vowels, as well as /æ/, were compatible with patterns of alpha-VISC but could also reflect the influence of the final /d/ locus.

Formant movement patterns were fairly consistent between adults and children and across age/sex groups in children, although younger children showed greater variability. There were systematic differences in the effects of age on the formant trajectories for different vowels, however, age-related changes in formant frequencies did not interact with VISC or context-directed formant movement patterns.

Pattern recognition analyses indicated that vowels were more accurately recognized when two analysis frames, sampled around 20 and 70 % of the vowel duration, were presented to the classifier, compared to any *single* frame; (2) adding a third analysis frame did not yield substantially higher recognition scores; and (3) the optimum locations for sampling the formant trajectory were consistent across different age groups of children.

Most of the formant trajectories with substantial movement between the 20 and 70 % points appeared to follow trajectories that did not differ markedly from a straight line approximation. The one exception to this was the diphthong /ɔɪ/, which showed a pronounced bowed pattern with F1 increasing until about 50 % of the duration. Overall, a simple two-point description appeared to characterize the vowels effectively, and no more complex information seems necessary to account for their distinctiveness. However, more elaborate characterizations such as the path length measure used by Fox and Jacewicz (2009, see also Jacewicz and Fox 2013 Chap. 4) may be needed to characterize subtle but consistent differences that are required to distinguish vowels in other dialects.

6.1 Limitations

One limitation of the present study is that the vowel recordings were restricted to a single context, /hVd/. An issue that has not been adequately addressed in the literature is how VISC interacts with the effects of adjacent consonants (generating different patterns of coarticulation) to produce distinctive patterns of formant frequency movement. The findings confirm that the interpretation of VISC patterns is complicated by the influence of the final consonant. A modeling approach is needed to unravel these effects in a range of consonant contexts.

It has been reported that children exhibit different patterns of coarticulation compared to adults (e.g., Sussman et al. 1992; Gibson and Ohde 2007), but the

interaction of consonant context with VISC has not been systematically investigated. We are currently recording an expanded set of vowels in a range of CVC contexts and analyzing the relative contribution of VISC and consonant context using a modeling approach (Nearey 2013 Chap. 4). We are also planning a series of perceptual experiments to determine whether vowel and consonant transitions contribute independently or interact in vowel perception.

Acknowledgments This work was supported by National Science Foundation grants 0318451 and 1124479 and Social Sciences and Humanities Research Council (Canada) grant 410-2000-1353. We thank Geoffrey Stewart Morrison and Robert Allen Fox for their comments on an earlier version of the chapter.

References

Andruski, J.E., Nearey, T.M.: On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables. J. Acoust. Soc. Am. **91**, 390–410 (1992). doi:10.1121/1.402781

Assmann, P., Ballard, W., Bornstein, L., Paschall, D.: Track-draw: A graphical interface for controlling the parameters of a speech synthesizer. Behav. Res. Methods Instrum. Comput. **26**, 431–436 (1994). doi:10.3758/BF03204661

Assmann, P.F., Katz, W.F.: Time-varying spectral change in the vowels of children and adults. J. Acoust. Soc. Am. **108**, 1856–1866 (2000). doi:10.1121/1.1289363

Assmann, P.F., Katz, W.F.: Synthesis fidelity and vowel identification. J. Acoust. Soc. Am. **117**, 886–895 (2005). doi:10.1121/1.1852549

Enzinger, E.: Characterizing formant tracks in Viennese diphthongs for forensic speaker comparison. In: Proceedings of the 39th Audio Engineering Society Conference—Audio Forensics: Practices and Challenges, Hillerød, Denmark, pp. 47–52. Audio Engineering Society, New York (2010)

Fant, G.: A note on vocal tract size factors and non-uniform F-pattern scalings. Speech Music Hear. Q. Prog. Status Rep. (STL-QPSR), **7**(4), 22–30 (1966)

Fitch, W.T., Giedd, J.: Morphology and development of the human vocal tract: A study using magnetic resonance imaging. J. Acoust. Soc. Am. **106**, 1511–1522 (1999). doi:10.1121/1.427148

Fox, R.A., Jacewicz, E.: Cross-dialectal variation in formant dynamics of American English vowels. J. Acoust. Soc. Am. **126**, 2603–2618 (2009). doi:10.1121/1.3212921

Gay, T.: Effect of speaking rate on diphthong formant movements. J. Acoust. Soc. Am. **44**, 1570–1573 (1968). doi:10.1121/1.1911298

Gibson, T., Ohde, R.N.: F2 locus equations: phonetic descriptors of coarticulation in 17- to 22-month-old children. J. Speech Lang. Hear. Res. **50**, 97–108 (2007). doi:10.1044/1092-4388(2007/008

Green, J.R., Moore, C.A., Reilly, K.J.: The sequential development of jaw and lip control for speech. J. Speech Lang. Hear. Res. **45**, 66–79 (2002). doi:10.1044/1092-4388(2002/005

Hillenbrand, J.M.: Static and dynamic approaches to understanding vowel perception. In: Morrison, G.S., Assmann, P.F (eds.) Vowel Inherent Spectral Change (Chap. 2). Springer, Heidelberg (2013)

Hillenbrand, J.M., Clark, M.J., Nearey, T.M.: Effects of consonant environment on vowel formant patterns. J. Acoust. Soc. Am. **109**, 748–763 (2001). doi:10.1121/1.1337959

Hillenbrand, J.M., Nearey, T.M.: Identification of resynthesized /hVd/ utterances: Effects of formant contour. J. Acoust. Soc. Am. **105**, 3509–3523 (1999). doi:10.1121/1.424676

Holbrook, A., Fairbanks, G.: Diphthong formants and their movements. J. Speech Hear. Res. **5**, 33–58 (1962)
Hollien, H., Green, R., Massey, K.: Longitudinal research on adolescent voice change in males. J. Acoust. Soc. Am. **96**, 2646–2654 (1994). doi:10.1121/1.411275
Jacewicz, E., Fox, R.A.: Cross-dialectal differences in dynamic formant patterns in American English vowels. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chapter 8). Springer, Heidelberg (2013)
Johnson, K.: Speaker normalization in speech perception. In: Pisoni, D.B., Remez, R. (eds.) The Handbook of Speech Perception, pp. 363–389. Blackwell Publishers, Oxford (2005). doi:10.1002/9780470757024.ch15
Johnson, K.: Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. J. Phonetics **34**, 485–499 (2006). doi:10.1016/j.wocn.2005.08.004
Katz, W.F., Assmann, P.F.: Identification of children's and adults' vowels: Intrinsic fundamental frequency, fundamental frequency dynamics, and presence of voicing. J. Phonetics **29**, 23–51 (2001). doi:10.1006/jpho.2000.0135
Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T., Irino, T.: Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. In: Proceedings of Interspeech 2005, Lisboa, pp. 537–540, Sept 2005
Klatt, D.H., Klatt, L.C.: Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. **87**, 820–857 (1990). doi:10.1121/1.398894
Labov, W., Ash, S., Boberg, C.: Atlas of North American English. Mouton de Gruyter, New York (2006)
Ladefoged, P., Johnson, K.: A Course in Phonetics, 6th edn. Thomson Wadsworth, Boston (2010)
Lee, S., Potamianos, A., Narayanan, S.: Acoustics of children's speech: Developmental changes of temporal and spectral parameters. J. Acoust. Soc. Am. **105**, 1455–1468 (1999). doi:10.1121/1.426686
Lee, S., Narayanan, S., Byrd, D.: A developmental acoustic characterization of English diphthongs. J. Acoust. Soc. Am. **115**, 2628 (A) (2004)
Lehiste, I., Peterson, G.E.: Transitions, glides and diphthongs. J. Acoust. Soc. Am. **38**, 268–271 (1961). doi:10.1121/1.1908681
Morrison, G.S.: Theories of vowel inherent spectral change: A review. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chapter 3). Springer, Heidelberg (2013a)
Morrison, G.S.: Vowel-inherent spectral change in forensic voice comparison. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chapter 11). Springer, Heidelberg (2013b)
Nearey, T.M.: Vowel inherent spectral change in the vowels of North American English. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chapter 4). Springer, Heidelberg (2013)
Nearey, T.M., Assmann, P.F.: Modeling the role of vowel inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986). doi:10.1121/1.394433
Nearey, T.M., Assmann, P.F., Hillenbrand, J.M.: Evaluation of a strategy for automatic formant tracking. J. Acoust. Soc. Am. **112**, 2323 (2002)
Nittrouer, S.: The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study. J. Speech Hear. Res. **36**, 959–972 (1993)
Perry, T.L., Ohde, R.N., Ashmead, D.H.: The acoustic bases for gender identification from children's voices. J. Acoust. Soc. Am. **109**, 2988–2998 (2001). doi:10.1121/1.1370525
Peterson, G.E., Barney, H.L.: Control methods used in a study of the vowels. J. Acoust. Soc. Am. **24**, 175–184 (1952). doi:10.1121/1.1917300
Peterson, G., Lehiste, I.: Duration of syllable nuclei in English. J. Acoust. Soc. Am. **32**, 693–703 (1960). doi:10.1121/1.1908183
Potamianos, A., Narayanan, S.: Robust recognition of children's speech. IEEE Trans. Speech Audio Proc. **11**, 603–616 (2003). doi:10.1109/TSA.2003.818026

Potter, R.K., Peterson, G.E.: The representation of vowels and their movements. J. Acoust. Soc. Am. **20**, 528–535 (1948). doi:10.1121/1.1906406

Prator Jr, C.H., Robinett, B.W.: Manual of American English Pronunciation, 4th edn. Harcourt Brace, Tokyo (1985)

Sussman, H.M., Hoemeke, K.A., McCaffrey, H.A.: Locus equations as an index of coarticulation for place of articulation distinctions in children. J. Speech Hear. Res. **35**, 769–781 (1992)

Sussman, H.M., Bessell, N., Dalston, E., Majors, T.: An investigation of stop place of articulation as a function of syllable position: A locus equation perspective. J. Acoust. Soc. Am. **101**, 2826–2838 (1997). doi:10.1121/1.418567

Thomas, E.R.: A rural/metropolitan split in the speech of Texas Anglos. Lang. Var. Change **9**, 309–332 (1997). doi:10.1017/S0954394500001940

Thomas, E.R.: Secrets revealed by southern vowel shifting. Am. Speech **78**, 150–170 (2003). doi:10.1215/00031283-78-2-150

Tiffany, W.: Vowel recognition as a function of duration, frequency modulation and phonetic context. J. Speech Hear. Disord. **18**, 289–301 (1953)

Trager, G.L., Smith, H.L.: An outline of English structure. Studies in Linguistics: Occasional Papers 3. Norman. Battenburg Press, Oklahoma (1951)

Vorperian, H., Kent, R.D.: Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. J. Speech Lang. Hear. Res. **50**, 1510–1545 (2007). doi:10.1044/1092-4388(2007/104

Watson, C., Harrington, J.: Acoustic evidence of dynamic formant trajectories in Australian English vowels. J. Acoust. Soc. Am. **44**, 458–468 (1999). doi:10.1121/1.427069

Zahorian, S., Jagharghi, A.: Spectral-shape features versus formants as acoustic correlates for vowels. J. Acoust. Soc. Am. **94**, 1966–1982 (1993). doi:10.1121/1.407520

Vowel Inherent Spectral Change and the Second-Language Learner

Catherine L. Rogers, Merete M. Glasbrenner, Teresa M. DeMasi and Michelle Bianchi

Abstract Because vowel inherent spectral change (VISC) is necessary for optimal identification of vowels by native English speakers, learners of English as a second language must acquire relevant information about VISC in order to achieve native-like levels of performance in both perception and production of vowels in English. This chapter reviews studies of both perception and production of VISC by learners of English as a second language, whose first language is Spanish, with either an earlier or later age of immersion in an English speaking environment. In perception, later learners of English appeared to rely more heavily on duration cues than monolinguals and early learners and, in some cases, to be less able to use VISC to discriminate near neighbors in the vowel space. In production, acoustic analyses were performed for American English vowels produced by participants in each group. The data are examined in terms of the degree of separation achieved by each talker group across the course of the vowel, as represented by three time points (20, 50 and 80 % of vowel duration). Additional analyses of productions by the most and least intelligible talkers in each group were used to explore individual talkers' strategies for using VISC to distinguish neighbor vowels from one another.

Abbreviations

CV	Consonant-vowel
EL	Early learners of English
f0	Fundamental frequency
F1	First formant
F2	Second formant

C. L. Rogers (✉) · M. M. Glasbrenner · T. M. DeMasi · M. Bianchi
Department of Communication Sciences and Disorders, University of South Florida, 4202 E. Fowler Avenue, PCD1017, Tampa, FL 33620, USA
e-mail: crogers2@usf.edu

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_10,

MO Monolingual English speakers
LL Later learners of English
VC Vowel-consonant
VISC Vowel inherent spectral change

1 Introduction

1.1 Dynamic Information in Speech

Speech is inherently dynamic. The jaw and other articulators move continuously in space and time through several articulatory postures per second. A speaking rate of two to three syllables per second, for example, implies approximately seven to eleven such articulatory postures per second, depending on the complexity of the syllable (i.e., in terms of traditional phonetic description, combinations of place of articulation, manner of articulation, and voicing for consonants, or tongue height, tenseness, and backness for vowels). In the rich stream of acoustic information generated by the process of articulation, many types of information are used by listeners to identify speech sounds—from the very brief, such as the transients generated by the release of stop consonants, to the more distributed, such as differences in direction and slope of formant transitions, which may cue both place and manner of articulation of consonants (cf. Pickett 1999 for an overview).

Until the early 1980s, much, but certainly not all, of the research on vowel perception rested on the assumption that the vast majority of information about the monophthongal vowels of English was to be conveyed by the formant values achieved near the middle of the vowel nucleus, or "steady state," despite Peterson and Barney's (1952) and others' acknowledgment of the insufficiency of formant values from a single time slice to distinguish among productions of multiple talkers (cf. Hillenbrand 2013 (Chap. 2) for a review of this and two other earlier findings). Thus, both the formant transitions to and from any neighboring phonemes and any systematic change in formant frequencies occurring during the vocalic nucleus were considered to be of secondary importance at best, with regard to conveying information about vowel identity. The notion of the necessity of static vowel nuclei as the locus of vowel perception was overturned, however, with the publication of the "silent center syllable" studies of Strange and colleagues (Jenkins et al. 1983; Strange et al. 1983). The silent center studies showed that vowel identification could proceed reasonably well, even in the absence of acoustic energy from the middle of CVC syllables, despite that energy being thought to most closely approximate a speaker's "target" steady state values.

Conversely, Nearey et al. (Assmann et al. 1982; Nearey and Assmann 1986) showed that vowel identification improved when dynamic formant information occurring during the vowel nucleus was provided than when static vowel formant

information was provided. They used the term "vowel inherent spectral change" to refer to the dynamic formant information occurring during the vowel nucleus. Ongoing debate over the interpretation of the findings of Nearey and colleagues vs. those of Strange and colleagues notwithstanding, there is no longer a question of whether formant dynamics play some role in vowel perception and production (cf. Strange and Jenkins 2013 (Chap. 4); Morrison 2013 (Chap. 3). A Review; Hillenbrand 2013 (Chap. 2); Nearey 2013 (Chap. 4)). As such, the work of these researchers has radically altered prevailing notions about vowel perception. Perhaps not surprisingly, however, the effects of this change in thinking about vowel perception were not immediately applied to the study of speech produced and/or perceived by populations with significant linguistic, motor or perceptual differences from adult native speaker norms. These populations include learners of a second language, persons with hearing loss or cochlear implants, persons with motor speech disorders, and children learning their first language. The last five to ten years have seen progress in this area, and, these results in turn will continue to inform the development of models of speech perception and production that can more realistically represent the use of dynamic information in speech. This chapter focuses on research on the use of dynamic spectral information in vowels by learners of English as a second language, but the goal is to view these results more generally, as they may apply to different populations.

1.2 What is Vowel Inherent Spectral Change?

Figure 1 illustrates productions of two /bVd/ syllables. The area between the arrows in each case roughly corresponds to the vocalic nucleus, containing the "target" formant frequencies, while the voiced regions preceding and following the arrows correspond roughly to the consonant–vowel (CV) and vowel–consonant (VC) formant transitions, respectively. An examination of Fig. 1 indicates substantial change in formant frequency during both the transition portions and the vowel nucleus. Since the initial research of Nearey and colleagues, vowel inherent spectral change has been accepted as a feature of the vowels of North American English and has been shown to (1) vary systematically across the vowels of a given dialect, (2) vary systematically across dialects (cf. Hillenbrand 2013 (Chap. 2); Jacewicz and Fox 2013 (Chap. 8), and (3) be necessary for optimal vowel-identification performance (Assmann and Katz 2005; Hillenbrand et al. 2001; Hillenbrand and Nearey 1999). Thus, vowel inherent spectral change (VISC) differs from CV and VC transitions and refers to systematic changes in formant frequency that occur largely between the CV and VC formant transitions.

First-formant (F1) frequencies of vowels can be described as corresponding roughly with tongue/jaw height (with lower first formant values for higher positions of the tongue/jaw), while second-formant (F2) frequencies are associated with frontness of tongue constriction (with higher second formant values associated with more front constrictions), although there may be some variations from

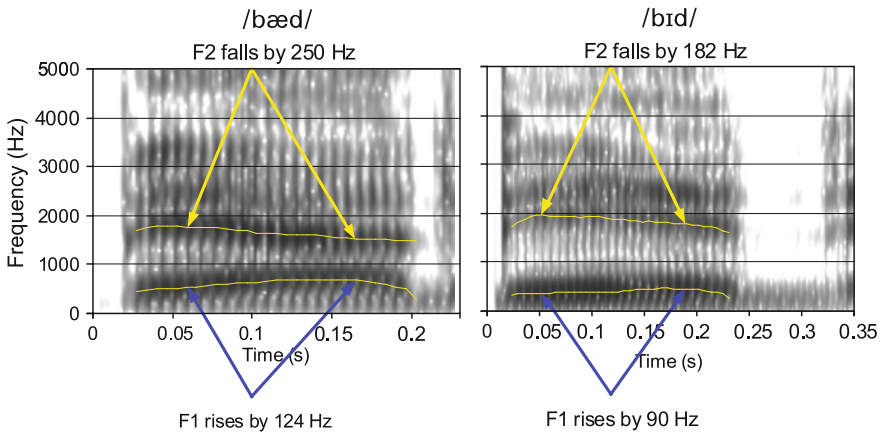


Fig. 1 Spectrograms of two syllables that exhibit similar degrees of vowel-inherent spectral change during the high-intensity vowel nucleus. The approximate region of vowel-inherent spectral change is indicated by the regions of F1 and F2 between the *arrows* in each case. The formant regions outside the *arrows* roughly indicate CV and VC formant transitions in each case

these patterns across speakers and contexts. Consequently, changes in F1 and F2 values over time may be roughly associated with direction of movement of the tongue constriction (up or down and forward or back, respectively). These relationships are evident in the CV and VC transitions in Fig. 1, with rising (CV) and falling (VC) F1 transitions indicating decreasing constriction following the release of the /b/ and increasing constriction prior to the closure for the /d/. For both syllables, VISC is most evident in the decrease in F2 over the vocalic nucleus, translatable as a slight backing gesture, or centralization of tongue position (or alpha-VISC in the vocabulary of Nearey 2013 (Chap. 4)). Note that this gesture is the opposite of what might be expected as a consequence of coarticulation for the upcoming /d/ consonant. Therefore, VISC may be hypothesized to be employed as a distinguishing feature in some languages but not others, while CV and VC formant transitions must be employed to some degree in every language. Consequently, learners of English as a second language may be assumed to "know" about the type of information encoded by CV and VC formant transitions, but not necessarily about the information encoded by VISC.

Specifically, with regard to formant transitions, some degree of change in formant frequency would seem to be an *unavoidable* consequence of following a consonant with a vowel, or vice versa, because articulators cannot move instantaneously from one posture to another. Talkers might exaggerate or minimize the degree of transition depending on speaking rate, style, etc., but some degree of transition from one articulatory posture to the next is essential and should be observed in every language. As noted above, the gesture underlying the VISC in "bad" and "bid" (see Fig. 1) appears to be the opposite of what would be anticipated from CV or VC coarticulatory forces alone. Rather than being an obligatory consequence of movement from one articulatory posture to the next,

VISC is a gesture that occurs *within* the vocalic nucleus for a given vowel, and which may be employed as a distinguishing feature of vowels in some languages but not others.

If VISC is assumed to be a more language-specific feature than formant movement due to coarticulation from consonants to vowels, then second-language learners may be expected to face challenges in learning to encode the vowel-specific information contained in VISC, if it is not a feature of the vowels of their native language. On the other hand, the information contained in formant movement due to consonant-vowel coarticulation might be expected to be more universally available to learners from different language backgrounds. While these hypotheses may or may not be supported empirically, the main point of the above discussion is to suggest that not all dynamic information in speech is necessarily created equal. The information encoded by some dynamic aspects of speech may be more readily learned or interpreted by second-language learners than other dynamic aspects. Now that some of the production-related aspects of VISC have been considered, VISC will be considered from the perspective of vowel perception.

1.3 VISC and Vowel Perception

Perceptually, VISC has been shown to be *used* for vowel identification, however, the consequences of removing VISC are moderate on average (about 15–25 % reduction in correct-classification rates overall), and highly variable across target vowels (near 0 % for some vowels and near 100 % for others), with most somewhere between the two extremes (Hillenbrand and Nearey 1999; Assmann and Katz 2000, 2005). Such effects are reasonable. For most vowels traditionally described as monophthongs in English, the formant ratios present throughout most portions of the vowel nucleus place the vowel in a relatively specific region of the vowel space, while VISC depicts movement *within* that region of the vowel space. Thus, the formant ratios narrow the field of possible vowels substantially, even if VISC is not appropriate for the vowel in question. On the other hand, appropriate VISC alone is unlikely to be useful in vowel identification if the format ratios present in the vowel nucleus place the vowel in an entirely inappropriate region of the vowel space.

Congruent with the idea of vowel formant information narrowing a field of competitors, another perspective on vowel recognition presented in this volume is that of "efficient coding," (cf. Kluender et al. 2013 (Chap. 6)), according to which speech sounds are depicted not in terms of a list of features defining their identity, but rather by how they contrast with one another. Under this view, speech sounds are not so much *defined* as distinguished. A speech sound is well identified if it is sufficiently distinct from all of its neighbors. According to this type of theory, formant ratios throughout the vowel nucleus would need to place the vowel in an appropriate region of the vowel space for correct identification, but somewhat

different strategies may be successful in achieving the goal of adequately distinguishing a vowel from its nearest neighbors (e.g., some talkers might employ duration more as a distinguishing feature than would others).

In a study that is compatible with this perspective, Neel (2008) examined the relationship between several acoustic characteristics of the American English vowels used by Hillenbrand et al. (1995) and the intelligibility of those vowels to native English-speaking listeners. Neel (2008) characterized the acoustic variables studied as either global or fine-grained. One result of her study was that fine-grained acoustic characteristics related to the distinctiveness of *neighboring pairs of vowels* predicted overall intelligibility of the vowel set better than global measures. Among those "neighbor distinguishing" variables, she found VISC and vowel duration differences between intrinsically long and short vowels to be particularly important in predicting intelligibility of the full set of vowels.

Whether under the view of efficient coding or a more traditional feature description, VISC has been interpreted as a means of enhancing separation among sets of vowels produced by multiple talkers (cf. Kluender et al. 2013 (Chap. 6); Hillenbrand 2013 (Chap. 2); Neel 2008). As shown by Peterson and Barney (1952) and others, substantial overlap is observed across talkers in the regions of formant space occupied by different, neighboring vowels. Perhaps, for the fairly densely populated vowel space of English, it is nearly impossible to achieve a degree of separation among all possible neighboring vowels that is adequate for reliable identification, given inter-talker variability in production. The role of VISC could then be to achieve adequate separation among all neighboring vowels over the course of the entire vowel, although each may not be distinguished from every possible neighbor at any single point in time. Some neighbors may be distinguished from one another early in the vowel, but overlap in production later, while other neighbors may be better distinguished later in the vowel but overlap in the early part of the vowel. This additional information may be particularly useful in difficult listening environments, such as noise, where portions of the signal may be obscured. In this way, VISC can be seen as the steps in a sort of 'dance,' in which each vowel moves to avoid overlapping with another, ultimately causing overlap with another and hence more movement to the next time point.

Assuming a goal of achieving good separation among all neighboring vowel regions of the vowel space over the entire course of vowel production, it is possible that different talkers may use different strategies, realized by different patterns of VISC, to achieve that goal. For this argument, let us assume that a perceptual region of vowel space is attuned to a particular vowel and that this region is roughly defined by one's lifetime's experience with exemplars of that vowel in one's native dialect. That is, the region attuned to a particular vowel must encompass a range of cross-talker differences, and may roughly define a "native speaker norm." However, a given talker's production of /ɪ/ might lie near the edge of the specified region for /ɪ/ that borders the region attuned to /i/. In order to achieve good separation among one's own vowels, and vis à vis native-speaker norms, a different pattern of VISC may be appropriate for this talker's /ɪ/

production than for a talker whose production of /ɪ/ lies nearer the region of the vowel space attuned to /ɛ/.

2 VISC and Vowel Perception by Second-Language Learners

With regard to second-language learners, recall that in the above discussion, VISC was hypothesized to be a more language-specific dynamic feature of American English vowels than CV and VC formant transitions, making it likely that second-language learners may face significant challenges in learning to encode the vowel-specific information contained in VISC. Nevertheless, learners of English as a second language must acquire relevant information about VISC in order to achieve native-like levels of performance in both perception and production of vowels in English. Second, the work of Neel (2008) has shown that acoustic features related to distinctiveness of neighboring vowel pairs were better predictors of overall vowel intelligibility than global features. With these observations in mind, it would seem that investigations of the use of VISC by second-language learners may enhance understanding of acquisition of native-like competence in vowel perception.

Relatively few studies have examined the role of VISC explicitly in either perception or production by second-language learners. One exception, however, is work conducted at the University of Alberta. Morrison (2009), for example, created a large set of synthetic stimuli with formant values ranging from those appropriate for English /i/ to those appropriate for English /ɛ/, and either diverging F1 and F2 (typical of English /eɪ/), zero formant movement (typical of English /i/) or converging F1 and F2 (typical of English /ɪ/) VISC properties. Labeling responses were collected from monolingual English speakers and non-native English speakers with an L1 of Spanish. Spanish labeling responses were also collected from monolingual Spanish speakers. Logistic regression analyses were used to create labeling maps showing the portions of the stimulus space assigned to each vowel category. Monolingual English speakers demonstrated the expected patterns, with more stimuli being labeled with /eɪ/ responses when diverging VISC formant trajectories were used than when zero VISC formant trajectories were used and /ɪ/ responses being obtained almost exclusively when converging VISC formant trajectories were used. Data for the non-native listeners showed a wide range of response patterns, with some very similar to those of the native listeners, others appearing to partition the space very similarly to the monolingual Spanish speakers, and many with response patterns lying between those two extremes. Although cross-sectional, these data support the hypothesis that non-native listeners' use of VISC for vowel identification differs from that of native English listeners, is highly variable across listeners, and may evolve over time to approximate native-speaker norms.

Morrison (2008) also found that monolingual Spanish speakers with different dialects (Mexican vs. Peninsular Spanish) partitioned the stimulus space

differently with regard to Spanish labeling, suggesting differing initial assimilation patterns across the two dialects, and thus possibly different patterns of acquisition for English vowels.

Finally, both Morrison (2006) and Thomson et al. (2009) integrated VISC data into representations of vowels produced by native and non-native English speakers. In both cases, the VISC data were integrated into statistical modeling procedures used to quantify distances between native and non-native productions and to predict their categorization by native English-speaking listeners.

The remainder of the chapter will focus on research conducted at the University of South Florida, comparing the perception and production of VISC and other acoustic features of vowels by native and non-native English speakers.

2.1 *Perception of Silent-Center Syllables by Native and Non-Native Listeners*

One working hypothesis of the perceptual studies described below was that the vowel-specific information encoded by VISC might be a particularly challenging feature for learners of English as a second language to acquire. This hypothesis was based on the suggestion, above, that VISC may be a more language-specific feature than changes in formant frequency arising primarily from coarticulatory effects and on evidence that duration may be a relatively salient cue even for early learners of a second language (Bohn and Flege 1990). An additional hypothesis arose from the literature on speech perception by children (Lowenstein and Nittrouer 2010; Edwards et al. 2002; Hazan and Barrett 2000) suggesting that children may be more adversely affected by syllable disruption or limited acoustic information than adults, even when children appear to use the acoustic information present in the syllable similarly to adults. In Edwards et al. (2002), for example, adults and 7–8 year old children performed nearly identically when the entire syllable was presented. When a portion of the final-consonant formant transition was removed, however, the performance of the children declined much more than that of the adults.

Based on the results described above for children learning their first language, a second hypothesis was that non-native listeners, especially those still in the process of acquiring phonetic categories, might be more adversely affected by syllable disruption than native English-speaking listeners, even when a large portion of the relevant acoustic cue remained available. Two perception studies carried out at the University of South Florida explored these hypotheses. In the first (Rogers and Lopez 2008), the effects of syllable disruption were examined by comparing the perception of silent-center syllables by native and non-native English-speaking listeners. In the second (Glasbrenner 2005), the effects of vowel isolation and neutralization of VISC and duration were compared for native and non-native English-speaking listeners.

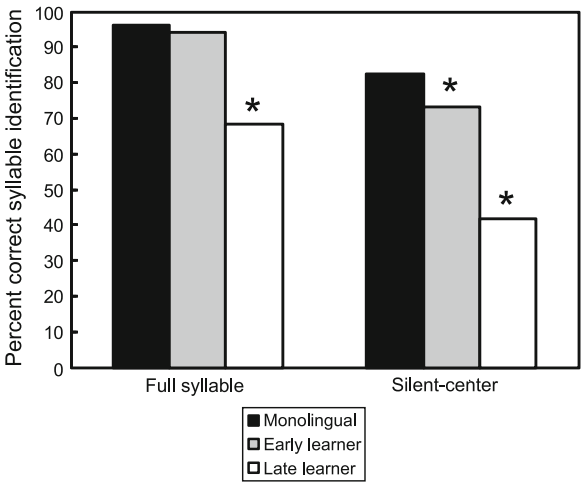


Fig. 2 Percent-correct syllable-identification rates for full and silent-center syllables by monolingual English-speaking listeners, relatively early learners of English as a second language and later learners of English as a second language. *Asterisks* indicate statistically significant differences in performance between listener groups, for each listening condition

Rogers and Lopez (2008) compared identification of six American English vowels by monolingual English speakers and by speakers of English as a second language, whose first language was Spanish and who had either an earlier or a later age of immersion in an English-speaking environment. Full (unedited) and silent-center versions of six target syllables (/bib/, /bɪb/, /beɪb/, /bɛb/, /bæb/ and /bɑb/) were created using methods similar to those used in previous silent-center syllable studies (cf. Strange 1989). Additional monolingual English-speaking participants rated the accentedness of sentences produced by the participants in the silent-center syllable-identification portion of the study. Accentedness ratings for the early learners fell within the bottom third of the 9-point rating scale, indicating little or no foreign accent, with several falling within the range of scores obtained for the native speakers. Ratings for the later learners all fell within the upper half of the scale, indicating a moderate to strong degree of foreign accent.

For each target syllable, four silent-center syllables were created by preserving 10, 20, 30 or 40 ms of the CV and VC formant transitions and reducing to silence the remainder of the syllable center. Figure 2 shows the average percent-correct identification rates for the full syllables and silent-center syllables for the monolingual, early-learner and later-learner participants, averaged across the six syllables and four silent-center conditions. Asterisks indicate statistically significant differences in performance across listener groups in each case.

As can be seen from Fig. 2, the monolingual English-speaking listeners and the early learners of English as a second language identified full syllables with a high degree of accuracy (90–95 % correct). No significant differences in performance were found between the two groups for the full syllable condition. Both groups, however, identified the full syllables about 25 % more accurately than the later learners, a difference that was statistically significant. For the silent-center syllables, on the other hand, the native listeners outperformed the early learners by about 10 %, a statistically significant difference. Both the native and early learner

groups outperformed the later learners by an even greater margin than in the full syllable condition (about 30–40 %). Thus, although the performance of the early learners declined more than that of the monolingual listeners, the later learners' performance declined most across the two conditions.

Furthermore, an analysis comparing performance only in the full and 40 ms transition conditions showed a significant three-way interaction of listener group, listening condition and vowel identity. A post hoc analysis of that interaction showed no significant differences in performance between the monolinguals and early learners, despite the significant difference obtained when performance was averaged across the four silent-center conditions. From these results, it seems that the early learners were more robust to the effects of syllable disruption than the later learners, but that they may have greater difficulty than the monolingual listeners in identifying syllables based on only partial cue information, as evidenced by their poorer performance at the shortest conditions (cf. Rogers and Lopez 2008 for details).

2.2 *Effects of Removing Consonant Context, Duration and VISC Information on Vowel Identification by Native and Non-Native Listeners*

The results of Rogers and Lopez (2008) suggested that at least some learners of English as a second language may be less robust to syllable disruption than monolingual English-speaking listeners. Apart from a greater sensitivity to any sort of disruption to syllable integrity, one reason for this difference may be that non-native listeners process dynamic information such as VISC differently than native English speakers. That is, if the non-native listeners were not perceptually attuned to the information provided by VISC that remained in the silent-center syllables, then their perception may have been more disrupted by the removal of the more distinctive, or "target," formant information contained in the vowel center than were the native English-speaking listeners. These hypotheses were investigated by comparing native and nonnative English-speaking listeners' identification of isolated vocalic nuclei, from which the CV and VC formant transitions (rather than the vocalic nucleus) had been removed. Listeners' identification of the isolated vowel nuclei was also compared following neutralization of VISC and vowel duration information (Glasbrenner 2005). Accuracy of identification and confusion patterns were examined for six /bVd/ syllables (/bib/, /bɪb/, /beɪb/, /bɛb/, /bæb/ and /bɑb/) as spoken by two female and two male monolingual English speakers. The syllables were presented to monolingual English-speaking listeners and both relatively early and later learners of English as a second language. As in Rogers and Lopez (2008), Spanish was the first language of all non-native English-speaking participants, and early learners were defined at

those with an age of immersion of age 12 or earlier and later learners as those with an age of immersion of age 15 or later.

To manipulate the availability to listeners of VISC and duration information, the vocalic nucleus was excised from each syllable and used to create five stimulus conditions, in addition to a full-syllable condition: (1) natural excised vowel (CV and VC formant transitions removed), (2) vocoded excised vowel, (3) duration neutralized, (4) flattened formant, and (5) flattened formant plus duration neutralized. The natural excised vowels were created by signal editing, preserving the middle 50 % of the vocalic nucleus only. The excised vowels were modified to create the remaining conditions using STRAIGHT software (Kawahara et al. 1999), a vocoder implementation that produces natural-sounding vocoded versions of speech input. Once coded in STRAIGHT, some acoustic characteristics (e.g., duration, f0) of an input speech sample may be modified using a graphical user interface. STRAIGHT-resynthesized speech stimuli have been shown to be highly natural sounding and to suffer little or no decrement in intelligibility due to vocoding process alone (cf. Liu and Kewley-Port 2004; Assmann and Katz 2005).

Relevant to the present study, one modification that is available within the STRAIGHT user interface is linear warping of the time axis to a specified value, while preserving the dynamic profile of both the formants and the fundamental frequency. Thus, temporal modifications were made using the STRAIGHT interface that preserved the extent and direction of formant dynamics (in this case primarily VISC); note, however, that these modifications may have altered the slope of formant trajectories. Thus, duration-neutralized (natural formant) stimuli were created within the interface by specifying one output duration for all input signals (i.e., the average vowel duration of all the original target vowels).

Formant flattened stimuli were created similarly to those of Neel (1998) according to the following steps: (1) identifying the most stable portion of the vowel, defined as the portion of the high-intensity vocalic nucleus in which the change in the ratio of the first two formants was minimized over successive temporal intervals (cf. Hillenbrand and Nearey 1999), (2) excising a small number of pitch periods around the stable region identified in 1, (3) replicating those pitch periods until the target duration was obtained, and (4) presenting these signals as input to STRAIGHT. Finally, duration-neutral flat-formant stimuli were created within STRAIGHT, from the original formant-flattened stimuli using the same duration-modification procedure described in the preceding paragraph.

Listeners were asked to identify the target vowel by indicating the corresponding original (full) syllable, displayed in orthographic transcription, using a six-alternative forced-choice task. All the excised vowel stimuli were presented in separate trial blocks from the full syllables. Figure 3 compares percent-correct performance across the three listener groups for the two natural speech conditions: full syllable and excised vowel. As can be seen in Fig. 3, both the monolinguals and early learners identified the syllables with a similarly high degree of accuracy in the whole syllable condition (about 90–95 % correct). For the early learners, but not the monolinguals, however, there was a modest but statistically significant decrement in performance from the whole syllable to the excised natural vowel

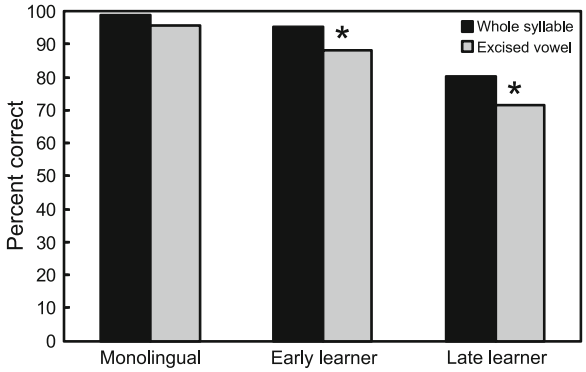


Fig. 3 Percent-correct identification rates for entire /bVd/ syllables and excised vowel nuclei by monolingual English-speaking listeners, relatively early learners of English as a second language and later learners of English as a second language. *Asterisks* indicate statistically significant differences in performance across the two listening conditions, within each listener group

condition (about 7 %). A similar and also statistically significant decrement in performance across the two conditions (about 9 %) was also observed for the later learners. The later learners also showed substantially lower performance than the other two listener groups in both conditions. The data presented in Fig. 3, combined with those obtained by Rogers and Lopez (2008), support the hypothesis outlined above that later, and sometimes early, learners of English as a second language may have greater difficulty in processing spectral cues to vowel identity when syllable integrity has been disrupted, whether in the syllabic nucleus or in the formant transitions.

The effects of duration neutralization and formant flattening are shown in Figs. 4a and b, respectively. Percent-correct syllable identification performance for the STRAIGHT-vocoded excised vowels (no modification) is shown in the left-hand bars in each panel. Performance for the modified stimuli is shown in the right-hand bars in each panel, with performance for the duration-neutral stimuli displayed in the upper panel (a) and that for the formant-flattened stimuli in the lower panel (b). With regard to the effects of duration neutralization, panel (a), only the later learners showed a significant (but small, 4 %) decrement in performance from the full cue to the duration neutralized condition. The slight decrement in performance observed for the other two groups was not significant for this condition. A post hoc analysis of a significant group by vowel identity by listening condition interaction showed that the later learners' decrement in performance with duration neutralization was greatest for the target vowel /ɪ/, suggesting that the later learners may over-rely on duration, rather than VISC, to distinguish this vowel from /i/.

With regard to the effects of formant flattening alone (Fig. 4b), a significant (4–7 %) decrease in performance was observed for the monolinguals and early learners, but not the later learners, from the full cue to the formant flattened

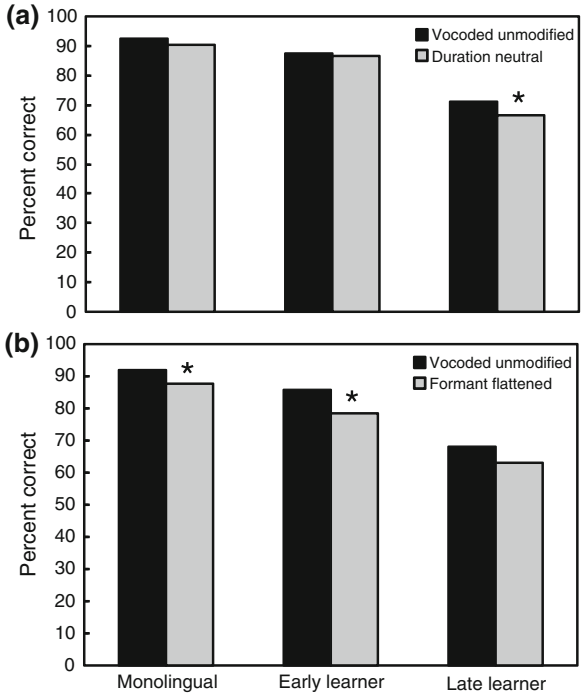


Fig. 4 Percent-correct identification rates for STRAIGHT-vocoded unmodified vowels, compared to duration neutralized vowels (**a**) and to formant flattened stimuli (**b**) for monolingual English-speaking listeners, relatively early learners of English as a second language and later learners of English as a second language. Asterisks indicate statistically significant differences in performance across the two listening conditions, within each listener group. Performance for /eɪ/ is excluded from the comparison of unmodified and formant flattened stimuli (**b**)

condition. Note that /eɪ/ was excluded from this analysis because it dominated the results when included. In contrast to the previous panel, the effect of formant flattening was similar in magnitude but not statistically significant for the later learners. Considering the two panels together, there is some support for the hypothesis outlined above, that VISC can be a relatively difficult cue to acquire for learners of English as a second language. That is, it does appear that the later learners tend to rely more heavily on duration cues and less heavily on VISC for vowel identification, compared to the monolinguals and early learners. The early learners, on the other hand, appear to use both cues similarly to the monolinguals. Together the results of this study support both of the hypotheses outlined above, that non-native listeners may be less robust to syllable degradation in any form, and that they may be less able to make use of the VISC information that is available when the vocalic nucleus is removed.

3 VISC and Vowel Production by Second-Language Learners

All of the participants in the Glasbrenner (2005) were also recorded speaking the same /bVd/ syllables that were presented in the perception task. Each syllable was spoken several times, first in a conversational speaking style and then in a clear speech style, all within a carrier phrase. Productions of two repetitions of each syllable from a subset of the participants (all female talkers) from each group were presented in multi-talker babble to native English-speaking listeners. Listeners were instructed to identify the target syllables in a six-alternative forced-choice task. With regard to intelligibility, the early learners were found to be at least as intelligible as the monolinguals and equally capable of improving intelligibility by speaking more clearly (Rogers et al. 2010). The later learners, on the other hand, were less intelligible overall and showed a smaller improvement in intelligibility when asked to speak more clearly. In fact, analysis of a significant group by speaking style by target syllable interaction showed that later learners' clear speech productions of /bɪd/ syllables were significantly less intelligible than their conversational speech productions.

In summary, even though the early learners had greater difficulty than monolinguals in identifying vowels in disrupted syllables (see Sect. 2.1–2.2), they were equally capable of *producing* those syllables intelligibly, and were equally capable of increasing intelligibility when asked to speak clearly (Rogers et al. 2010). Later learners, however, produced /bVd/ syllables that were less intelligible overall than those produced by the other two talkers, and were less capable of increasing intelligibility when asked to speak clearly.

Acoustic analyses were made for two productions of each of the six /bVd/ syllables in each speaking style, for the majority of the speakers used in Rogers et al. (2010). Measurements of f0 and formant frequencies were made at 20, 50 and 80 % of the vowel duration (Bianchi 2007). Relevant to the hypothesis that VISC aids in distinguishing among neighboring vowels, three broad factors were hypothesized to be relevant to evaluating the degree to which a native or non-native talker may succeed in using formant frequencies and/or changes in formant frequency during the vowel (i.e., VISC) to appropriately differentiate neighboring vowels within the vowel space: (1) location of items in the vowel space and direction of VISC, relative to native-speaker norms; (2) changes over time in the degree of overlap of neighbor vowels across all tokens produced by a group of talkers; and (3) the use of VISC by individual talkers to distinguish a vowel from neighboring vowels, relative to the *position* of that token within the vowel space, as defined by native-speaker norms. The acoustic analyses that follow are roughly organized with regard to each of the three factors outlined above and will be used to compare the productions of the three talker groups described above.

3.1 Comparison of Average Formant Values Across the Talker Groups

First, it is assumed that listeners' expectations are attuned to some region of the vowel space for each distinctive vowel sound in a given language, as shaped by their lifetime's experience with their native language. That is, although inter-speaker variability and a listener's experience with social or regional variation allow for some latitude, it is assumed that a vowel cannot be well recognized if it lies substantially outside the appropriate region of the vowel space. Thus, spectral distinctiveness is necessary but not sufficient for accurate recognition; reasonable agreement with native-speaker norms is also necessary for accurate recognition. Thus, even if features such as VISC or duration allow a talker to reliably distinguish neighboring vowels in production, those vowels are unlikely to be well recognized by native listeners if they are not placed in roughly the appropriate region of the vowel space.

Figure 5 depicts the average values of the productions of each vowel for each talker group. Average Bark-scaled values of F2 are plotted on the abscissa for each group, with average Bark-scaled F1 values on the ordinate. Conversational speech productions are presented in the upper panel (a) and clear speech productions in the lower panel (b). The average values at 20, 50 and 80 % are connected by arrows, with the arrowhead at the later time point in each case.

As pertains to the first factor, the goal of this analysis is to examine the degree to which the average productions of the non-native talker groups occupy similar regions of the vowel space, relative to the native-speaker data. As can be seen from the figure, the means obtained for the monolinguals (MO) and the early learners (EL) tend to parallel one another across the three time points in both the

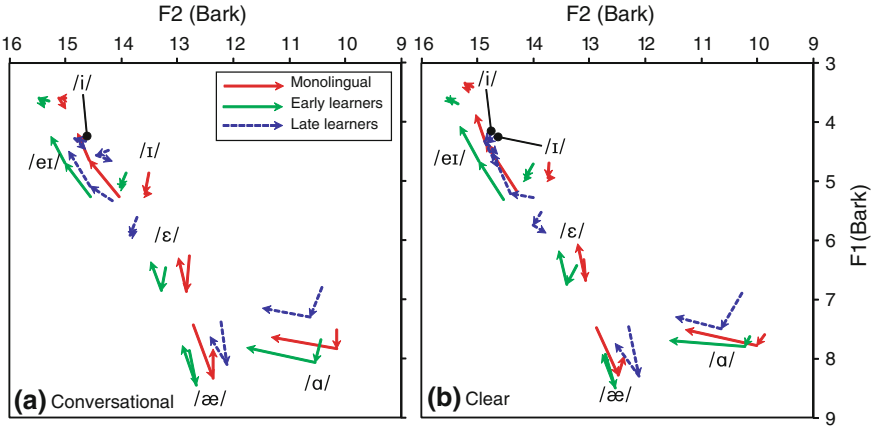


Fig. 5 Average F1 and F2 frequencies (in Barks) at 20, 50 and 80 % of vowel duration for vowels produced in conversational (**a**) and clear speech (**b**) styles by monolinguals, early learners and later learners

conversational and clear speech conditions, rather than matching precisely. Broadly, a parallel between the two vectors implies that similar formant ratios are maintained across the groups and that VISC proceeds in a similar manner across the two groups. Furthermore, the differences in the early learners' productions do not tend to move their productions closer to a region occupied by another vowel, relative to the monolingual talkers. Instead, the early learners' productions tend to be placed in more extreme portions of the vowel space in most cases. That is, the average values for the early learners' front vowels have slightly higher F2 values, implying a more fronted tongue constriction, than those of the monolinguals. Similarly, the early learners' average F1 value for the low back vowel /ɑ/ is somewhat higher than that obtained for the monolinguals, implying a somewhat lower tongue/jaw position. These trends are similar across the two speaking styles for the two talker groups. In all cases, the averages for the EL talkers' productions are within about 0.5 Barks of those for the MO talkers for both speaking styles.

For the later learners (LL), the situation is quite different. The average values for the LL talkers' productions of the vowels /i/, /ɪ/, and /ɛ/ occupy a quite different region of the vowel space than those for the MO talkers, and have a different trajectory of VISC. In both conversational and clear speech, the average F1 values for the LL talkers' /i/ productions are substantially higher than those for the MO talkers (by about 0.7–1.0 Barks), implying a lower tongue/jaw position. The average F2 values for the LL talkers' /ɪ/ productions, on the other hand are substantially higher and the average F1 values somewhat lower than those of the MO talkers (by about 1.0 and 0.5 Barks, respectively), implying a more front tongue constriction and somewhat higher tongue/jaw position, relative to those of the MO talkers. As a result, in both speaking styles, the positions of the LL talkers' average /i/ and /ɪ/ productions are much nearer to one another in the vowel space than for the MO and EL talkers, and both are very near or overlap with the trajectory for /eɪ/, as produced by the MO talkers. Furthermore, the two vowels are even closer in the clear than in the conversational speech style, most likely accounting for the significant decrease in intelligibility for the LL talkers' production of /ɪ/ from conversational to clear speech (Rogers et al. 2010).

Similarly, the average F2 values for the LL talkers' /ɛ/ productions are substantially higher and the average F1 values lower than those of the MO talkers (by about 0.7–1.0 and 0.5–1.0 Barks, respectively), again implying a more front tongue constriction and somewhat higher tongue/jaw position, for the LL talkers, compared to the MO talkers. As a result of the combined differences for the three vowels /i/, /ɪ/, and /ɛ/, the vowel space occupied by the four vowels /i/, /ɪ/, /eɪ/ and /ɛ/ is greatly reduced for the LL talkers, compared to the MO and EL talkers, by a factor of two or more. That is, the Bark-scaled Euclidean distance between the later learners' average /i/ and /ɛ/ productions (about 1.9 Barks in conversational and 1.7 in clear speech) is less than half of that achieved by the other two talker groups (about 3.9 Barks or more in both conversational and clear speech). Finally, the average values for the LL talkers' productions of the two low vowels, /æ/ and /ɑ/, indicate somewhat centralized articulation, compared to the productions of the other two talker groups, as indicated by their lower F2 values for

/æ/ and higher F1 values for /ɑ/). Again, the effect is to bring the two vowels closer to one another within the vowel space than was observed for the other two talker groups.

3.2 *Comparison of Between-Talker Variation and Vowel Overlap, Across Temporal Locations and Talker Groups*

The second factor to be examined is the degree of overlap of neighbor vowels across all tokens produced by a group of talkers. Rather than comparing average values across groups, as in the previous analysis, this analysis examines the role of inter-speaker variability. Some degree of overlap among multiple talkers' productions of neighboring vowels may be unavoidable at any given time point, at least for a language such as English, with a relatively dense vowel space; however, changes in formant frequency over time (i.e., VISC) tend to result in different vowels overlapping at different time points, allowing for better distinctions among all vowels over the whole course of the vowel than is possible at any single time point.

For this analysis, the degree of overlap among neighboring vowels is compared across the three time points measured by Bianchi (2007) for each of the three talker groups. However, this analysis is not meant to assert that these particular time points (20, 50 and 80 % of the vowel duration) are the only or the best time points for such an analysis, merely that they represent values from near the beginning, middle and end of the vocalic portion of the syllables in question (roughly excluding the CV and VC transitions). Data for the MO, EL and LL talkers are presented in Fig. 6, respectively. Data are presented for the conversational speech tokens only because the patterns observed were largely similar across the two speaking styles. In each figure, the values for individual tokens produced by each talker are presented, with ellipses surrounding the full set of tokens for each vowel (with the exception of a few outliers). Bark-scaled F1 and F2 formant values measured at 20, 50 and 80 % of vowel duration are presented in the upper, middle and lower panels, respectively.

For the MO talkers (Fig. 6 left panels), the vowel ellipses appear to rearrange themselves quite substantially across the three time points. That is, both the distance between neighboring vowels and the degree of overlap between any given pair of neighbor-vowel ellipses varies substantially across the three time points. Thus, different sets of neighbor vowels do appear to overlap at the different time points, so that each vowel ellipse *is distinct from those of all nearest neighbor vowels for at least one time point*.

Specifically, at 20 % of vowel duration, the ellipses for /ɪ/, /eɪ/ and /ɛ/ overlap substantially with one another, while the ellipse for /ɛ/ also overlaps substantially with that for /æ/. However, the ellipses for /i/ and /ɑ/ are well separated from all of those for the other four vowels at this time point, by about 0.5 Bark or more. At the

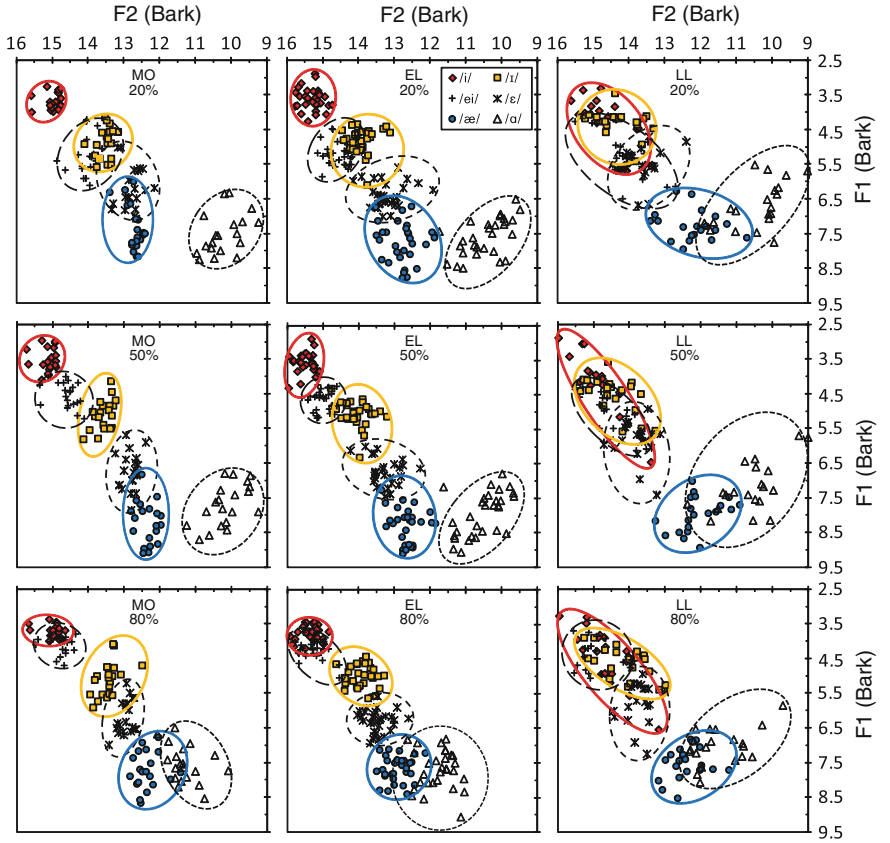


Fig. 6 Individual F1 and F2 frequencies (in Barks) at 20 % (*upper panels*), 50 % (*center panels*) and 80 % (*lower panels*) of vowel duration for individual target vowels spoken in conversational style by monolinguals (MO, *left panels*), early learners (EL, *middle panels*), and late learners (LL, *right panels*). *Ellipses* surround the full set of two tokens per talker for each of the six target vowels (except for one outlier in EL)

50 % time point, the ellipse for /eɪ/ has moved substantially up and left to overlap with those for both /i/ and /ɪ/, but no longer with that for /ɛ/. At the same time, the ellipses for /ɛ/ and /æ/ have both moved substantially lower in the vowel space; thus, the ellipses for /ɛ/ and /æ/ still overlap at 50 % of the vowel duration, but those for /ɛ/ and /ɪ/ no longer overlap with one another. Finally, at 80 % of the vowel duration, the ellipse for /eɪ/ has moved even farther up, so that it no longer overlaps with that for /ɪ/, which has also moved somewhat to the right and down across the time points. Also at this time point, the ellipses for /ɛ/ and /æ/ are no longer overlapping, but the ellipse for /æ/ now overlaps with that for /ɑ/, due to some upward movement of the ellipse for /ɛ/, backward movement of the ellipse for /æ/ and forward movement of the ellipse for /ɑ/.

It is worth noting that such rearranging of the locations of vowels within the vowel space is quite reminiscent of a well-documented type of linguistic change—chain shifts in vowel location over historical time (cf. Labov 1998). Similar to the movement seen in Fig. 6 left panels, as one vowel shifts position due to dialectal change, a neighbor vowel must move to make space for the first shifted vowel, and so on. Imagining these successive movements playing out in time, one can envision a sort of dance among the vowels; each target shifts to avoid another, thus necessitating a shift in the next. Furthermore, under this view, it is also possible that the regular differences in VISC observed across dialects (cf. Jacewicz and Fox 2013 (Chap. 8)) are a consequence of differences across the dialects in patterns or degree of overlap among the "populations" of tokens for specific neighbor vowels, necessitating a different set of rearrangements to distinguish neighbor-vowel tokens for the larger group of talkers from one another, across the duration of the vowel.

The data for the EL talkers are presented in Fig. 6 middle panels, and the patterns observed are similar but not identical to those for the MO talkers. Broadly, the approximate locations and patterns of overlap among the vowel ellipses are quite similar for the MO and EL talkers, especially at the 20 and 80 % time points. In fact, there is a somewhat greater distance across the ellipses in the F1 dimension for the EL talkers than for the MO talkers, particularly at the 20 % time point. On the other hand, the ellipses for /ɑ/ and /æ/ are better distinguished for the MO talkers than for the EL talkers, particularly at the 20 % time point. Furthermore, the ellipse for /ɪ/ overlaps at least slightly with those for both /eɪ/ and /ɛ/ at all three time points for the EL talkers, but not for the MO talkers (note, however, that tokens were analyzed for a larger number of EL than MO talkers).

The data for the LL talkers are presented in Fig. 6 right panels, and the patterns observed are quite different than those for the other two talker groups. Overall, there is much greater overlap among the ellipses for the different vowels at all time points. The ellipses for /i/, /ɪ/, /eɪ/, and /ɛ/ all overlap with one another at all three time points, while those for /ɛ/ and /æ/ also overlap at all three time points. The areas within the ellipses are also substantially larger for the LL talkers than for the other two talker groups, indicating greater between-talker variability.

3.3 Comparison of VISC Relative to Native-Speaker Norms for the Most and Least Intelligible Talkers in Each Talker Group

The third factor to be examined is the use of VISC by individual talkers to distinguish a vowel from neighboring vowels, relative to native-speaker norms. The purpose of this analysis is to explore whether different talkers employ VISC in different ways, depending on the relative position of their productions in the vowel space. That is, the canonical measure of direction and extent of VISC appears to be

a good strategy for achieving separation among closely spaced neighbor vowels for talkers whose productions lie near the mean values for a given population. Other talkers' productions, however, must lie near the edge of the ellipses defined by inter-talker variability (see Fig. 6 for example). For these talkers, the nearest neighbor to a particular production, from among the canonical or "mean" vowels may be different than for a talker whose productions lie near the opposite edge of the ellipse. As a consequence, different trajectories of VISC may be appropriate for different talkers, in order to best distinguish their productions of a given vowel from different nearest neighbor vowels.

For this analysis, the emphasis was not just on whether VISC was different across neighbor vowels, but on the degree to which it is used *by individual talkers* to distance a vowel from its nearest neighbors, as defined by canonical productions of native speakers, estimated from group averages. The focus on distance between neighbor vowels is relevant because while the direction of VISC is quite different for the vowels /æ/ and /ɑ/ for the MO talkers in Fig. 5, the effect is that the two vowels are somewhat closer in the vowel space at 80 % than at 20 % of the vowel duration. Figures 7, 8 and 9 present data for the individual conversational speech productions of the three most intelligible (dashed lines) and the three least intelligible (solid gray lines) talkers in the MO, EL and LL talker groups, respectively (two tokens for each talker for each vowel). The talkers used for these analyses were selected from those used for the acoustic analysis, based on percent-correct scores for their conversational speech tokens in Rogers et al. (2010).

Four panels are presented in each figure, one each for the vowels /ɪ/, /eɪ/, /ɛ/, and /æ/. In each figure, the vowel space is centered on the MO talker group's average value for the vowel in question at 20 % of the vowel duration, with a vector (heavy solid black arrow) extending from that point to the average value for the MO talkers at 80 % of the vowel duration. The vectors for individual talkers' productions are presented as two-point vectors, with formant values at 20 and 80 % of vowel duration scaled relative to the origin (i.e., the average value for the MO talker group at 20 % of the vowel duration). Within each panel, the MO talker group's average vectors are also shown for the nearest neighbor vowels to the vowel in question, with formant values again scaled relative to the origin (solid black arrows); each of these additional vectors is labeled with its phonetic symbol.

Figure 7 presents data for the three most and least intelligible of the MO talkers. For all four vowels, nearly all of the productions of the least intelligible talkers are produced with lower values of F1 (/ɛ/) or both F1 and F2 (/ɪ/, /eɪ/, and /æ/) than those of the most intelligible talkers. Together, these data suggest a somewhat higher and/or more centralized production for the less intelligible talkers and more extreme productions than the average for the most intelligible talkers.

With regard to VISC, some interesting patterns do emerge. For the vowel /ɛ/ in particular (lower left panel), there is some support for the hypothesis that individual talkers may use different directions of VISC to achieve good separation from neighbor vowels, based on the position of their particular productions relative to those of mean or "canonical" vowels within the vowel space. Specifically, while many of the more intelligible talkers' /ɛ/ productions occupy a similar

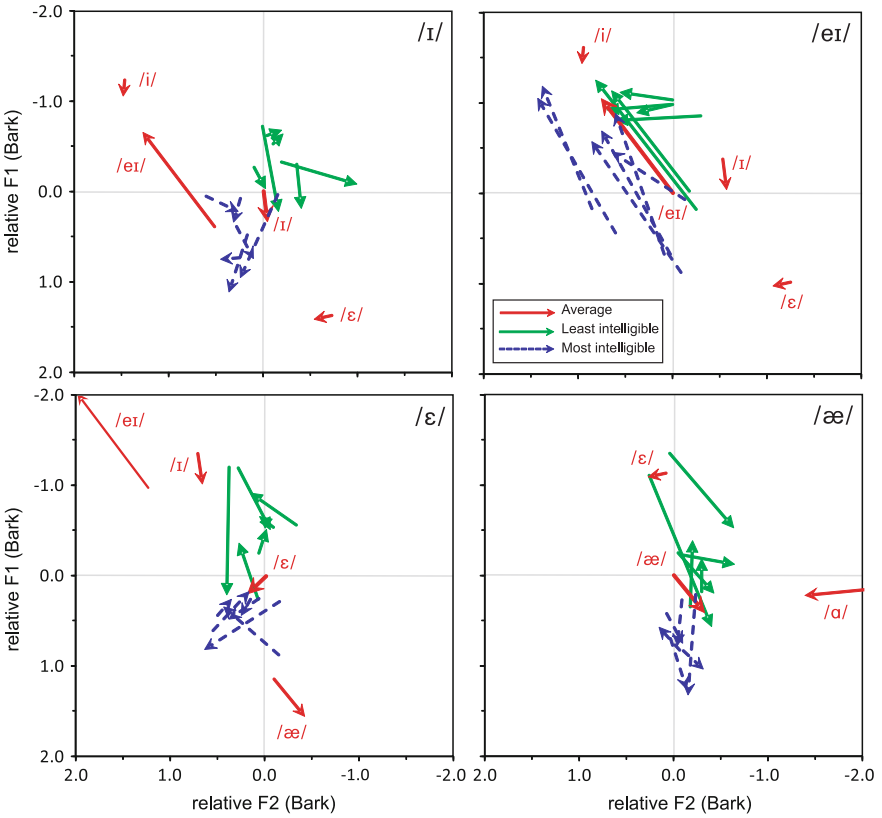


Fig. 7 F1 and F2 frequencies (in Barks) at 20 and 80 % of vowel duration for individual productions of the three most intelligible and three least intelligible talkers in the MO talker group (two tokens per talker per vowel). Productions for the vowels /ɪ/, /eɪ/, /ɛ/ and /æ/ are shown in separate panels. The origin in each panel is the average value for the MO talkers at 20 % of the vowel duration, with a vector extending to the average value at 80 %. Individual talkers' productions are scaled relative to the value used for the origin. The average productions of the MO talkers for near neighbor vowels are also scaled relative to the vowel in question

portion of the vowel space as the average vector for /ɛ/, one talker's production (lowest dashed line) begins much closer to the average vector for /æ/ but decreases substantially in F1 and increases in F2 from 20 to 80 % of the vowel duration, so that it is well separated from the average value for /æ/ at 80 %. The increase in F1 is opposite that of the average vector for /ɛ/, but is what is needed for that talker to distinguish that production of /ɛ/ from the average /æ/. Furthermore, because the vector starts lower in the vowel space, this "inappropriate" VISC does not move the talker's /ɛ/ into the region of any other vowel. There is also some dispersion in direction of the vectors for the rest of the most intelligible talkers but the movement largely keeps their productions well separated from the average vectors for both /ɪ/ and /æ/.

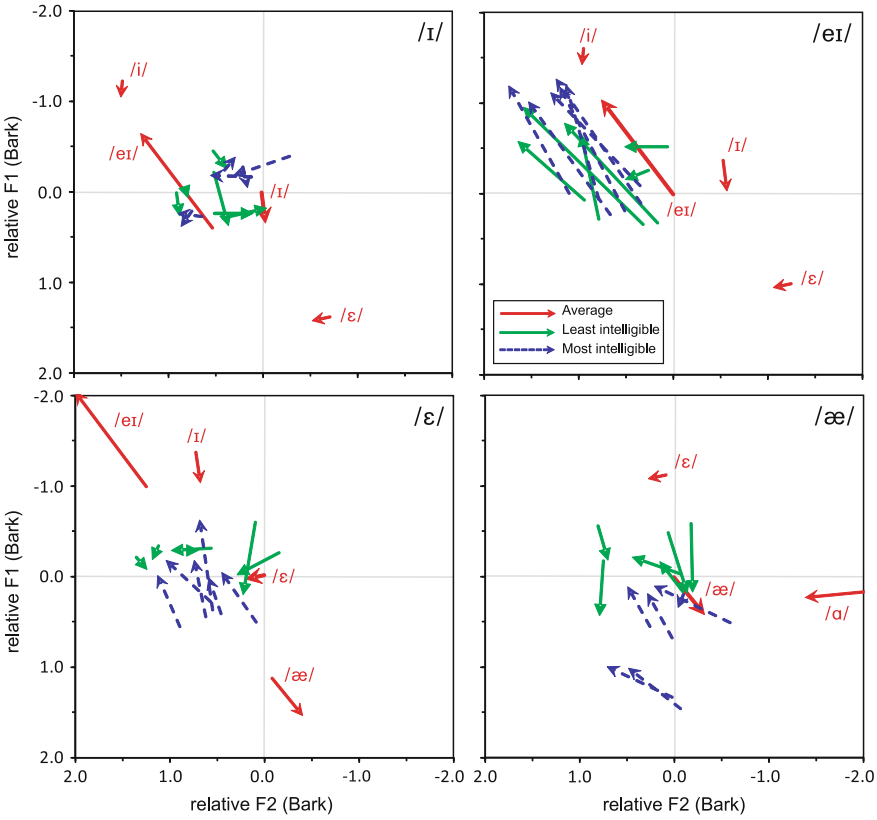


Fig. 8 F1 and F2 frequencies (in Barks) at 20 and 80 % of vowel duration for individual productions of the three most intelligible and three least intelligible talkers in the EL talker group (two tokens per talker per vowel). Productions for the vowels /ɪ/, /eɪ/, /ɛ/ and /æ/ are shown in separate panels. The origin in each panel is the average value for the MO talkers at 20 % of the vowel duration, with a vector extending to the average value at 80 %. Individual talkers' productions are scaled relative to the value used for the origin. The average productions of the MO talkers for near neighbor vowels are also scaled relative to the vowel in question

The less intelligible MO talkers' /ɛ/ productions tend to move toward or lie closer to the average vector for /ɪ/ than those for the more intelligible talkers, and there is again a fair bit of dispersion in direction of the vectors. The direction of VISC for the two talkers whose vectors start closest to /ɪ/ is nearly opposite that for the talker whose production started near /æ/, but is appropriate to maintain separation among the vowels, due to the difference in nearest neighbors. For all three of these 'atypical' talkers, the extent of VISC is larger than average, perhaps because there is more need to move these productions away from a relatively near neighbor.

The situation is similar for /æ/ (bottom right panel), with most of the less intelligible talkers' vectors starting closer to /ɛ/ and moving more towards /ɑ/ than those

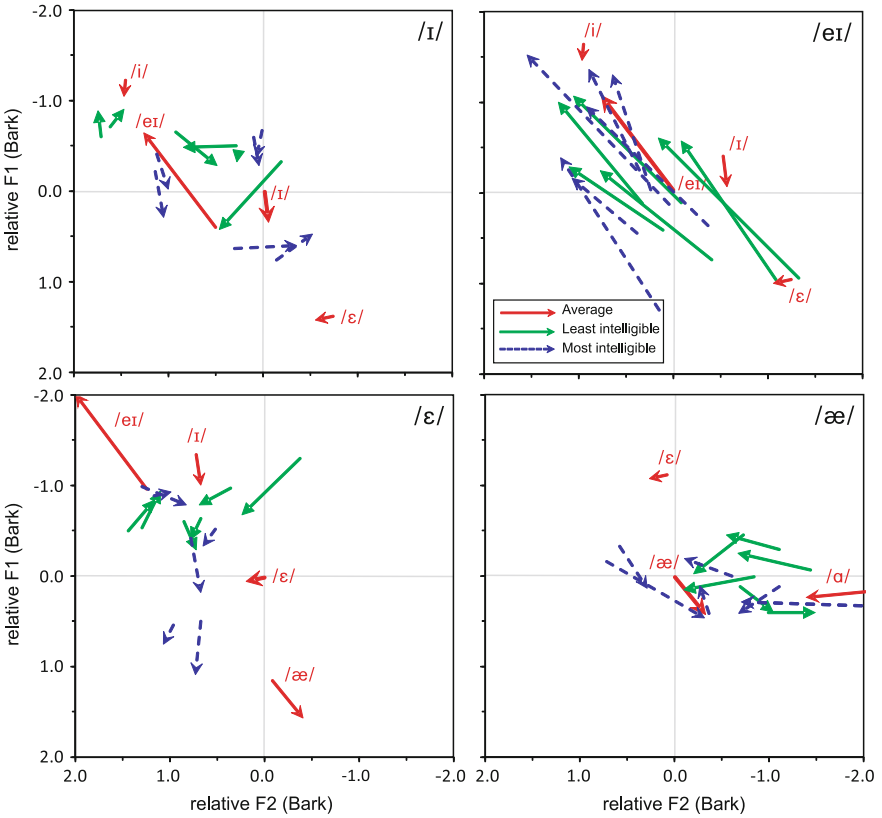


Fig. 9 F1 and F2 frequencies (in Barks) at 20 and 80 % of vowel duration for individual productions of the three most intelligible and three least intelligible (*solid gray vectors*) talkers in the LL talker group (two tokens per talker per vowel). Productions for the vowels /ɪ/, /eɪ/, /ɛ/ and /æ/ are shown in separate panels. The origin in each panel is the average value for the MO talkers at 20 % of the vowel duration, with a vector extending to the average value at 80 %. Individual talkers' productions are scaled relative to the value used for the origin. The average productions of the MO talkers for near neighbor vowels are also scaled relative to the vowel in question

for the more intelligible talkers. Again, some of the more intelligible talkers' vectors are quite different in direction than that for the average vector for /æ/, but in most cases this movement maintains better separation from the average vector for /ɑ/ than would be observed if the average vector direction and extent had been used.

For the vowel /ɪ/ (upper left panel), VISC for at least two of the most intelligible talkers (bottom left dashed lines) moves their productions into a region of the vowel space which had been quite close to the average value for /eɪ/ at 20 % but which has been vacated at 80 %; this movement distances these talkers' /ɪ/ productions from the average value for /eɪ/ over time, without moving them closer to the average vector for /ɛ/. This movement is also somewhat different from the direction of the *average vector* for /ɪ/. Thus, direction of VISC for these talkers is

different from the average VISC, but is at least as successful at maintaining separation from near neighbors in the vowel space. The less intelligible talkers' productions of /ɪ/ tend to result in a somewhat more centralized production, moving away from the average vectors for /i/ and /eɪ/ but somewhat closer to /ɛ/, even though they agree somewhat better in direction with the mean or "canonical" vector. For /eɪ/, the vectors for the most intelligible talkers and three of those for the less intelligible talkers largely parallel the average vector for /eɪ/, although with somewhat higher values of F2 for the more intelligible talkers. Three of the vectors for the less intelligible talkers' productions are less well distinguished from the average vectors for /i/ and /ɪ/, however, due much lower values of F1 at 20 %, and an increase only in F2 from 20 to 80 % of vowel duration.

While average values for direction and extent of VISC are, by definition, the best representation of a talker group's strategies for maintaining separation among neighbor vowels, the above analyses suggest that individual talkers may successfully implement VISC using strategies that differ from mean or "canonical" values but nevertheless achieve the goal of maintaining separation among neighbor vowels. For this reason, models that fit talkers' productions to average values of VISC may be limited in their predictive power because they do not take into account that alternative strategies of using VISC to maintain vowel separation may be equally successful, particularly for talkers whose average formant ratios do not lie near the group mean. Perhaps a model that includes such information might better predict intelligibility based on measures of VISC for individual talkers.

Figure 8 presents the data for the EL talkers, with quite a few differences observable from the typical patterns for the MO talkers. Given these differences, it is worth noting that these EL talkers were found to be, on average, at least as intelligible as the MO talkers (cf. Rogers et al. 2010). For /ɛ/ (lower left panel), many of the vectors for the more intelligible talkers' productions begin with a higher F1 and higher F2 than the average value for the MO talkers, placing many of the EL talkers' productions somewhat close to the average MO-talker vector for /æ/. Similar to one of the more intelligible MO talkers, the effect of VISC for these talkers is to decrease F1 and somewhat increase F2. Thus VISC for these talkers tends to be greater and in a different direction than the average MO-group vector for /ɛ/, but serves to move these talkers' productions away from the nearest neighbor vowel (/æ/). The less intelligible EL talkers' vectors for /ɛ/ tend to have a smaller degree of VISC than those of the more intelligible EL talkers and also have a lower F1 and higher F2 than that of the average vector for /ɛ/ for the MO talkers. As a result, some of the less intelligible EL talkers' /ɛ/ productions lie closer to the average MO-group vectors for both /eɪ/ and /ɪ/ than do the vectors for most of the more intelligible EL talkers.

For /ɪ/ (upper left panel), the vectors for the most intelligible EL talkers show a much smaller extent of VISC than those for the more intelligible MO talkers and in no case do they move to occupy the space vacated by /eɪ/, as did those for some of the MO talkers. The productions of the EL talkers also tend to lie closer to the

average MO-talker vector for /eɪ/ than do those of the MO talkers. Both of these patterns *may* be attributable to an avoidance of Spanish /e/, which has been measured to have a substantially lower F2 but similar F1 to American English /eɪ/, although for a different dialect (Bradlow 1995). The pattern for the EL talkers' /eɪ/ productions (upper right panel) is quite similar to that of the MO talkers, except that even the less intelligible talkers' productions tend to be produced with higher F2 values than the average MO-group vector for /eɪ/.

For /æ/, the vectors for the less intelligible EL talkers' productions start closer to the average MO-group vector for /ɛ/ than do those for the more intelligible talkers, but move to become closer to the average MO-group vector for /ɑ/, similar to the pattern for some of the less intelligible MO talkers' productions. Most of the vectors for the more intelligible EL talkers' productions for /æ/ begin with higher F1 values than average MO-group vector for /æ/ and increase in F2 and decrease in F1 across the vowel duration, maintaining separation from the average MO-group vectors for both /æ/ and /ɑ/. As a consequence, the direction of VISC for /æ/ for the more intelligible EL talkers tends to be quite different from that for the average MO-group vector for /æ/, but allows for good separation among /æ/ and /ɑ/. Note that the observed direction of VISC would *not* be effective for maintaining vowel separation if these talkers' vowels had begun at a lower value of F1. Thus, these patterns also provide some evidence that the location of a talker's production within the vowel space occupied by a given vowel is used by the talker to determine the direction of VISC that is most appropriate for maintaining separation from neighbor vowels.

Figure 9 presents data for the LL talkers. Overall, both the most and least intelligible LL talkers' productions are more dispersed than for the other two groups. For /ɪ/, most of the talkers' productions have a lower F1 and/or a higher F2 than the average MO-group vector for /ɪ/. As a result, most of the vectors for these LL talkers' productions approach the average MO-group vector for either /i/ or /eɪ/.

For /eɪ/, two of the less intelligible LL talkers' vectors begin in the region of the average MO-group vector for /ɛ/, but those for rest of the LL talkers roughly parallel the average MO-group vector for /eɪ/. For /ɛ/, all the vectors for the less intelligible LL talkers' productions start and end with substantially lower values for F1 and, in some cases, higher values for F2 than for the average MO-talker vector for /ɛ/, so that they approach the average MO-group vector for /ɪ/ or /eɪ/. The vectors for three productions of /ɛ/ by the most intelligible LL talkers do lie at higher F1 values than those of the other LL talkers, placing them closer to the average MO-group vector for this vowel.

For /æ/, the vectors for most of least intelligible LL talkers' productions lie quite close to the MO group average value at 80 % for /ɑ/. For most of these, the direction of VISC is to increase F2, which helps to distinguish them from /ɑ/ but is opposite of the direction of the average vector for /æ/ for the MO talkers. Two of the productions of the most intelligible LL talkers, on the other hand, are well distinguished from the average vector for /ɑ/ and have a direction of VISC that is similar to that for the MO talkers' average vector for /æ/.

For /ɪ/ and /ɛ/ especially, patterns of VISC did not appear to help to separate the LL talkers' productions from the nearest MO-group average neighbor vowels very effectively. These talkers, of course, are most likely still in the process of acquiring these sounds. To some degree then, the data for the LL talkers support the hypothesis that *appropriate* use of VISC is a relatively difficult feature of American English vowel production to acquire for learners with a first language of Spanish. Performance for /eɪ/, however, suggests that this sound may be transferred to English from the corresponding Spanish diphthong because the direction and extent of formant movement are reasonably similar to that of the monolinguals, even for the later learners.

4 Conclusion

One goal of the research described in this chapter is to examine *variation* in the use of acoustic cues in both speech perception and production among second-language learners with different ages of onset of immersion in order to better understand the ways in which the human speech perception and production systems achieve the remarkable robustness and apparent flexibility demonstrated during everyday language use. Indeed, the comparisons across monolingual and second-language learner groups, both in perception and production, gave us many insights into both similarities and differences across the monolingual, early learner and later learner subject groups we studied. In perception, differences in the effects of modification of duration and VISC cues to vowel identity support the hypothesis outlined in the Introduction that VISC might be a relatively difficult feature for non-native listeners to acquire (Glasbrenner 2005). However, those results, combined with those of Rogers and Lopez (2008) suggest that both early and later learners of English as a second language with Spanish L1 may have greater difficulty than native listeners in adapting to disruptions in syllable integrity. These results may in part account for the greater difficulty that non-native listeners have been shown to experience when processing speech in noisy environments.

In production, analyses exploring multi-talker variability explored the hypotheses that (1) VISC enhances separation among neighbor vowels by allowing for overlap for different subsets of neighbor vowels at different points in time and (2) different talkers may use different strategies in implementing VISC to achieve better separation among neighbor vowels, depending on the position of their own productions, relative to native-speaker norms. These analyses showed that, on average, the early learners of English produced vowels that were similar to those of the native speakers in both location within the formant space and direction and extent of VISC. The average values for the later learners showed a contracted distance within the vowel space between the vowels /i/ and /ɛ/, compared to the other two talker groups. The analyses of between-talker variability examined overlap among the ellipses defined by the individual tokens of all the talkers across the three time points measured, for each of the three talker groups. While there was

some overlap among neighbor-vowel ellipses at each time point, the analyses for the native speakers and early learners showed separation between ellipses of all neighbor vowels for at least one time point. In the analyses for the later learners, however, some neighbor-vowel ellipses were never well distinguished from one another. Instead, only two subsets of vowels were found to be distinguished from one another, with considerable overlap among the ellipses within each set.

Finally, analyses of VISC were performed for individual tokens produced by the talkers within each group with the three highest and three lowest intelligibility scores in Rogers et al. (2010). For each vowel, tokens were analyzed relative to the average values for the vowels produced by the native speakers. For the native speaker and early learner groups, these analyses showed some support for the hypothesis that talkers may use differing strategies to achieve good separation among target vowels, based on their position in the vowel space, relative to native-speaker norms. Together, these results suggest that statistical analyses of VISC that take into account the degree to which an individual talkers' productions maximize spectral distance from neighbor vowels over the course of the vowel, both within a talker and relative to native-speaker norms, may add to our understanding of how VISC is used to enhance intelligibility.

Acknowledgments Research supported by NIH-NIDCD Grant #5R03 DC005561-01. Thanks to Peter F. Assmann, Geoffrey Stewart Morrison, Winifred Strange, and Keith Kluender for many helpful discussions, suggestions and encouragement.

References

Assmann, P.F., Katz, W.F.: Time-varying spectral change in the vowels of children and adults. J. Acoust. Soc. Am. **108**, 1856–1866 (2000). doi:10.1121/1.1289363

Assmann, P.F., Katz, W.F.: Synthesis fidelity and time-varying spectral change in vowels. J. Acoust. Soc. Am. **117**, 886–895 (2005). doi:10.1121/1.1852549

Assmann, P.F., Nearey, T.M., Hogan, J.T.: Vowel identification: orthographic, perceptual, and acoustic aspects. J. Acoust. Soc. Am. **71**, 975–989 (1982). doi:10.1121/1.387579

Bianchi, M.: Effects of clear speech and linguistic experience on acoustic characteristics of vowel production, Unpublished masters thesis, University of South Florida, Tampa (2007)

Bohn, O.-S., Flege, J.E.: Interlingual identification and the role of foreign language experience in L2 vowel perception. Appl. Psycholinguistics **11**, 303–328 (1990). doi:10.1017/S0142716400008912

Bradlow, A.R.: A comparative acoustic study of English and Spanish vowels. J. Acoust. Soc. Am. **97**, 1916–1924 (1995). doi:10.1121/1.412064

Edwards, J., Fox, R.A., Rogers, C.L.: Final consonant discrimination in children: effects of phonological disorder, vocabulary size, and articulatory accuracy. J. Speech Lang. Hear. Res. **45**, 231–242 (2002). doi:10.1044/1092-4388(2002/018

Glasbrenner, M.M.: Vowel identification by monolingual and bilingual listeners: use of spectral change and duration cues. Unpublished masters thesis, University of South Florida, Tampa (2005)

Hazan, V., Barrett, S.: The development of phonemic categorization in children aged 6–12. J. Phonetics **28**, 377–396 (2000). doi:10.1006/jpho.2000.0121

Hillenbrand, J.M.: Static and dynamic approaches to understanding vowel perception. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Ch. 2). Springer, Heidelberg (2013)
Hillenbrand, J.M., Clark, M.J., Nearey, T.M.: Effect of consonant environment on vowel formant patterns. J. Acoust. Soc. Am. **109**, 748–763 (2001). doi:10.1121/1.1337959
Hillenbrand, J.M., Getty, L.A., Clark, M.J., Wheeler, K.: Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. **97**, 3099–3111 (1995). doi:10.1121/1.411872
Hillenbrand, J.M., Nearey, T.M.: Identification of resynthesized/hVd/syllables: effects of formant contour. J. Acoust. Soc. Am. **105**, 3509–3523 (1999). doi:10.1121/1.424676
Jacewicz, E., Fox, R.A.: Cross-dialectal differences in dynamic formant patterns in American English vowels. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Ch. 8). Springer, Heidelberg (2013)
Jenkins, J.J., Strange, W., Edman, T.R.: Identification of vowels in 'vowel less' syllables. Perception and Psychophysics **34**, 441–450 (1983). doi:10.3758/BF03203059
Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous frequency-based F0 extraction. Speech Commun. **27**, 187–207 (1999). doi:10.1016/S0167-6393(98)00085-5
Kluender, K.R., Stilp, C.E., Kiefte, M.: Perception of vowel sounds within a biologically realistic model of efficient coding. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 6). Springer, Heidelberg (2013)
Labov, W.: The three dialects of English. In: Linn, M.D. (ed.) Handbook of Dialects and Language Variation, pp. 39–81. Academic, San Diego (1998)
Liu, C., Kewley-Port, D.: Vowel formant discrimination for high-fidelity speech. J. Acoust. Soc. Am. **116**, 1224–1233 (2004). doi:10.1121/1.1768958
Lowenstein, J.H., Nittrouer, S.: Learning to perceptually organize speech signals in native fashion. J. Acoust. Soc. Am. **127**, 1624–1635 (2010). doi:10.1121/1.3298435
Morrison, G.S.: Perception of synthetic vowels by monolingual Canadian-English, Mexican-Spanish, and Peninsular-Spanish listeners. Can Acoust **36**(4), 17–23 (2008). (Typesetting errata published in 37(1), 34)
Morrison, G.S.: L1 and L2 production and perception of English and Spanish vowels: a statistical modelling approach. University of Alberta, PhD dissertation (2006)
Morrison, G.S.: L1-Spanish speakers' acquisition of the English/i/–/I/contrast II: perception of vowel inherent spectral change. Lang. Speech **52**, 437–462 (2009). doi:10.1177/0023830909336583
Morrison, G.S.: Theories of vowel inherent spectral change: a review. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 3). Springe, Heidelberg (2013)
Nearey, T.M.: Vowel inherent spectral change in the vowels of North American English. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 4). Springer, Heidelberg (2013)
Nearey, T.M., Assmann, P.: Modeling the role of inherent spectral change in vowel identification. J. Acoust. Soc. Am. **80**, 1297–1308 (1986). doi:10.1121/1.394433
Neel, A.T.: Factors influencing vowel identification in elderly hearing-impaired listeners. Ph.D. dissertation, Indiana University. From Dissertations & Theses: Full Text (Publication No. AAT 3005387), Retrieved 12 May 2011
Neel, A.T.: Vowel space characteristics and vowel identification accuracy. J. Speech, Language, Hear. Res. **51**, 574–585 (2008). doi:10.1044/1092-4388(2008/041
Nittrouer, S.: The relation between speech perception and phonemic awareness: evidence From low-SES children and children with chronic OM. J. Speech Hear. Res. **39**, 1059–1070 (1996)
Peterson, G., Barney, H.: Control methods used in a study of vowels. J. Acoust. Soc. Am. **24**, 175–184 (1952). doi:10.1121/1.1906875
Pickett, J.M.: The acoustics of speech communication: Fundamentals, speech perception theory, and technology. Allyn and Bacon, Boston (1999)

Rogers, C.L., DeMasi, T.M., Krause, J.C.: Conversational and clear speech intelligibility of /bVd/ syllables produced by native and non-native English speakers. J. Acoust. Soc. Am. **128**, 410–423 (2010). doi:10.1121/1.3436523

Rogers, C.L., Lopez, A.S.: Perception of silent-center syllables by native and nonnative English speakers. J. Acoust. Soc. Am. **124**, 1278–1293 (2008). doi:10.1121/1.2939127

Strange, W.: Dynamic specification of co articulated vowels spoken in sentence context. J. Acoust. Soc. Am. **85**, 2135–2153 (1989). doi:10.1121/1.397863

Strange, W., Jenkins, J.J.: Dynamic specification of co articulated vowels: Research chronology, theory, and hypotheses. In: Morrison, G.S., Assmann, P.F. (eds.) Vowel Inherent Spectral Change (Chap. 4). Springer, Heidelberg (2013)

Strange, W., Jenkins, J.J., Johnson, T.L.: Dynamic specification of co articulated vowels. J. Acoust. Soc. Am. **74**, 695–705 (1983). doi:10.1121/1.389855

Thomson, R.I., Nearey, T.M., Derwing, T.M.: A modified statistical pattern recognition approach to measuring the cross linguistic similarity of Mandarin and English vowels. J. Acoust. Soc. Am. **126**, 1447–1460 (2009). doi:10.1121/1.3177260

Part IV
VISC Applied

Vowel Inherent Spectral Change in Forensic Voice Comparison

Geoffrey Stewart Morrison

Abstract The onset + offset model of vowel inherent spectral change has been found to be effective for vowel-phoneme identification, and not to be outperformed by more sophisticated parametric-curve models. This suggests that if only simple cues such as initial and final formant values are necessary for signaling phoneme identity, then speakers may have considerable freedom in the exact path taken between the initial and final formant values. If the constraints on formant trajectories are relatively lax with respect to vowel-phoneme identity, then with respect to speaker identity there may be considerable information contained in the details of formant trajectories. Differences in physiology and idiosyncrasies in the use of motor commands may mean that different individuals produce different formant trajectories between the beginning and end of the same vowel phoneme. If within-speaker variability is substantially smaller than between-speaker variability then formant trajectories may be effective features for forensic voice comparison. This chapter reviews a number of forensic-voice-comparison studies which have used different procedures to extract information from formant trajectories. It concludes that information extracted from formant trajectories can lead to a high degree of validity in forensic voice comparison (at least under controlled conditions), and that a whole trajectory approach based on parametric curves outperforms an onset + offset model.

Abbreviations

C_{llr}	Log-likelihood-ratio cost
DCT	Discrete cosine transform

G. S. Morrison (✉)
Forensic Voice Comparison Laboratory, School of Electrical Engineering & Telecommunications, University of New South Wales, Sydney, Australia
e-mail: geoff-morrison@forensic-voice-comparison.net

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3_11,

DNA	Deoxyribonucleic acid
DTW	Dynamic time warping
F1	First formant
F2	Second formant
F3	Third formant
LPC	Linear predictive coding
LR	Likelihood ratio
MFCC	Mel-frequency cepstral coefficient
MVKD	Multivariate kernel density
VISC	Vowel inherent spectral change

1 Introduction

This chapter provides a review of the literature on the extraction of information about vowel inherent spectral change (VISC) for the purpose of forensic voice comparison. The chapter begins by providing some basic information about forensic science. It then provides a theory as to why information extracted from VISC might be effective for forensic voice comparison. It then reviews procedures which have been used to extract information from VISC for forensic voice comparison. Finally it summarizes recent studies which have demonstrated that information extracted from VISC can lead to a high degree of validity in forensic voice comparison (at least under controlled conditions).

2 Forensic Science

2.1 The New Paradigm for the Evaluation of Forensic Evidence

We are in the midst of what Saks and Koehler (2005) have called a *paradigm shift* in the evaluation and presentation of forensic evidence, although in Kuhnian terms (Kuhn 1962) it may be better described as a shift from a pre-paradigm stage of scientific investigation towards the establishment of the first generally accepted paradigm. The new paradigm applies to the branches of forensic science which deal with the comparison of the quantifiable properties of samples of known and questioned origin (the sample of known origin is typically associated with a suspect and the sample of questioned origin is typically associated with the offender). Examples include the comparison of DNA profiles, finger marks, glass fragments, rifling on bullets, handwriting, and voice recordings. The new paradigm requires that the validity and reliability of the evaluation of forensic evidence be demonstrated, and that the results of a forensic evaluation be presented in a

logically correct manner. A framework which allows researchers and practitioners to meet the latter requirement, and which is recommended by forensic statisticians (e.g., Aitken and Taroni 2004; Aitken et al. 2010; Balding 2005; Berger et al. 2011; Buckleton 2005; Evett 1998, 2009; Evett et al. 2011; Lucy 2005; Morrison 2012b; Robertson and Vignaux 1995) and which has been adopted as standard for DNA-profile comparison (Foreman et al. 2003), is the *likelihood-ratio framework*. A brief history of the (as-yet far from universal) adoption of the likelihood-ratio framework for forensic voice comparison is provided in Morrison (2009a). The present chapter will not reiterate arguments in favor of the adoption of the likelihood-ratio framework for forensic voice comparison (the interested reader may consult Champod and Meuwly 2000; González-Rodríguez et al. 2006; González-Rodríguez et al. 2007; Jessen 2008; Morrison 2009a, b; Rose 2002, 2003, 2005, 2006; Rose and Morrison 2009). For an extended introduction to the evaluation of forensic-voice-comparison evidence within the new paradigm, see Morrison (2010, 2011b) and Morrison et al. (2012).

Although some important ideas on the forensic use of VISC predate the new paradigm, their discussion in the present chapter will be simplified by working from the perspective of the new paradigm. The chapter's primary focus will also be on the paradigm-neutral techniques employed for the extraction of information from the acoustic signal.

2.2 The Forensic Likelihood-Ratio

A forensic scientist is provided with two (or more) samples, one of known origin (e.g., an audio recording of a police interview with a suspect), and the other of questioned origin (e.g., an audio recording of a telephone call made by the offender). In the likelihood-ratio framework the task of the forensic scientist is to present the court with a *strength-of-evidence* statement in answer to the question:

> How much more likely are the observed properties of the known and questioned samples under the hypothesis that the questioned sample has the same origin as the known sample than under the hypothesis that it has a different origin?

The answer to this question is quantitatively expressed as a *likelihood ratio*, calculated using Eq. 1.

$$\text{likelihood ratio} = \frac{p(\text{observed sample properties} \mid \text{same origin hypothesis})}{p(\text{observed sample properties} \mid \text{different origin hypothesis})} \tag{1}$$

The likelihood ratio (LR) is a numeric expression of the strength of the evidence with respect to the competing hypotheses. If the forensic scientist testifies that one would be 100 times more likely to observe the properties of the known and questioned samples under the same-origin hypothesis than under the different-origin hypothesis (LR = 100), then whatever the trier of fact's belief prior to

hearing this, they should now be 100 times more likely than before to believe that the samples have the same origin (the *trier of fact* is the judge, panel of judges, or the jury depending on the legal system). Likewise, if the forensic scientist testifies that one would be 1,000 times more likely to observe the evidence under the different-origin hypothesis than under the same-origin hypothesis (LR = 1/1000), then whatever the trier of fact's prior belief, they should now be 1,000 times more likely than before to believe that the samples have different origins. Likelihood-ratios which are further from 1 represent greater strengths of evidence and will therefore be of greater assistance to the trier of fact in their task of determining whether the recordings were produced by the same speaker or different speakers.

3 Will VISC Provide Good Acoustic Features for Forensic Voice Comparison?

An important research question in forensic voice comparison is which acoustic features have the potential to lead to the most valid and reliable strength-of-evidence results. An ideal forensic-voice-comparison system would produce a high strength of evidence in favor of the same-speaker hypothesis (LR $\gg$ 1) when the inputs are same-speaker samples, and a high strength of evidence in favor of the different-speaker hypothesis (LR $\ll$ 1) when the inputs are different-speaker samples. Upon contemplation of Eq. 1, it should be apparent that acoustic features of voices which have the potential for producing appropriate LR $\gg$ 1 and LR $\ll$ 1 will be those with relatively low within-speaker variability and relatively high between-speaker variability. If the difference between the suspect and offender samples is relatively small, and the likelihood of obtaining such a small difference assuming same origin is high (because within-speaker variability is small) and the likelihood of obtaining such a small difference assuming different origin is low (because between-speaker variability is large), then this will result in LR $\gg$ 1. Ceteris paribus, if the difference between the suspect and offender samples is relatively large then this will result in LR $\ll$ 1. If, however, the likelihood of obtaining the difference between the suspect and offender samples is approximately equal whether assuming same or different origin (because within-speaker and between-speaker variability are similar), then this will result in LR $\approx$ 1, the strength of evidence will not be high in favor of either hypothesis. Given this, will information about VISC provide good features for forensic voice comparison?

As outlined in Morrison (2013a, Chap. 3), simple parameterizations of VISC such as the onset + offset or onset + slope models (each using two parameters per formant) have been found to be highly correlated with human listeners' vowel-phoneme identification (Gottfried et al. 1993; Nearey and Assmann 1986), and in pattern-recognition experiments more sophisticated parametric-curve models do not appear to outperform simple two-parameter models (Hillenbrand et al. 2001). If only simple cues such as initial and final formant values are necessary for signaling phoneme identity, then speakers may have considerable freedom in the

exact trajectories produced between initial and final values, and an individual speaker may produce the formant trajectories which best suit the idiosyncrasies of the physiology of their vocal tract and their learning of motor commands. Vowel formant trajectories could, therefore, have relatively high between-speaker variability and relatively low within-speaker variability and thus potentially lead to high strengths of evidence. This theoretical argument was previously made in Morrison (2008, 2009c). McDougall (2004, 2006) made a similar theoretical argument, but in terms of articulation rather than perception, on the premise that what is important for phoneme production is the attainment of articulatory targets.

In addition, features extracted from VISC may fulfill other desiderata for acoustic features in forensic voice comparison: (a) Voice samples provided for forensic analysis often consist of relatively short audio recordings, and it may be that realizations of only the most common phonemes occur in sufficient number to allow for reliable statistical analysis. Some diphthongs occur relatively frequently in speech and may therefore make good candidates for forensic comparison, e.g., /aɪ/ occurs frequently in English because it occurs in common words such as "hi" and "bye". (b) Questioned voice samples often consist of audio recordings of telephone conversations, and acoustic features which are robust to degradation due to telephone transmission are therefore preferable. First formant (F1) trajectories are prone to being compromised by the lower cutoff of the bandpass of a landline telephone system, and third formant (F3) trajectories are prone to being compromised by the codecs employed by mobile telephone systems (Guillemin and Watson 2008), but second formant (F2) trajectories may be relatively robust to both. One can therefore expect to have at least F2 from each vowel available for forensic comparison. (c) Acoustic features which are relatively easy to extract are more practical for forensic voice comparison. Although error-free formant measurement remains elusive and human supervision of semi-automatic algorithms may be necessary, formant measurements of vowels are relatively easy to extract from the acoustic signal, standard algorithms such as linear predictive coding (LPC) being widely employed. Cepstral coefficients are easier to extract, but are more susceptible to transmission-channel effects.

4 Procedures for the Extraction of Information from VISC

This section consists of a review of procedures which have been used to extract information from VISC for the purpose of conducting forensic voice comparison.

4.1 Ad hoc Measurements

The first published article to focus on extracting information from VISC for the purpose of forensic voice comparison appears to have been Goldstein (1976). The

quantification of the properties of VISC was rather ad hoc, the procedure involved two steps: first track and plot the formant trajectories of a number of vowels and vowel plus /r/ sequences, then visually inspect the plots and identify landmarks where acoustic features could be measured. Goldstein extracted 199 different features (the same measurement on tokens of different phonemes, e.g., maximum F1 of /e/ and maximum F1 of /ar/ , were counted as different features). A few examples of these features are:

- maximum F1 in /e/
- minimum F2 in /ar/
- F2 at the beginning of /e/
- F3 at the midpoint of /ɝ/
- F2 at 20 ms before the end of /aɪ/
- F1 of /ɔɪ/ at its maximum F2 point
- mean F4 of /aʊ/ excluding the last 20 ms
- mean F3 minus F2 of /rɛ/ excluding the last 20 ms

Features with the greatest potential for forensic voice comparison were identified via a direct calculation of their ratio of within-speaker to between-speaker variance (*F* ratio). By looking at formant measures from different parts of the vowel, rather than a single static vowel target, Goldstein pioneered forensic application of VISC.

4.2 Multi-Point Comparisons

For two decades after the publication of Goldstein (1976) there seems to have been little interest in using VISC features for forensic voice comparison. The next published papers on the topic appeared in the mid 1990s, and applied variations on a technique which I will label *multi-point comparison*.

Ingram et al. (1996) analyzed the formant trajectories in sonorant stretches of speech using LPC. The formant measurements at each analysis frame were used in the calculation of a dissimilarity metric (the spacing of the frames/window step size was not mentioned in the paper). Pairs of formant tracks were aligned at the first frame and the mean of the *signed differences* between the two tracks at each frame from the first frame to the last frame of the shorter track was calculated, see Fig. 1. No time normalization was applied, and the excess frames of the longer track played no part in the calculation of the dissimilarity metric.

A theoretical example will demonstrate that the Ingram et al. (1996) metric conflates different types of differences: If two formant tracks are identical, then they will have a dissimilarity value of zero, but a symmetrical pair of formant tracks, one rising over time and the other falling with the crossover point mid way between the first and last frame, see Fig. 2, will also have a dissimilarity value of zero.

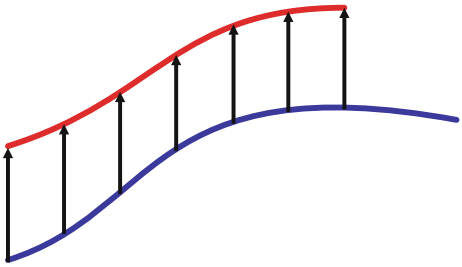


Fig. 1 Multi-point comparison of Ingram et al. (1996)

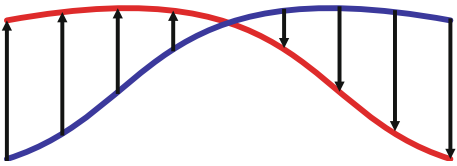


Fig. 2 Two formant tracks with an Ingram et al. (1996) dissimilarity metric of zero

Greisbach et al. (1995) also used multi-point comparison of formant tracks from a number of vowels, but first *linearly time normalized* the tracks then measured the *unsigned differences* between pairs of formants at five equally-spaced points, see Fig. 3. The Euclidean distance in the F1–F2 space between a pair of vowels was then calculated at each of the five time points and these values summed to obtain a dissimilarity metric.

A theoretical example will demonstrate that the Greisbach et al. (1995) metric conflates different types of differences: Imagine a pair of vowels with identical F2 tracks and a parallel pair of F1 tracks separated by 10 Hz at each of the five time points (Fig. 4a), the dissimilarity value will be 50. Now distort one of the F1 tracks so that it crosses the other and has relative separations of −10, −5, +5, +10, +20 Hz at the five time points respectively (Fig. 4b), the dissimilarity value will also be 50.

McDougall (2004) also applied multi-point comparison to linearly time-normalized formant tracks of vowels, but measured at 9 points in the vowel: 10–90 % of the duration of the vowel in intervals of 10 % of the duration of the vowel.

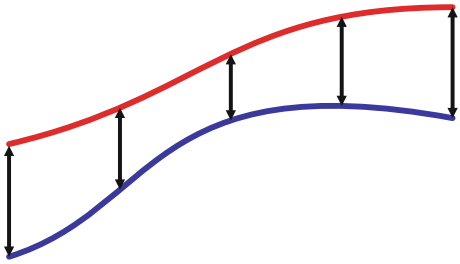


Fig. 3 Linear time normalization (compare with Fig. 1) and multi-point comparison of Greisbach et al. (1995)

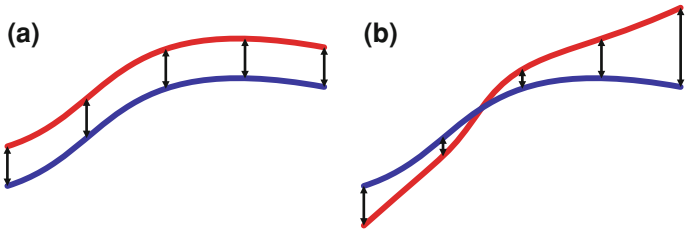


Fig. 4 Two pairs of formant tracks, each with the same Greisbach et al. (1995) dissimilarity metric value

Some of the weaknesses of the Ingram et al. (1996) and Greisbach et al. (1995) approaches were avoided by analyzing the data using discriminant analysis which treats each time point as a separate dimension rather than conflating measurements from different time points into a single metric. Variants of this procedure were also applied by Thaitechawat and Foulkes (2011) and Zuo and Mok (2011). The output of a discriminant analysis is, however, posterior probabilities for a close set of test samples, which would be incompatible with the likelihood-ratio framework (see also Morrison 2008).

4.3 Dynamic Time Warping

Although not specifically aimed at forensic application, a study by Kasuya et al. (1994) on speaker identification has been mentioned in the forensic-voice-comparison literature. Kasuya et al. used *dynamic time warping* (DTW) to calculate a distance metric between pairs of formant tracks. DTW non-linearly changes the time dimension of one track, repeating some points and skipping others, to make it correspond as closely as possible to another track (a template), see Fig. 5. The procedure includes the calculation of the smallest achievable sum of point-by-point distances between the two tracks (the minimum cost alignment, the lowest cost path between opposite corners of the cost matrix), and can thus be considered a form of multi-point comparison. It is therefore subject to the shortcomings of the multi-point comparisons outlined above.

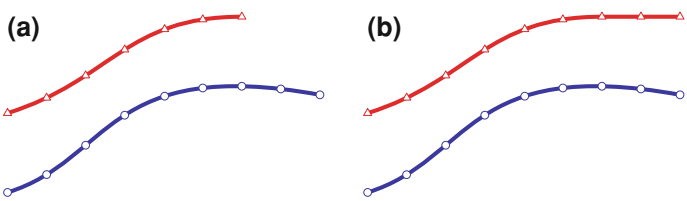


Fig. 5 **a** Original formant track (*top*) and template (*bottom*). **b** Dynamic time warped track (*top*) and template (*bottom*). Compare with linear time normalization in Fig. 3

Although DTW may be a simple and effective tool for closed-set classification, a DTW-based metric is not clearly adaptable for use within the open-set probability-of-evidence-given-competing-hypotheses likelihood-ratio framework. It is also not clear whether DTW as a preprocessing stage would be appropriate for forensic voice comparison. The potential problem is that formant tracks must be normalized towards a template and it is not immediately clear how one would select an appropriate template. The choice of template could potentially skew the results depending on if and how it affects within-speaker versus between-speaker variability.

4.4 Minimal Enclosing Rectangle

A very different approach was adopted in Rodman et al. (2002), Eriksson et al. (2004a, b). A pitch synchronous Fourier analysis was performed, and the first and second moments (means and variances) of the spectra extracted. Next the series of measurements across time were plotted as a track in a two-dimensional first- by second-moment space. Finally a *minimal enclosing rectangle* was drawn around the track (see Fig. 6), and the following features were used for forensic voice comparison: The x and y values of the centre of the rectangle, the length L of the longest side of the rectangle, and the angle of orientation of the rectangle α. The minimum and maximum x and y values of the track were also used. The features above were said to characterize the location of the track. Rodman et al. (2002) also used rather complex calculations to produce two additional features which were said to characterize the shape of the track.

The set of features in this approach appears to be arbitrary, and I am not aware of any comparisons of the minimal-enclosing-rectangle approach with more conventional parametric-curve approaches (see Sect. 4.7) to quantifying the shape of complex curves. Metrics based on spectral moments would also likely be highly susceptible to transmission-channel effects.

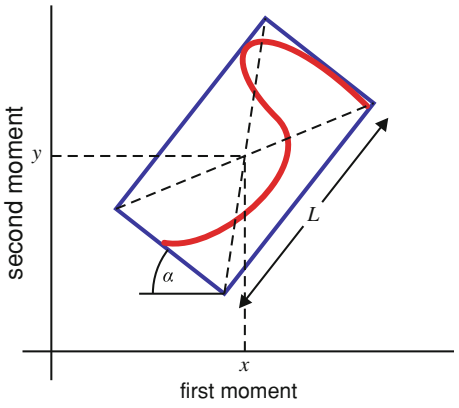


Fig. 6 Minimal enclosing rectangle of Rodman et al. (2002)

4.5 Onset + Offset Models

Some studies have used variants of multi-point comparison with the number of points being two, and thus these are versions of the *onset + offset model* which has been found to be successful for vowel phoneme identification (see Morrison 2013a, Chap. 3). Nolan and Grigoras (2005) used two points from the formant tracks of a number of diphthongs. The first point was 20 ms after the beginning of the vowel and the second point 20 ms before the end. Rose et al. (2006) also used two points to quantify VISC in /aɪ /tokens, but selected the two points according to the traditional idea (e.g., Lehiste and Peterson 1961) that diphthongs canonically have initial and final steady-state portions corresponding to the achievement and holding of articulatory targets. In practice they measured at points where F2 was stable, i.e., three or more adjacent formant measurements with "visually" the same value on a plot (the step size between measurements was not specified in the paper), or, when there was no stable portion, at the earliest point after the consonant transition or at the F2 maximum for initial and final targets respectively. According to the theory advanced in Sect. 3, models which capture the shape of the whole trajectory are expected to be more effective for forensic voice comparison than onset + offset models.

4.6 Slope Models

Another model which has been found to be effective for vowel phoneme identification (although perhaps marginally less so than onset + offset) is *onset + slope*, where slope is defined as the rate of change of formant frequency over time (see Morrison 2013a Chap. 3). Although not designed for forensic voice comparison, an early use of slope in speaker recognition was Sambur (1975) who used the value of the slope parameter from a linear least-squares fit of F2 against time for tokens of the diphthong /aɪ/, see Fig. 7a, Kinoshita and Osanai (2006) also made use of

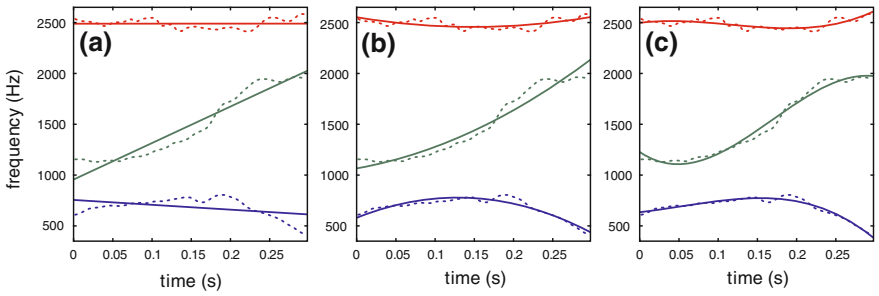


Fig. 7 **a** First-order [linear], **b** second-order [quadratic], and **c** third-order [cubic] polynomials (*solid lines*) fitted to the formant trajectories of a token of Australian English /aɪ/ (*dashed line*). The linear fit **a** is a slope model

the linear slope of F2 against time in /aɪ/, in combination with the same initial and final target values as were used in Rose et al. (2006). There is a conceptual problem with the combination of two canonical steady-state targets and a linear fit to the whole trajectory: If the formant trajectory has initial and final steady states then its shape is sigmoidal, not linear, and information about the shape of the trajectory could be better captured by a parametric-curve model which can fit a sigmoid. On the other hand, if the shape of the trajectory is linear, then one of initial formant value, slope, or final formant value is redundant.

4.7 Parametric Curves

Fitting *parametric curves* such as *polynomials* and *discrete cosine transforms* (DCTs) is the standard procedure for quantifying the shape of complex curves in statistical modeling and engineering (e.g., Hastie et al. 2009), and has previously been applied in research on VISC (Harrington 2006; Hillenbrand et al. 2001; Watson and Harrington 1999; Zahorian and Jahargi 1991, 1993). The slope model described in the previous section is a first-order polynomial, and the order of a polynomial can be increased to allow it to have more complex shapes. McDougall (2006), McDougall and Nolan (2007), Morrison (2008, 2009c), and Morrison and Kinoshita (2008) fitted first, second, and third order (linear, quadratic, and cubic) polynomials to formant trajectories of a number of vowels (see Fig. 7), and used the parameter estimates as features for forensic voice comparison. Morrison and Kinoshita (2008), Morrison (2009c, 2011a), and Morrison et al. (2010) fitted and used the parameter estimates from first, second, and third order DCTs, see Fig. 8. Morrison (2009c) concluded that, for a number of Australian English diphthongs (from among /aɪ/, /eɪ/, /aʊ/, /oʊ/, and /ɔɪ/), third order parametric curves generally outperformed second order, but differences in performance between polynomials and DCTs were minimal. Morrison (2008, 2009c) and Morrison and Kinoshita (2008) also experimented with linear and logarithmic scaling of frequency, and raw and linearly normalized time scales. No single combination of

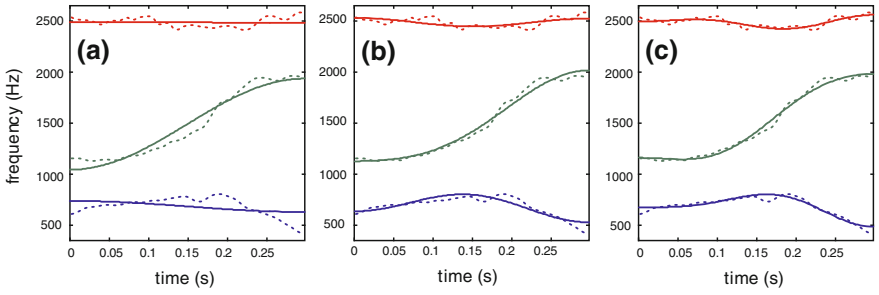


Fig. 8 First (**a**), second (**b**), and third (**c**) order DCTs (*solid lines*) fitted to the formant trajectories of a token of Australian English /aɪ/ (*dashed lines*)

scales clearly outperformed the others across vowel phonemes. Enzinger (2010) applied polynomials, DCTs, *B-splines*, and *bent-cable* models to formant trajectories of Viennese-German /aɛ/. B-splines are piecewise polynomial and potentially more flexible than DCTs or single polynomials (note that a quadratic or cubic B-spline uses more degrees of freedom than a single quadratic or cubic polynomial, see Hastie et al. 2009 Chap.5). Bent-cables are piecewise linear with a quadratic-polynomial elbow. If one had a hypothesis that diphthongs consisted of an initial steady-state portion followed by a linear glide, then the bent-cable model would be a parameterization of that hypothesis. When time normalization was applied, all models performed approximately equally well, with the exception of the bent-cable model which performed considerably worse. The validity of third-order DCTs and polynomials was slightly better than second-order in terms of C_{llr} (see Sect. 5.2) whereas this was reversed for cubic versus quadratic B-splines. The latter models being more complex than the former, it may be the case that (for this particular data set, but also perhaps in general) a parametric-curve model with complexity equal to a third-order polynomial or DCT has saturated on the amount of speaker-specific information which can be extracted from the formant trajectories of diphthongs, and that more complex models begin fitting noise. Zhang et al. (2011), however, found that fourth order DCTs on F2 and F3 resulted in the best performance for a forensic-voice-comparison system based on tokens of the Standard Chinese triphthong /iau/.

5 Does a Whole-Trajectory Model Outperform an Onset + Offset Model?

5.1 Data and Methodology

Morrison (2008) set out to directly test the hypothesis that, for forensic voice comparison, a parametric-curve model quantifying the whole of a formant trajectory would outperform an onset + offset model. The study made use of the same data as had been used in Rose et al. (2006): /aɪ/ tokens extracted from laboratory recordings of 25 male Australian English speakers recorded on two separate occasions separated by approximately 2 weeks (24 tokens per recording session). (Morrison 2008 also conducted experiments using data from all 27 speakers in the database, but direct comparisons with Rose et al. 2006, only made use of data from the same 25 speakers as used in the latter.) The use of the initial and final formant target values from Rose et al. (2006) was compared with the use of the estimated coefficient values from quadratic to cubic polynomials fitted to the formant trajectories. Cross-validated likelihood-ratio values were calculated for all possible same-speaker and different-speaker comparisons of the first versus the second session recordings using Aitken and Lucy's (2004) multivariate kernel density (MVKD) formula.

5.2 Measuring Validity

What follows is an updated presentation of the results from Morrison (2008). These results are based on the likelihood ratios calculated in Morrison (2008), but they have subsequently been calibrated, and a numeric as well as a graphical representation of system validity is also presented. Although an LR $>$ 1 favors the same-speaker hypothesis and an LR $<$ 1 favors the different-speaker hypothesis, forensic voice comparison is not a binary decision task; rather, the task is to calculate a strength of evidence. Thus, all else being equal, if a comparison is known to be a same-speaker comparison, the performance of the forensic-comparison system which outputs a likelihood ratio which is much larger than one is better than a system which outputs a likelihood ratio which is only a little larger than one. Also, all else being equal, if a comparison is known to be a same-speaker comparison, the performance of the forensic-comparison system which outputs a likelihood ratio which is much less than one is worse than a system which outputs a likelihood ratio which is only a little less than one. Mutatis mutandis for a comparison which is known to be a different-speaker comparison. A metric which captures the gradient nature of the validity of a forensic-comparison system is the log-likelihood-ratio cost, C_{llr} (Brümmer et al. 2007; Brümmer and du Preez 2006; van Leeuwen and Brümmer 2007; see also González-Rodríguez et al. 2007; Morrison 2010, 2011b). The smaller the C_{llr} value, the better the system performance. As a metric of system validity C_{llr} favors systems which minimize contrary-to-fact support for incorrect hypotheses, even if this is correlated with them being more conservative in their support for correct hypotheses. System accuracy is improved by calibration (Brümmer et al. 2007; Brümmer and du Preez 2006; Pigeon et al. 2000; Ramos Castro 2007 §6.5; van Leeuwen and Brümmer, 2007; Morrison 2012c) which can have desirable effects on the output of a forensic voice comparison system, reversing the direction of support or reducing the magnitude of likelihood ratios which provide strong support for contrary-to-fact hypotheses (Morrison and Kinoshita 2008; Morrison 2009b). The likelihood-ratio outputs from both the onset + offset and parametric-curve systems were linearly calibrated using cross-validation.

5.3 Results

Post-calibration numeric C_{llr} and graphical Tippett-plot results are discussed below (for a longer general guide to interpreting Tippett plots see the appendix to this chapter, which is an extract from Morrison 2010). For brevity only the results of tests using all three of F1, F2, and F3 are discussed. Also, of the parametric-curve systems, only the results of the system using hertz frequency, linearly normalized time, and cubic polynomials are discussed (formant frequencies in the onset + offset system were also in hertz).

Post-calibration C_{llr} for the onset + offset system was 0.43, and for the cubic-polynomial system it was 0.10. This indicates that, consistent with the theory outlined above (Sect. 3), a forensic voice comparison system which made use of information about the whole of the formant trajectory outperformed a system which made use of only the initial and final formant targets.

Figure 9 provides a Tippett plot of the calibrated likelihood ratios. The scale on the bottom of the Tippett plot is in base-ten logarithms of the likelihood-ratio values. The logarithmic scale introduces isometry with respect to support for the same- and different-speaker hypotheses: likelihood ratios in the range 0..1 are converted to log likelihood ratios in the range $-\infty..0$, and likelihood ratios in the range $1..+\infty$ are converted to log likelihood ratios in the range $0..+\infty$. Log likelihood ratios greater than zero support the same-speaker hypothesis and log likelihood ratios less than zero support the different-speaker hypothesis, and the strength of evidence is the distance from the log LR = 0 line. The curves rising to the left represent the cumulative proportion of different-speaker comparisons which resulted in log likelihood ratios equal to or greater than the value indicated on the x axis. Likewise the curves rising to the right represent the cumulative proportion of same-speaker comparisons which resulted in log likelihood ratios equal to or less than the value indicated on the x axis. For the cubic-polynomial system (solid curves) compared to the onset + offset system (dashed curves) the same-speaker curve always had larger log-likelihood-ratio values and the different-speaker curve generally had smaller log-likelihood-ratio values. The cubic-polynomial system also produced fewer log likelihood ratios which supported contrary to fact hypotheses (8 compared to 39 for the onset + offset system). The results are therefore consistent with the theory outlined in Sect. 3 that a forensic voice comparison system which made use of information about the whole of the formant trajectory would outperform a system which made use of only the initial and final formant targets.

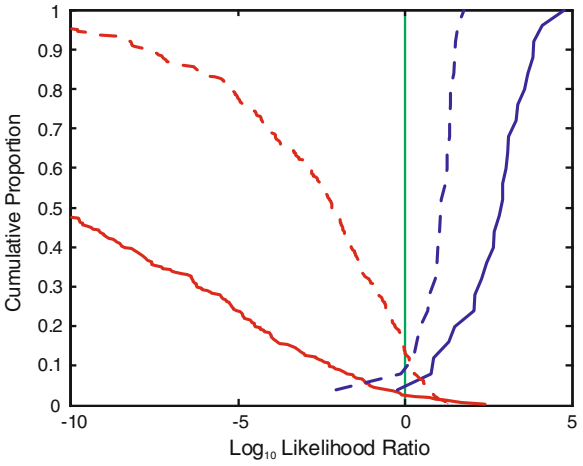


Fig. 9 Tippett plot of likelihood ratios generated from Australian English /aɪ/ tokens produced by 25 speakers by onset + offset system (*dashed curves*) and cubic-polynomial system (*solid curves*)

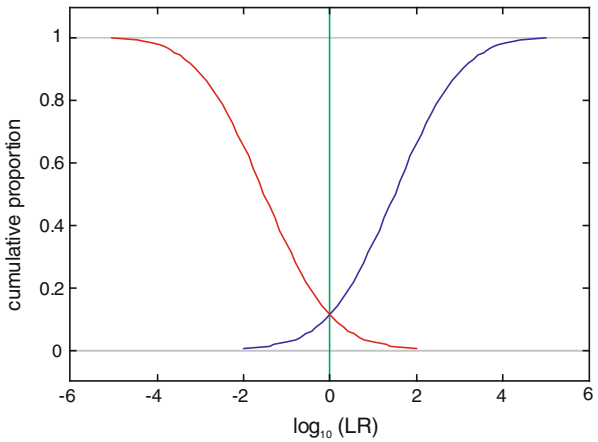


Fig. 10 Tippett plot of hypothetical test results

Fig. 11 Tippett plot of hypothetical test results

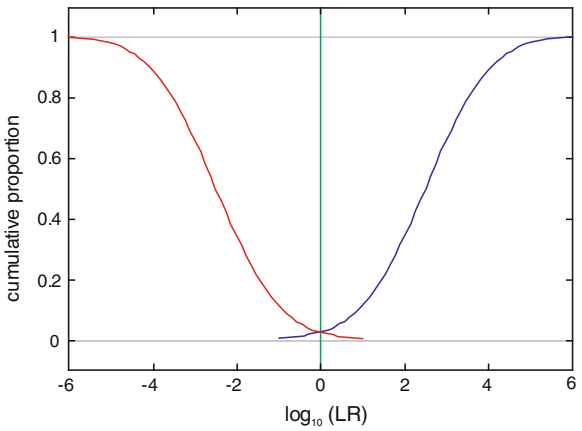
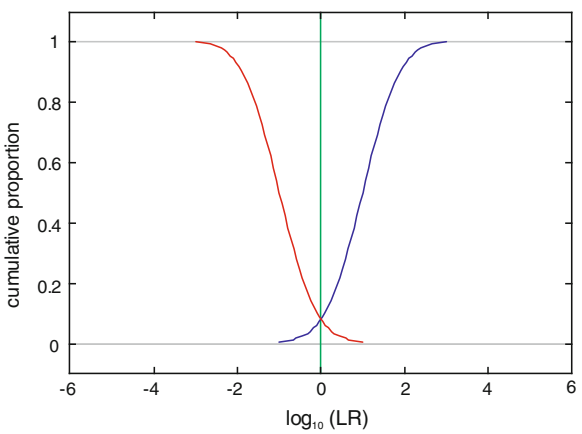


Fig. 12 Tippett plot of hypothetical test results



6 Conclusion

A review of the literature indicates that information extracted from VISC can lead to high validity in forensic-voice-comparison results. Information about the shape of the whole of the formant trajectory, particularly information extracted using parametric curves, leads to better performance than only using the initial and final formant values. It may be that similar techniques can also lead to improvements in systems based on other parameterizations of acoustic spectra, such as mel-frequency cepstral coefficients (MFCCs). This is a potential future direction for research. Experiments conducted to date have been based on relatively small databases of relatively controlled laboratory-quality voice recordings. Additional experiments will be necessary to test the validity and reliability of parametric-curve systems when applied to databases of voice recordings from a larger number of speakers and speech which is more representative of forensically realistic recording quality and speech styles, e.g., spontaneous telephone-conversation speech.

Acknowledgments Thanks to Philip Rose, Peter F. Assmann, and Stephen A. Zahorian for comments on earlier versions of this chapter. The writing of this chapter was supported by the Australian Research Council, the Australian Federal Police, New South Wales Police, Queensland Police, the National Institute of Forensic Science, the Australasian Speech Science and Technology Association, and the Guardia Civil via Linkage Project LP100200142. Unless otherwise explicitly attributed, the opinions expressed herein are those of the author and do not necessarily represent the policies or opinions of any of the above mentioned organizations or individuals.

Appendix: Interpretation of Tippett Plots

A graphical method for presenting the results of running a likelihood-ratio forensic-comparison system on a set of test data is a Tippett plot. Tippett plots were introduced in Meuwly (2001) (inspired by the work of C. F. Tippett and by Evett and Buckleton 1996), and are now a standard method for presenting results in likelihood-ratio forensic-voice-comparison research. Tippett plots provide more detailed information about the results than is available from a summary measure such as C_{llr}. This appendix is an extract from Morrison (2010 Sect. 99.930) and provides a guide to the interpretation of Tippett plots.

Figures 10, 11, 12 provide a series of Tippett plots drawn on the basis of hypothetical sets of output from forensic-comparison systems. The lines rising to the right represent the results from same-speaker comparisons in the test set, the cumulative proportion of log likelihood ratios less than or equal to the value indicated on the x axis. The lines rising to the left represent the results from different-speaker comparisons in the test set, the cumulative proportion of log likelihood ratios greater than or equal to the value indicated on the x axis. (Some authors draw both same-speaker and different-speaker lines as the cumulative proportion of log likelihood ratios greater than or equal to the value indicated on

the x axis.) In these hypothetical results the same-speaker and different-speaker lines are symmetrical and cross at a log likelihood ratio of zero; this need not be the case for real test results.

An ideal forensic-comparison system should produce a large positive log likelihood ratio for a same-origin comparison, and a large negative log likelihood ratio for a different-origin comparison. Large-magnitude log likelihood ratios which support the consistent-with-fact hypothesis are better than small-magnitude log likelihood ratios which support the consistent-with-fact hypothesis. Log likelihood ratios which support the contrary-to-fact hypothesis are bad, and the larger their magnitude the worse they are. Therefore, in Tippet plots the further apart the same-speaker and different-speaker lines (the further to the right the same-speaker line and the further to the left the different-speaker line) the better the results. The results presented in the Tippett plot in Fig. 11 are therefore better than those presented in the Tippett plot in Fig. 10.

Note, however, that (consistent with the C_{llr} metric) log-likelihood-ratio results which support contrary-to-fact hypotheses are of greater concern than whether the consistent-with-fact log-likelihood-ratio results are relatively small or large—a system which minimizes support for contrary-to-fact hypotheses is preferable even if this leads to a reduction in its strength of support for consistent-with-fact hypotheses. The results presented in the Tippett plot in Fig. 12 are therefore also better than those presented in the Tippett plot in Fig. 10.

References

Aitken, C.G.G., Lucy, D.: Evaluation of trace evidence in the form of multivariate data. Appl. Stat **54**, 109–122 (2004). doi:10.1111/j.1467-9876.2004.02031.x

Aitken, C.G.G., Roberts, P., Jackson, G.: Fundamentals of probability and statistical evidence in criminal. In: Proceedings guidance for judges, lawyers, forensic scientists and expert witnesses, Royal Statistical Society, London (2010)

Aitken, C.G.G., Taroni, F.: Statistics and the evaluation of evidence for forensic scientists. Wiley, Chichester (2004)

Balding, D.J.: Weight of evidence for forensic DNA profiles. Wiley, Chichester (2005)

Berger, C.E.H., Buckleton, J., Champod, C., Evett, I.W., Jackson, G.: Evidence evaluation: A response to the court of appeal judgment in R v T. Sci. Justice **51**, 43–49 (2011). doi:10.1016/j.scijus.2011.03.005

Brümmer, N., Burget, L., Cernocký, J.H., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D.A., Matejka, P., Schwarz, P., Strasheim, A.: Fusion of heterogenous speaker recognition systems in the STBU submission for the NIST SRE 2006. IEEE. Trans. Audio. Speech. Lang. Process **15**, 2072–2084 (2007). doi:10.1109/TASL.2007.902870

Brümmer, N., du Preez, J.: Application independent evaluation of speaker detection. Comput. Speech. Lang **20**, 230–275 (2006). doi:10.1016/j.csl.2005.08.001

Buckleton, J.: A framework for interpreting evidence. In: Buckleton, J.,Triggs, C.M.,Walsh S.J. (eds.) Forensic DNA Evidence Interpretation. Boca Raton, FL: CRC, pp. 27–63 (2005)

Champod, C., Meuwly, D.: The inference of identity in forensic speaker recognition. Speech. Commun **31**, 193–203 (2000). doi:10.1016/S0167-6393(99)00078-3

Enzinger, E.: Characterising formant tracks in Viennese diphthongs for forensic speaker comparison. In: Proceedings of the 39th Audio Engineering Society Conference—Audio Forensics: Practices and Challenges, Hillerød, Denmark. Audio Engineering Society, New York, pp. 47–52 (2010)

Eriksson, E.J., Cepeda, L.F., Rodman, R.D., McAllister, D.F., Bitzer, D., Arroway, P.: Cross-language speaker identification using spectral moments. In: Branderud, P., Traunmüller, H. (eds.) Proceedings of FONETIK 2004: The XVIIth Swedish Phonetics Conference. Stokholm, Sweden: Department of Linguistics, Stockholm University, pp. 76–79 (2004a)

Eriksson, E.J., Cepeda, L.F., Rodman, R.D., Sullivan, K.P.H., McAllister, D.F., Bitzer, D., Arroway, P.: Robustness of spectral moments: A study using voice imitations. In: Cassidy, S., Cox, F., Mannell, R., Palethorpe. (eds.) Proceedings of the 10th Australian International Conference on Speech Sciences & Technology. Australian Speech Science & Technology Association, Canberra, pp. 259–264 (2004b)

Evett, I.W.: Towards a uniform framework for reporting opinions in forensic science case-work. Sci. Justice **38**, 198–202 (1998). doi:10.1016/S1355-0306(98)72105-7

Evett, I.W.: Evaluation and professionalism. Sci. Justice **49**, 159–160 (2009). doi:10.1016/j.scijus.2009.07.001

Evett, I.W., Buckleton, J.S.: Statistical analysis of STR data. In: Carraredo, A., Brinkmann, B., Bär, W. (eds.) Advances in Forensic Haemogenetics, vol. 6, pp. 79–86. Springer, Heidelberg (1996)

Evett, I.W., and other signatories Expressing evaluative opinions: A position statement. science & justice. **51**, 1–2 (2011). doi:10.1016/j.scijus.2011.01.002

Foreman, L.A., Champod, C., Evett, I.W., Lambert, J.A., Pope, S.: Interpreting DNA evidence: A review. Int. Stat. J **71**, 473–495 (2003). doi:10.1111/j.1751-5823.2003.tb00207.x

Goldstein, U.G.: Speaker-identifying features based on formant tracks. J. Acoust. Soc. Am **59**, 176–182 (1976). doi:10.1121/1.380837

González-Rodríguez, J., Drygajlo, A., Ramos-Castro, D., García-Gomar, M., Ortega-García, J.: Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Comput. Speech. Lang **20**, 331–355 (2006). doi:10.1016/j.csl.2005.08.005

González-Rodríguez, J.,Ramos, D.: Forensic automatic speaker classification in the coming paradigm shift. In: Müller, C. (ed.) Speaker Classification I: Selected Projects. Springer-Verlag, Berlin, pp. 205–217 (2007)

González-Rodríguez, J., Rose, P., Ramos, D., Toledano, D.T., Ortega-García, J.: Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. IEEE. Trans. Audio. Speech. Lang. Process **15**, 2104–2115 (2007). doi:10.1109/TASL.2007.902747

Gottfried, M., Miller, J.D., Meyer, D.J.: Three approaches to the classification of american english diphthongs. J. Phonetics **21**, 205–229 (1993)

Greisbach, R., Esser, O., Weinstock, C.: Speaker identification by formant contours. In: Braun, A., Köster, J.-P. (eds.) Studies in Forensic Phonetics, pp. 49–55. Wissenschaftlicher, Trier, Germany (1995)

Guillemin, B.J., Watson, C.: Impact of the GSM mobile phone network on the speech signal: Some preliminary findings. Int. J. Speech, Lang. Law **15**, 193–218 (2008). doi:10.1558/ijsll.v15i2.193

Harrington, J.: An acoustic analysis of happy-tensing in the Queen's Christmas broadcasts. J. Phonetics **34**, 439–457 (2006). doi:10.1016/j.wocn.2005.08.001

Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York (2009)

Hillenbrand, J.M., Clark, M.J., Nearey, T.M.: Effect of consonant environment on vowel formant patterns. J. Acoust. Soc. Am. **109**, 748–763 (2001). doi:10.1121/1.1337959

Ingram, J.C.L., Prandolini, R., Ong, S.: Formant trajectories as indices of speaker identification. Forensic. Linguist. Int. J. Speech. Lang. Law **3**, 129–145 (1996)

Jessen, M.: Forensic phonetics language and linguistics. Compass **2**, 671–711 (2008). doi:10.1111/j.1749-818x.2008.00066.x

Kasuya, H., Tan, X., Yang, C.-S.: Voice source and vocal tract characteristics associated with speaker individuality. In: Proceedings of the 3rd International Conference on Spoken-Language Processing, Yokohama, pp. 1459–1462 (1994)
Kinoshita, Y., Osanai, T.: Within speaker variation in diphthongal dynamics: What can we compare?. In: Warren, P., Watson, C.I. (eds.) Proceedings of the 11th Australasian International Conference on Speech Science & Technology, Auckland, New Zealand. Australia: Australasian Speech Science & Technology Association, Canberra, pp. 112–117 (2006)
Kuhn, T.S.: The Structure of Scientific Revolutions. University of Chicago Press, Chicago (1962)
Lehiste, I., Peterson, G.E.: Transitions, glides, and diphthongs. J. Acoust. Soc. Am **33**, 268–277 (1961). doi:10.1121/1.1908681
Lucy, D.: Introduction to Statistics for Forensic Scientists. Wiley, Chichester (2005)
McDougall, K.: Speaker-specific formant dynamics: an experiment on Australian English /aɪ/. Int. J. Speech. Lang. Law **11**, 103–130 (2004)
McDougall, K.: Dynamic features of speech and the characterization of speakers. Int. J. Speech. Lang. Law **13**, 89–126 (2006)
McDougall, K., Nolan F.: Discrimination of speakers using the formant dynamics of /u/ in British English. In: Trouvain, J., Barry, W.J. (eds.) Proceedings of the 16th International Congress on Phonetic Sciences, Saarbrücken. Saarbrücken, Germany, pp. 1825–1828 (2007)
Meuwly, D.: Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique. Dissertation, University of Lausanne, Switzerland (2001)
Morrison, G.S.: Forensic voice comparison using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English /aɪ/. Int. J. Speech. Lang. Law **15**, 247–264 (2008). doi:10.1558/ijsll.v15i2.249
Morrison, G.S.: Comments on Coulthard & Johnson's (2007) portrayal of the likelihood-ratio framework. Aust. J. Forensic. Sci **41**, 155–161 (2009a). doi:10.1080/00450610903147701
Morrison, G.S.: Forensic voice comparison and the paradigm shift. Sci. Justice **49**, 298–308 (2009b). doi:10.1016/j.scijus.2009.09.002
Morrison, G.S.: Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. J. Acoust. Soc. Am. **125**, 2387–2397 (2009c). doi:10.1121/1.3081384
Morrison, G.S.: Forensic voice comparison. In: Freckelton, I., Selby, H. (eds.) Expert Evidence (Ch. 99). Sydney, Australia: Thomson Reuters (2010)
Morrison, G.S.: A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multvariate kernel density (MVKD) versus gaussian mixture model—universal background model (GMM-UBM). Speech. Commun **53**, 242–256 (2011a). doi:10.1016/j.specom.2010.09.005
Morrison, G.S.: Measuring the validity and reliability of forensic likelihood-ratio systems. Sci. Justice **51**, 91–98 (2011b). doi:10.1016/j.scijus.2011.03.002
Morrison, G.S.: Static and dynamic approaches to understanding vowel perception. In: Morrison, G.S., Assmann, P.F. (eds.) Theories of vowel inherent spectral change (ch. 3). Springer Verlag, Heidelberg (2013a)
Morrison, G.S.: The likelihood-ratio framework and forensic evidence in court: A response to R v T. Int. J. Evid. Proof **16**, 1–29 (2012b). http://vathek.org/doi/abs/10.1350/ijep.2012.16.1.390
Morrison, G.S.: Tutorial on logistic regression calibration and fusion: Converting a score to a likelihood ratio. Aus. J. Forensic Sci. online 31 Oct 2012 (2012c). doi:10.1080/00450618.2012.733025
Morrison, G.S., Kinoshita, Y.: Automatic-type calibration of traditionally derived likelihood ratios: Forensic analysis of Australian English /o/ formant trajectories. In: Proceedings of Interspeech Incorporating *SST,* International Speech Communication Association, pp. 1501–1504 (2008)
Morrison, G.S., Ochoa, F., Thiruvaran, T.: Database selection for forensic voice comparison. In: Proceedings of Odyssey 2012: The Language and Speaker Recognition Workshop, Singapore International Speech Communication Association, pp. 62–77 (2012)

Morrison, G.S., Thiruvaran, T., Epps, J.: Estimating the precision of the likelihood-ratio output of a forensic-voice-comparison system. In: Proceedings of Odyssey 2010: The Language and Speaker Recognition Workshop, Brno, Czech Republic. International Speech Communication Association (2010)

Nearey, T.M., Assmann, P.F.: Modeling the role of vowel inherent spectral change in vowel identification. J. Acoust. Soc. Am **80**, 1297–1308 (1986). doi:10.1121/1.394433

Nolan, F.: Speaker recognition and forensic phonetics. In: Hardcastle, W.J., Laver, J. (eds.) The Handbook of Phonetic Sciences, pp. 744–767. Blackwell, Oxford (1997)

Pigeon, S., Druyts, P., Verlinde, P.: Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. Digital Signal Process **10**, 237–248 (2000). doi:10.1006/dspr.1999.0358

Ramos Castro, D.: Forensic evaluation of the evidence using automatic speaker recognition systems. Dissertation, Universidad Autónoma de Madrid, Madrid, Spain (2007)

Robertson, B., Vignaux, G.A.: Interpreting Evidence. Wiley, Chichester (1995)

Rodman, R., McAllister, D., Bitzer, D., Cepeda, L., Abbitt, P.: Forensic speaker identification based on spectral moments. Int. J. Speech. Lang. Law **9**, 22–43 (2002)

Rose, P.: Forensic Speaker Identification. Taylor & Francis, London (2002)

Rose, P.: *The technical comparison of forensic voice samples*. In: Freckelton, I., Selby, H. (eds.) Expert Evidence (Ch. 99). Sydney, Australia: Thomson Lawbook (2003)

Rose, P.: Forensic speaker recognition at the beginning of the twenty-first century: An overview and a demonstration. Aust. J. Forensic. Sci **37**, 49–72 (2005). doi:10.1080/00450610509410616

Rose, P.: Technical forensic speaker recognition: Evaluation, types and testing of evidence. Comput. Speech. Lang **20**, 159–191 (2006). doi:10.1016/j.csl.2005.07.003

Rose, P., Kinoshita, Y., Alderman, T.: Realistic extrinsic forensic speaker discrimination with the diphthong /aɪ/. In: Warren, P., Watson, C.I. (eds.) Proceedings of the 11th Australasian International Conference on Speech Science & Technology, Auckland, New Zealand. Canberra, Australia: Australasian Speech Science & Technology Association, pp. 329–334 (2006)

Rose, P., Morrison, G.S.: A response to the UK position statement on forensic speaker comparison. Int. J. Speech. Lang. Law **16**, 139–163 (2009). doi:10.1558/ijsll.v16i1.139

Saks, M.J., Koehler, J.J.: The coming paradigm shift in forensic identification science. Science **309**, 892–895 (2005). doi:10.1126/science.1111565

Sambur, M.R.: Selection of acoustic features for speaker identification. IEEE. Trans. Acoust. Speech. Signal. Process **23**, 176–182 (1975). doi:10.1109/TASSP.1975.1162664

Taitechawat, S., Foulkes, P.: Discrimination of speakers using tone and formant dynamics in Thai. In: Lee, W.-S., Zee, E. (eds.) Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China. Hong Kong: Organizers of ICPhS XVII at the Department of Chinese, Translation and Linguistics, City University of Hong Kong, pp. 1975–1981 (2011)

van Leeuwen, D.A., Brümmer, N.: An introduction to application-independent evaluation of speaker recognition systems. In: Müller, C. (ed.) Speaker Classification I: Selected Projects, pp. 330–353. Springer-Verlag, Berlin (2007)

Watson, C., Harrington, J.: Acoustic evidence of dynamic formant trajectories in Australian English vowels. J. Acoust. Soc. Am. **106**, 458–468 (1999). doi:10.1121/1.427069

Zahorian, S.A., Jagharghi, A.J.: Speaker normalization of static and dynamic vowel spectral features. J. Acoust. Soc. Am **90**, 67–75 (1991). doi:10.1121/1.402350

Zahorian, S.A., Jagharghi, A.J.: Spectral-shape features versus formants as acoustic correlates for vowels. J. Acoust. Soc. Am **94**, 1966–1982 (1993). doi:10.1121/1.407520

Zhang, C., Morrison, G.S., Thiruvaran, T.: Forensic voice comparison using Chinese/iau/. In: Lee, W.-S., Zee, E. (eds.) Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China. Hong Kong: Organizers of ICPhS XVII at the Department of Chinese, Translation and Linguistics, City University of Hong Kong, pp. 2280–2283 (2011)

Zuo, D., Mok, P.P.K.: Formant dynamics of/ua/in the speech of Mandarin-Shanghainese bilingual identical twins. In: Lee, W.-S., Zee, E. (eds.) Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China. Hong Kong: Organizers of ICPhS XVII at the Department of Chinese, Translation and Linguistics, City University of Hong Kong, pp. 2332–2335 (2011)

Index

G. S. Morrison and P. F. Assmann (eds.), *Vowel Inherent Spectral Change*, Modern Acoustics and Signal Processing, DOI: 10.1007/978-3-642-14209-3,