Prosody, Phonology and Phonetics

Keikichi Hirose
Jianhua Tao *Editors*

Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis

Springer

Prosody, Phonology and Phonetics

The series will publish studies in the general area of Speech Prosody with a particular (but non-exclusive) focus on the importance of phonetics and phonology in this field. The topic of speech prosody is today a far larger area of research than is often realised. The number of papers on the topic presented at large international conferences such as Interspeech and ICPhS is considerable and regularly increasing. The proposed book series would be the natural place to publish extended versions of papers presented at the Speech Prosody Conferences, in particular the papers presented in Special Sessions at the conference. This could potentially involve the publication of 3 or 4 volumes every two years ensuring a stable future for the book series. If such publications are produced fairly rapidly, they will in turn provide a strong incentive for the organisation of other special sessions at future Speech Prosody conferences.

More information about this series at http://www.springer.com/series/11951

Keikichi Hirose • Jianhua Tao
Editors

Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis

Springer

Editors
Keikichi Hirose
Graduate School of Information
Science and Technology
University of Tokyo
Tokyo
Japan

Jianhua Tao
Institute of Automation
Chinese Academy of Sciences
Beijing
China

ISSN 2197-8700 ISSN 2197-8719 (electronic)
Prosody, Phonology and Phonetics
ISBN 978-3-662-45257-8 ISBN 978-3-662-45258-5 (eBook)
DOI 10.1007/978-3-662-45258-5

Library of Congress Control Number: 2014955166

Springer Berlin Heidelberg Dordrecht London

Printed on acid-free paper

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

Contents

Part III Control of Prosody in Speech Synthesis

Contributors

Daniel Aalto University of Helsinki, Helsinki, Finland

Nick Campbell Trinity College Dublin, The University of Dublin, Dublin, Ireland

Yoko Greenberg Waseda University, Tokyo, Japan

Keikichi Hirose The University of Tokyo, Tokyo, Japan

Daniel J. Hirst CNRS & Aix-Marseille University, Aix-en-Provence, France

Tongji University, Shanghai, China

Hirokazu Kameoka The University of Tokyo, Tokyo, Japan/NTT Communication Science Laboratories, Atsugi, Japan

Hideki Kawahara Wakayama University, Wakayama, Japan

Takao Kobayashi Tokyo Institute of Technology, Tokyo, Japan

Pierre Lanchantin Cambridge University, Cambridge, UK

Ya Li Institute of Automation, Chinese Academy of Sciences, Beijing, China/Trinity College Dublin, The University of Dublin, Dublin, Ireland

Hansjörg Mixdorff Beuth-Hochschule für Technik Berlin, Berlin, Germany

Nicolas Obin IRCAM, UMR STMS IRCAM-CNRS-UPMC, Paris, France

Santitham Prom-on King Mongkut's University of Technology Thonburi, Thailand

Yoshinori Sagisaka Waseda University, Tokyo, Japan

Antti Suni University of Helsinki, Helsinki, Finland

Jianhua Tao Institute of Automation, Chinese Academy of Sciences, Beijing, China

Martti Vainio University of Helsinki, Helsinki, Finland

Christophe Veaux Centre for Speech Technology Research, Edinburgh, UK

Miaomiao Wang Toshiba China R&D Center, Beijing, China

Yi Xu University College London, London, UK

Kai Yu Shanghai Jiao Tong University, Shanghai, China

Part I
Modeling of Prosody

Chapter 1
ProZed: A Speech Prosody Editor for Linguists, Using Analysis-by-Synthesis

Daniel J. Hirst

Abstract This chapter describes a tool designed to allow linguists to manipulate the prosody of an utterance via a symbolic representation in order to evaluate linguistic models. Prosody is manipulated via a Praat TextGrid which allows the user to modify the rhythm and melody. Rhythm is manipulated by factoring segmental duration into three components: (i) intrinsic duration determined by phonemic identity (ii) local modifications encoded on the rhythm tier and (iii) global variations of speech rate encoded on the intonation tier. Melody is similarly determined by tonal segments on the tonal tier (= pitch accents) and on the intonation tier (= boundary tones) together with global parameters of *key* and *span* determining changes of pitch register. The TextGrid is used to generate a manipulation object that can be used either for immediate interactive assessment of the prosody determined by the annotation, or to generate synthesised stimuli for more formal perceptual experiments.

1.1 Introduction

The interaction between linguists and engineers has always been a productive area of exchange. This is particularly evident in the area of speech prosody. The analysis by synthesis paradigm is an attractive one for linguists, since it provides an empirical solution to the problem of validating an abstract model. If the representation derived from a model can be used as input to a speech synthesis system, and if the contrasts represented in the model are correctly rendered in the synthetic speech, then the representation can be assumed to contain all the information necessary to express that contrast.

Although speech technology has become more and more accessible in recent years, it remains nonetheless true that the gap between applications and users is still far too wide. This is unfortunate, since there are a great number of linguists throughout the world who are particularly interested in developing and assessing different models of prosodic structure.

D. J. Hirst (✉)
LPL, UMR 7309, CNRS & Aix-Marseille University, Aix-en-Provence, France
e-mail: daniel.hirst@lpl-aix.fr

School of Foreign Languages, Tongji University, Shanghai, China

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_1

Providing linguists with better tools will surely result in the availability of more and better data on a wide variety of languages, and such data will necessarily be of considerable interest to engineers working with speech technology.

In this presentation, I introduce the latest implementation of *ProZed*, a program specifically designed to allow linguists to manipulate the prosody of utterances on a symbolic level, providing an acoustic output which is directly controlled by a symbolic level of representation.

The implementation of ProZed is designed to be entirely language-independent and as theory-neutral as possible, although it obviously integrates a number of non-trivial principles which I have adopted over the years. It is hoped, however, that while it is never, of course, possible to be entirely theory-neutral, this software will at least prove to be *theory-friendly* in that it will be compatible with a number of different theoretical frameworks, and it may prove capable of providing enough evidence to allow a user to choose between various different theoretical options.

The prosody of speech can be defined for the purposes of this presentation as the explicit characterisation of the length, pitch and loudness of the individual sounds that make up an utterance. Even this fairly wide definition may be found to be too restrictive for some, who may regret the absence of any consideration of e.g. voice quality here. In the current implementation, only the length and pitch of speech sounds are treated, since it seems likely that an efficient manipulation of loudness will require modification of the distribution of energy in the spectrum rather than simply increasing or decreasing the overall intensity of the sound. There is, of course, nothing in the ProZed framework itself that is incompatible with the representation of voice quality and this could well be integrated into the same framework, as and when algorithms for the manipulation of these characteristics of speech become more generally available.

1.2 The General Framework

ProZed is implemented as a plugin to the Praat software (Boersma and Weenink 2014). It allows the manipulation of the rhythmic and tonal aspects of speech as defined on two specific tiers, respectively named the *rhythm* tier and the *tonal* tier. These two tiers control the short term variability of prosody. Longer-term variations are controlled via a third tier named the *intonation* tier. For discussion of the rationale behind the choice of these units cf. Hirst (2012).

The program also includes a function for the automatic display of the prosody of an utterance, which is designed to provide an intuitive way to compare the prosody of two utterances produced by the same or by different speakers, or to compare the prosody of utterances in different languages or dialects.

The speech input to the program can be natural recorded speech, the prosodic characteristics of which may then be modified by the software, oralternatively, it could be the output of a speech synthesis system with, for example, fixed (or mean) durations for each speech segment.

The current version of the program is designed as the resynthesis step of what is planned to be a complete analysis-by-synthesis cycle. This will subsequently be directly integrated with the output of the Momel-Intsint and ProZed Rhythm analysis models which are described below as well as with the automatic alignment of phonemes and syllables as provided by the recently developed *SPPAS* tool described in Bigi and Hirst (2012, 2013).

1.3 Using a TextGrid to Modify the Prosody of Utterances

The annotation of the prosody of an utterance is encoded via three interval tiers. These are:

- the rhythm tier
- the tonal tier
- the intonation tier

While it is hoped that linguists will find these tiers appropriate and useful levels for modeling the rhythm and melody of speech, no assumptions are made as to the phonological units corresponding to the intervals of these tiers. *Rhythm Units*, *Tonal Units* and *Intonation Units* are consequently *defined*, respectively, as the domains of short term lengthening/shortening, short-term pitch control and longer-term variation in both duration and pitch.

For different linguists, these units may correspond to different phonological entities. Thus, for example, for some linguists, the Rhythm Units and/or Tonal Units may be identified with the phonological *syllable*, while for others they may correspond to larger units such as the *stress foot* or the *phonological word*.

Work with my students (Bouzon and Hirst 2004) suggests that, as originally proposed by Wiktor Jassem (1952), for English at least, the *Narrow Rhythm Unit* and *Anacrusis* are the most appropriate domains for rhythm, while the slightly larger stress foot (= Jassem's *Tonal Unit*) seems more appropriate for modeling pitch accents.

The ProZed editor is designed to provide a means to implement any of these different interpretations in order to evaluate the effect of the different choice of units.

1.3.1 Determining Segmental Duration via the Rhythm Tier

The implementation of rhythmic characteristics in the ProZed environment makes the fairly consensual assumption that segmental duration in speech is the result of the combination of at least two factors. The first of these is the intrinsic duration of individual speech sounds. A /ʃ/ sound, for example, is intrinsically much longer than a /l/ sound.

The second factor is a domain-specific lengthening, which, in this implementation, following Hirst (2005), is modelled as a scalar lengthening by an integral number

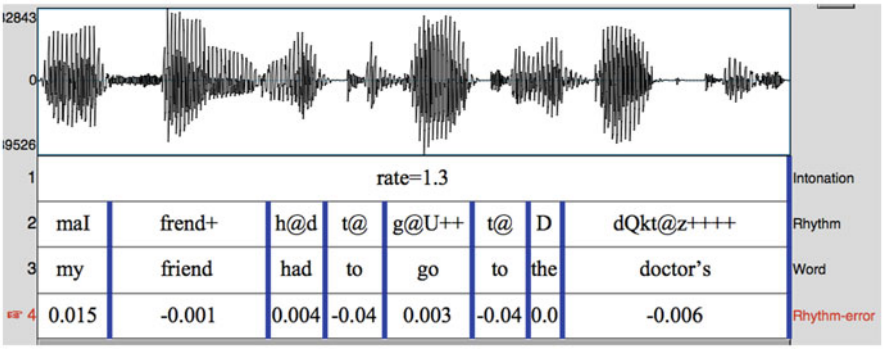


Fig. 1.1 TextGrid for the sentence "My friend had to go to the doctor's" showing the Rhythm tier and the Word tier together with a third tier, Rhythm-error, generated by the program, displaying the difference between the predicted and the observed durations

of quantal units. The quantal units, by default 50 ms each, are added to the sum of the intrinsic durations of the speech segments transcribed within the given rhythm unit. These can be thought of as "dummy" phonemes, which each add the duration of a very short phoneme to the rhythm unit, the sum duration of which is spread out uniformly across the whole unit.

The resulting value is finally corrected to take into account the current value of speech rate.

The formula given in Hirst and Auran (2005) is:

$$\hat{d}_{ru} = (\sum_{i=1}^{m} \bar{d}_{i/p} + k * q) * t \tag{1.1}$$

where $\hat{d}_{ru}$ is the predicted duration of the Rhythm Unit, $\bar{d}_{i/p}$ corresponds to the mean value of the phoneme p in the corpus, q is the quantal unit of lengthening and k the scalar value of that lengthening. The final variable t, for tempo, (corresponding to $\frac{1}{rate}$), is applied to the duration of the Rhythm Unit as a whole.

To take an example, the word "go" (in Fig. 1.1), is represented on the rhythm tier as: [g@U++].

The predicted duration of the Rhythm Unit is determined by a combination of the individual mean durations of its constituent segments, plus the lengthening factor annotated by the two plus signs.

Assuming that the individual mean durations of the phonemes /g/ and /@U/, as found in an external table are, respectively, 90 and 57 ms, the total duration of the *Tonal Unit* will be calculated as 147 ms plus 100 ms of lengthening as determined by the 2 + s, i.e. a total of 247 ms, which will then be further modified by dividing by the specified speech rate factor of 1.3. The resulting predicted value of 190 ms is very close to the observed value of 187 ms.

The difference between the predicted and the observed durations of each rhythm unit is calculated and displayed on a new tier: (*Rhythm-error*).

The duration of the Rhythm Unit can be manipulated linearly so that the synthesised duration is made to correspond to that determined by the symbolic representation. Thus, in the above example, the duration of the Rhythm Unit containing the segment corresponding to the phonemes /g@U/ would be globally adjusted to a duration of 190 ms.

The user is, of course, encouraged to experiment with different values of lengthening and speech rate in order to test various hypotheses concerning their interaction, as well as to experiment with different domains for the implementation of the lengthening.

In the current version of the program, there is no specific mechanism to implement final lengthening, other than by creating an ad hoc Rhythm Unit which is coextensive with the domain in which final lengthening is assumed to apply (such as the final syllable for example). This is an area in which the implementation could be improved in future versions in the light of work in progress on this type of lengthening, some preliminary results of which were reported in Hirst (2009).

1.3.2 Determining Pitch via the Tonal Tier

Pitch in ProZed is determined by a representation of the contour using the *INTSINT* alphabet (Hirst 2005). This assumes that a pitch contour can be adequately represented by a sequence of target points, each pair of which is linked by a continuous smooth monotonic interpolation (defining a quadratic spline function).

This, in turn, assumes that the shape of a pitch-accent, for example, is entirely determined by the time and frequency values of the relevant target points. I have never seen a convincing example of an pitch contour which cannot be adequately modelled in this way. Readers are invited to contact me if they believe they have such an example.

The pitch height of a target is determined by the symbolic "tonal" symbol from the INTSINT alphabet which is defined either globally with respect to the speaker's current register (see below) or locally, with respect to the preceding target.

Globally, the target may be at the *top*, *middle* or *bottom* of the current pitch range, as defined by the parameters *key* and *span* (see below), and is consequently marked respectively as *t*, *m* or *b*. Locally, the pitch targets may be interpreted as being *higher*, the *same*, or *lower* than the preceding target (respectively coded as *h*, *s* or *l*). They may also be coded as *upstepped* or *downstepped* (*u* or *d*), corresponding to a smaller interval up from or down from the preceding target. Note that in this implementation, the INTSINT tones are represented with lowercase letters rather than uppercase, as used in previous work. This helps to avoid confusion with other more abstract coding schemes such as ToBI (Pierrehumbert 1980; Silverman et al. 1992), or the even more abstract underlying representation used in Hirst (1998), both of which use some of the same symbols as INTSINT.

The actual fundamental frequency of the pitch targets is determined by the following formulas (where *p* is the value of the preceding target) and where pitch range

is defined by the current values of two parameters: key (in Hertz) and $span$ (in octaves):

absolute tones:
- t: $key * \sqrt{2^{span}}$ (i.e. half the value of span above key)
- m: key
- b: $key / \sqrt{2^{span}}$ (i.e. half the value of span below key)

relative tones:
- h: $\sqrt{p * t}$ (i.e. the (geometric) mean of p and t)
- s: p
- l: $\sqrt{p * b}$ (i.e. the (geometric) mean of p and b)

iterative tones:
- u: $\sqrt{p * \sqrt{(p * t)}}$ i.e. the (geometric) mean of p and the mean of p and t $(= h)$
- d: $\sqrt{p * \sqrt{(p * b)}}$ i.e. the (geometric) mean of p and the mean of p and b $(= h)$

The timing of the target points is assumed to be determined with respect to the boundaries of the corresponding Tonal Unit. In previous work (e.g. Hirst 1999), I suggested that this timing might be limited to a restricted inventory of positions within the Tonal Unit, such as initial, early, mid, late and final.

In this implementation, I adopt a more general solution and allow, in fact, an arbitrary precision of alignment via the use of "dummy" targets represented by the symbol "-". Using this annotation, a tonal target X which is alone in the middle of a unit will be coded [X]. When there are more than one tonal targets in a Tonal Unit, they are assumed to be spread out evenly, so that [W X] will have one target occurring at the first quarter of the duration and one at the third quarter of the duration.

Consequently, for two consecutive Tonal Units (of the same duration) each containing two targets, the four targets will be all be spaced equally apart. In order to represent a target at the third quarter of the duration with no preceding target, the annotation [-X] can be used. The symbol "-" is thus used to influence the timing of the other target but does not itself correspond to a pitch target.

The formula for calculating the timing of the ith target of a sequence of n targets in a Tonal Unit beginning at time $start$ and ending at time end is:

$$t = start + \frac{(2i - 1) * [end - start]}{2n} \tag{1.2}$$

In practice, it is assumed that a linguist will want to make fairly sparse use of these dummy symbols, but the annotation in fact allows the specific timing of a target or targets to be coded to an arbitrary degree of precision. Thus, a representation like [- - X - - Y - Z - - - -], for example, could be used to specify timing very precisely, where in this case the three targets would occur at 0.208, 0.458 and 0.625 of the duration of the interval, respectively, (calculated as $(2 * i - 1)/2n$, with n =12 (number of tonal positions defined) and $i = 3$, 6 and 8).

The actual precision of the timing is consequently left to the user to determine. It is particularly interesting to use an annotation system which can be rendered as precise or as general as wished, so that the same annotation can be used in the analysis and in the synthesis steps of the analysis-by-synthesis procedure.

1.3.3 *Defining Long-Term Parameters with the* Intonation *Tier*

The short term values obtained from the Rhythm and Tonal tiers are modified by the long-term parameters defined on the Intonation tier. In the current implementation, these are *rate* for rhythm and *key* and *span* for pitch. The three parameters are initialised with default values:

$$rate = 1;\ key = 150;\ span = 1$$

and then, any of the values can be modified for subsequent Intonation Units by simply including a specification of the value of the corresponding parameter or parameters. Thus, for example:

$$rate = 1.5;\ span = 0.8$$

on the Intonation tier, will make the speaking rate faster and the pitch span more reduced from that Intonation Unit onwards.

Each modification of a long-term value remains valid unless it is modified in a later Intonation Unit. The implementation makes the assumption that changes of these parameters only occur at the onset of an Intonation Unit.

The program also allows the definition of pitch targets at the extreme ends of an Intonation Unit; using the annotation $[mb]$, for example, will place a *mid* target located at the beginning of the unit and a *bottom* target located at the end. Dummy targets can also be used here, so $[-b]$ will place only a bottom target at the end of the unit with nothing at the beginning whereas $[m-]$ will place a target at the beginning of the unit with nothing at the end. This corresponds essentially to the targets interpreted as *boundary tones* in many phonological prosodic models.

The pitch targets defined on the Tonal and Intonation tiers are output in the form of a Pitch Tier which is then converted to a quadratic spline function using the Praat function *Interpolate quadratically*. The resulting Pitch Tier can then be used to replace the original Pitch via a Manipulation object, allowing the resynthesised version of the utterance to be compared with the original sound.

1.4 Integrating the Synthesis with the Output from Automatic Analysis

1.4.1 *Automatic Analysis of Rhythm*

The automatic alignment software, SPPAS (Bigi and Hirst 2012, 2013) can produce a TextGrid automatically from a sound and an orthographic transcription of the sound, provided that the appropriate resources (an orthography-to-phoneme dictionary and an acoustic model for the language) are available for the language.

The TextGrid produced contains interval tiers for both *phonemes* and *words*. An algorithm for automatically producing a syllable tier from the phoneme tier has been developed for French. The software is currently implemented for French, English,

<parameter tempo=0.761><parameter quant=50>
I have a problem1 with my water3 softener7. The1 water3 level1 is1 too4 high5 and the2 over-flow2 keeps2 dripping4. Could you1 a-rrange3 to send2 an engi-neer2 on Tuesday morning2 please6. It's the2 only day1 I1 can manage1 this1 week3. I'd be grateful if you could confirm2 the a'rrangement in1 writing6.

Fig. 1.2 A sample passage from the Eurom1 corpus coded for duration using the automatic coding scheme described above. Rhythm Units are delimited by spaces or hyphens, numbers correspond to the scalar lengthening factor k, applied to the preceding Rhythm Unit

Italian and Chinese (Mandarin), as well as partially for Taiwanese. Other languages are also currently in the process of being implemented and collaboration to extend the coverage of the system will be extremely welcome. The software and a tutorial describing the program's implementation and usage are freely downloadable from:

http://www.lpl.univ-aix.fr/∼bigi/sppas/

The latest version of ProZed (Hirst 2014) integrates an algorithm for creating a syllable tier from a phoneme tier and a word tier for English. The algorithm implements the Maximum Onset Principle (Kahn 1980), according to which, when two vowels in a word are separated by a sequence of consonants, the syllable boundary occurs before the maximum sequence of consonants which can constitute a well-formed word onset. Thus in a word like "extra" /'ekstrə/ the syllable boundary between /e/ and /ə/ might be: /'e.kstrə/, /'ek.strə/, /'eks.trə/, /'ekst.rə/ or /'ekstr.ə/. Since /str/ is a well-formed word onset but /kstr/ is not, the Maximum Onset Principle will choose the syllable boundary /'ek.strə/.

The model of rhythm described in Hirst and Auran (2005) models segmental duration by optimising the function given above in Eq. (1.1). An automatic algorithm to optimise this function is currently being tested and will be integrated in the ProZed plugin.

The output of the algorithm will be a TextGrid which can be used as input to the synthesis module as described above. There will also be an optional textual output as in Fig. 1.2:

The Rhythm Units are delimited by spaces or hyphens, which for the majority of English words results in a maximally economical annotation for the rhythm; a hyphen is only needed when a lexical stress occurs on a noninitial syllable, as in *engi-neer*, *a-rrange* or *over-flow*. Note that for rhythm, no distinction is made between primary and secondary stress so the secondary stress on the second syllable of "overflow" is preceded by a hyphen just like the primary stress on the third syllable of "engineer". The numbers after the Rhythm Units are a shortcut annotation for the equivalent number of (+) signs so that "water3" corresponds to "wO:t@+++" in the TextGrid.

The scalar lengthening, as described above, applies linearly to the whole preceding Rhythm Unit. Because of this, it is not necessary to have an alignment of the individual phonemes. Informal experimentation with this program suggests that this type of global lengthening of Rhythm Units is quite an efficient way of obtaining

a synthetic output and that listeners are much more sensitive to differences in the duration of the Rhythm Units themselves than they are to internal differences within the Rhythm Units. This, however, is clearly an area which needs more experimental research, to which it is hoped that this program may contribute.

1.4.2 Automatic Analysis of Pitch

The output of the Momel and INTSINT algorithms, as described in Hirst and Espesser (1993) and Hirst (2005, 2011) and implemented in Hirst (2007) can be directly used as input to the resynthesis module as described above. The temporal alignment of the targets currently needs to be determined manually with respect to the boundaries of the Tonal Units, but this step will also be automated in the near future.

1.5 Displaying Speech Prosody

It is, of course, desirable to compare the output of the prosody modelled with the editor with that obtained by the automatic analysis. The linguist may also wish to compare the production of different speakers, or that of a speaker and a synthetic voice, in order to try to decide what aspects of the production are critical for the interpretation of the utterance and what aspects are merely individual variations without any interpretative significance.

Such comparisons can be statistical, comparing for example the root mean-square error of the duration of the phonemes, syllables or Rhythm Units, or comparing the time and frequency values of the pitch points which define the Momel curve. Such statistical comparisons, however, despite their objectivity, need to make a number of underlying assumptions about what exactly is to be compared. There is, today, no general consensus on what statistical measurements best reflect prosodic similarity between two versions of an utterance.

A more intuitive comparison can be obtained by comparing visual displays of the prosody combined with listening to the different versions. In the more informal stages of developing a model of the prosody of a language, this can be an extremely useful tool for a linguist.

Nevertheless there are still a number of assumptions that need to be made concerning the nature of the visual display.

What units of duration should be displayed? What type of display gives a visual impression which is closest to the oral comparison? How can we compare pitch patterns produced by two different speakers, possibly a male speaker and a female speaker, for example?

ProZed implements a function (*Display Prosody* . . .) to display the prosody of an utterance. The display is designed to illustrate both the rhythm and the melody.

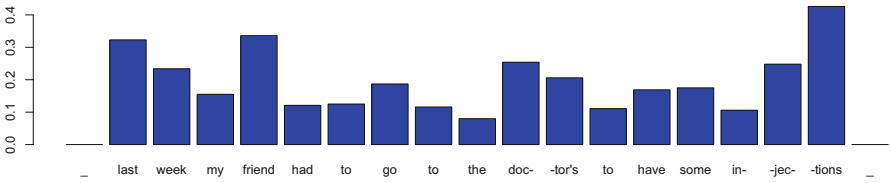


Fig. 1.3 Barplot for the syllable durations of the sentence "Last week my friend had to go to the doctor's". The height of each bar is proportional to the duration of the corresponding syllable

1.5.1 Displaying Rhythm

Gustafson (1987) notes that, traditionally, there are two ways of graphically representing the duration of the different units which make up the rhythm of the utterance.

The first is to represent the duration vertically, as for example, by a barplot, consisting of a sequence of equally spaced bars, the height of each bar being proportional to the duration of the corresponding unit (cf. Fig. 1.3).

The second way is to represent the duration horizontally. An example of this is the representation given in a TextGrid (as in Fig. 1.1), where the horizontal length of a given interval is proportional to the duration of that interval.

Gustafson (1987) suggests that a more appropriate representation would be to use both the horizontal and the vertical dimensions. He proposed representing the temporal dimension of an utterance as a sequence of *squares*, where both the horizontal and the vertical dimensions of the square are proportional to the duration of the unit. The justification for this double display of duration is that:

> the boxes display two different aspects of durations, 1) the dura/tion in itself (the vertical dimension), and 2) the timing of the beginning and end of each element whose duration is displayed (the horizontal dimension); another way of looking at this is to say that the boxes display durations as a function of true time. (pp. 107–108)

ProZed implements this idea with the possibility of using either squares or circles to represent the duration of the units. The use of squares follows the suggestion of Gustafson, while the use of different sized circles follows a practice common in handbooks teaching L2 intonation, see for example O'Connor and Arnold (1961) and Wells (2006), where intonation is often displayed using a representation of stressed and unstressed syllables by large and small circles respectively.

ProZed also offers two other possible displays which are not illustrated here. The first consists of using ovals instead of circles, where the width of the oval corresponds to the observed duration and the height of the oval corresponds to the predicted duration, as calculated from a table of mean phoneme durations. This is intended to give an impression, at the same time, of the observed duration and the degree of lengthening/shortening of that interval.

The other possibility is to use rectangles, where the vertical dimension of the rectangle corresponds to the pitch range of the unit so that the pitch excursion is always entirely contained in the rectangle.

1.5.2 Displaying Melody

When displaying the melody of an utterance, it is desirable to abstract away, on the one hand, from the microprosodic effects of the individual speech sounds, particularly the consonants, and, on the other hand, from the effect of the speaker's specific pitch range. This is particularly important, of course, if we wish to compare the melody of utterances produced by different speakers.

The first of these effects, the microprosodic effect of consonants, is to a large extent neutralised by modelling the f_0 curve with the Momel algorithm. For the second effect, that of the speaker's specific pitch range, we implement the findings of De Looze (2010) and De Looze and Hirst (2014) where it is claimed that the *OMe* (octave-median) scale is a natural scale for the analysis and display of speech melody.

The OMe scale is obtained by representing individual pitch points as the distance in octaves from the speaker's median pitch. This corresponds to the formula:

$$ome = log_2\left(\frac{f0}{median}\right) \tag{1.3}$$

In De Looze and Hirst (2014), it is claimed that a range of one octave centered on the speaker's median pitch is a good approximation to the limits of the speaker's unemphatic pitch range. While speakers will often use a pitch range that is wider than this (in particular for higher pitch values), when they do so, this is generally perceived as emphatic speech.

ProZed, then, displays the Momel pitch-curve on the OMe scale, as a continuous curve with discontinuities occurring only during silent pauses. The dimension of the pitch range is materialised by horizontal lines representing the median pitch and the limits corresponding to $\pm$ 0.5 octaves with respect to this median, i.e. the central octave of the speaker's voice centred on the median pitch. The absolute values of these three reference lines are given both in Hertz and in musical notes, assuming standard concert pitch where A4 (the A above middle C) corresponds to 440 Hz. This makes the melody display meaningful, for example, for linguists who have no formal training in acoustic phonetics, since the musical notes can easily be produced on a piano or any other instrument.

The pitch curve is overlaid on the representation (by circles, etc) of the duration of the selected units. The centre of the circle corresponds to the mean pitch of the unit.

An example of the ProZed display using this scale is given in Figs. 1.4 and 1.5 where the first figure corresponds to a male speaker and the second to a female

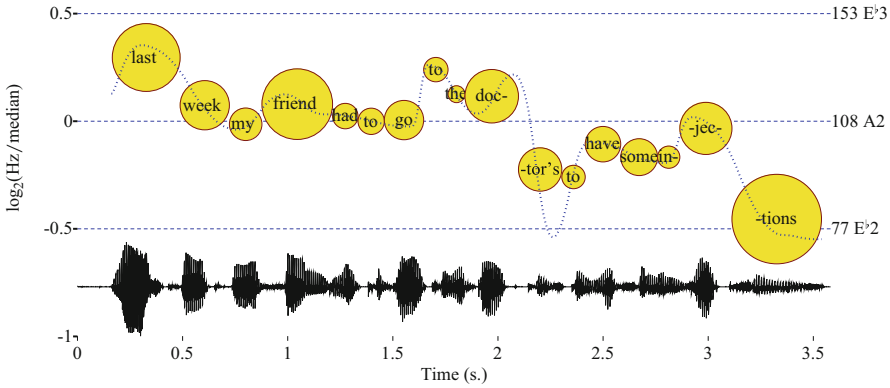


Fig. 1.4 ProZed prosody display for the sentence "Last week my friend had to go to the doctor's". The sizes of the circles represent the durations of the syllables, the *blue line* represents the Momel modelling of the f0 curve displayed using the OMe scale

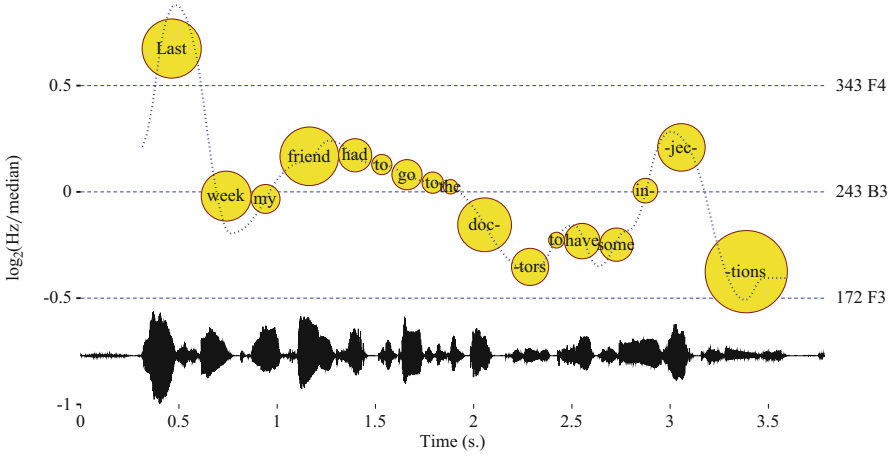


Fig. 1.5 ProZed prosody display for the sentence "Last week my friend had to go to the doctor's to have some injections". The sizes of the circles represent the durations of the syllables, the *blue line* represents the Momel modelling of the f0 curve displayed using the OMe scale

speaker. The displays were obtained automatically from the Sound and the TextGrid obtained via the SPPAS program, which originally contained interval tiers corresponding to words and phonemes. The syllable tier was obtained automatically by the function *Create syllable tier* . . . described above.

As can be seen, the female speaker's voice rises higher than the top reference line which can be taken as an indication that this is a more emphatic rendering of the sentence than that of the male speaker.

This display can also be useful to illustrate the difference of prosody across languages. Figure 1.6 illustrates a reading of the French version of the sentence in Figs. 1.4 and 1.5.

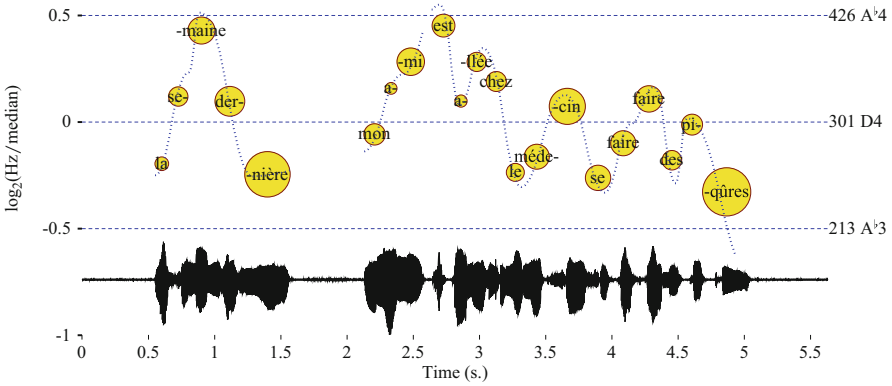


Fig. 1.6 ProZed prosody display for the sentence "La semaine derrière, mon amid set allée chez le médecin se faire faire des piqûres". The sizes of the circles represent the durations of the syllables, the *blue line* represents the Momel modelling of the f0 curve displayed using the OMe scale

The distinctive rising patterns that are characteristic of French intonation (Di Cristo 1998) are clearly apparent from this display.

Here, the speaker's voice is globally higher that that of Fig. 1.5, whereas the maximum peaks for the two speakers are practically at the same pitch. The display using the OMe scale shows, however, that the French female speaker's reading is in fact less emphatic than that of the English female speaker.

1.6 Conclusions

The ProZed plugin is designed as a tool to enable linguists to manipulate the rhythmic and tonal properties of an utterance by means of a symbolic representation, in order to evaluate the appropriateness of different phonological models of pitch and rhythm. It can be used either for immediate interactive experimentation with prosodic annotation, or to generate synthetic stimuli for more formal perceptual experiments. It also allows the user to automatically produce an intuitively easily interpretable display of the prosody of an utterance which can be compared to that of other speakers.

The plugin is freely downloadable from the *Speech and Language Data Repository*:

http://sldr.org/sldr000778/en.

This software, in conjunction with the algorithms for the automatic analysis of speech prosody described in Bigi and Hirst (2012, 2013), Hirst (2011) and Hirst and Auran (2005), aim to provide a complete analysis-by-synthesis environment, which, I have argued, is crucial for the development and testing of empirical models of speech prosody.

References

Bigi, Brigitte, and Daniel Hirst. 2012. SPPAS: A tool for the automatic analysis of speech prosody. *6th International Conference on Speech Prosody*, Shanghai, PRC..

Bigi, Brigitte, and Daniel Hirst. 2013, August. What's new in SPPAS 1.5? In *Proceedings of the International Workshop on Tools and Resources for the Analysis of Speech Prosody*, 62–65. Aix-en-Provence, France.

Boersma, Paul, and David Weenink. 1992–2010. Praat: A system for doing phonetics by computer. http://www.praat.org.

Bouzon, Caroline, and Daniel Hirst. 2004, March 23–26. Isochrony and prosodic structure in British English. *Proceedings of the Second International Conference on Speech Prosody 2004*, Nara.

De Looze, Céline. 2010, January. Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais. PhD thesis, Université de Provence, Aix-en-Provence, France.

De Looze, Céline, and Daniel Hirst. 2014, May. The OMe (Octave-Median) scale: A natural scale for speech prosody. In *Proceedings of the 7th International Conference on Speech Prosody (SP7),* ed. N. Campbell, D. Gibbon, and D. J. Hirst. Trinity College: Dublin.

Di Cristo, Albert. 1998. Intonation in French. In *Intonation Systems. A Survey of Twenty Languages*, ed. D.J. Hirst and A. Di Cristo, 195–218. Cambridge: Cambridge University Press.

Gustafson, Kjell. 1987. A new method for displaying speech rhythm, with illustrations from some nordic languages. In *Nordic Prosody IV. Papers from a Symposium*, 105–114, Odense, Odense University Press.

Hirst, Daniel. 1998. Intonation in British English. In *Intonation systems. A survey of twenty languages*, ed. D. J. Hirst and A. Di Cristo, 56–7. Cambridge: Cambridge University Press.

Hirst, Daniel. 1999, September. The symbolic coding of duration and alignment. An extension to the INTSINT system. *Proceedings Eurospeech '99*. Budapest.

Hirst, Daniel. 2005. Form and function in the representation of speech prosody. In *Quantitative prosody modeling for natural speech description and generation*, ed. K. Hirose, D. J.Hirst, and Y. Sagisaka, 334–347 (=*Speech Communication* 46 (3–4).

Hirst, Daniel. 2007. A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. In *Proceedings of the 16th International Congress of Phonetic Sciences*, 1233–1236. Saarbrüken, Germany.

Hirst, Daniel. 2009. The rhythm of text and the rhythm of utterances: From metrics to models. *Proceedings of Interspeech 2009*, 1519–1523. Brighton.

Hirst, Daniel. 2011. The analysis by synthesis of speech melody: From data to models. *Journal of Speech Sciences* 1 (1): 55–83.

Hirst, Daniel. 2012. Looking for empirical evidence for prosodic structure. In *Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem. = Speech and language technology*. vol. 14/15, ed. D. Gibbon, D. J. Hirst and N. Campbell, 23–33, Poznan: Polish Phonetic Association.

Hirst, Daniel. 2014, September. Automatic analysis and display of speech prosody with ProZed. *Interspeech*. Singapore (submitted)

Hirst, Daniel, and Robert Espesser. 1993. Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix* 15:71–85.

Hirst, Daniel, and Cyril Auran. 2005. Analysis by synthesis of speech prosody. The ProZed environment. *Proceedings of Interspeech/Eurospeech*. Lisbon.

Jassem, Wiktor. 1952. Intonation of conversational English: (educated Southern British). Nakl. Wroclawskiego Tow. Naukowego; skl. gl.: Dom Ksiazki. The Speech and Language Data Repository. http://sldr.org/sldr000777/en.

Kahn, Daniel. 1980. Syllable-based generalizations in English phonology. PhD dissertation, MIT. 1976. Garland, New York.

O'Connor, J. D., and G. Arnold. 1961. *Intonation of colloquial English. A practical handbook*. London: Longmans.

Pierrehumbert, Janet B. 1980. The phonology and phonetics of English intonation. Ph.D. dissertation, Massachusetts Institute of Technology. (Published in 1987 by Indiana University Linguistics Club, Bloomington.)

Silverman, Kim, Mary E. Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet B. Pierrehumbert, Julia Hirschberg. 1992, October 12–16. ToBI: A standard for labelling English prosody. In *Proceedings, Second International Conference on Spoken Language Processing* 2. Banff. Canada. 867–870.

Wells, John. 2006. *English intonation: An introduction*. Cambridge: Cambridge University Press.

Chapter 2
Degrees of Freedom in Prosody Modeling

Yi Xu and Santitham Prom-on

Abstract Degrees of freedom (DOF) refer to the number of free parameters in a model that need to be independently controlled to generate the intended output. In this chapter, we discuss how DOF is a critical issue not only for computational modeling, but also for theoretical understanding of prosody. The relevance of DOF is examined from the perspective of the motor control of articulatory movements, the acquisition of speech production skills, and the communicative functions conveyed by prosody. In particular, we explore the issue of DOF in the temporal aspect of speech and show that, due to certain fundamental constraints in the execution of motor movements, there is likely minimal DOF in the relative timing of prosodic and segmental events at the level of articulatory control.

2.1 Introduction

The ability to model speech prosody with high accuracy has long been the dream of prosody research, both for practical applications such as speech synthesis and recognition and for theoretical understanding of prosody. A key issue in prosody modeling is degrees of freedom (henceforth interchangeable with DOF). DOF refers to the number of independent parameters in a model that needs to be estimated in order to generate the intended output. So far there has been little serious discussion of the issue of DOF in prosody modeling, especially in terms of its theoretical implications. Nevertheless, DOF is often implicitly considered, and it is generally believed that, other things being equal, the fewer degrees of freedom in a model the better. For example, in the framework of intonational phonology, also known as the AM theory or the Pierrehumbert model of intonation, it is assumed that, at least for nontonal languages like English, "sparse tonal specification is the key to combining accurate phonetic modeling with the expression of linguistic equivalence

Y. Xu (✉)
University College London, London, UK
e-mail: yi.xu@ucl.ac.uk

S. Prom-on
King Mongkut's University of Technology, Thonburi, Thailand

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_2

of intonation contours of markedly different lengths" (Arvaniti and Ladd 2009, p. 48). The implication of such sparse tonal representation is that there is no need to directly associate F_0 events with individual syllables or words, and for specifying F_0 contours between the sparsely distributed tones. This would mean high economy of representation. Sparse F_0 specifications are also assumed in various computational models (e.g., Fujisaki 1983; Taylor 2000; Hirst 2005).

A main feature of the sparse tonal specification is that prosodic representations are assigned directly to surface F_0 events such as peaks, valleys, elbows, and plateaus. As a result, each temporal location is assigned a single prosodic representation, and an entire utterance is assigned a single string of representations. This seems to be representationally highly economical, but it also means that factors that do not directly contribute to the major F_0 events may be left out, thus potentially missing certain critical degrees of freedom. Another consequence of sparse tonal specification is that, because major F_0 events do not need to be directly affiliated with specific syllables or even words, their timing relative to the segmental events has to be specified in modeling, and this means that temporal alignment constitutes one or more (depending on whether a single point or both onset and offset of the F_0 event need to be specified) degrees of freedom. Thus many trade-offs need to be considered when it comes to determining DOF in modeling.

In this chapter we take a systematic, though brief look at DOF in prosody modeling. We will demonstrate that DOF is not only a matter of how surface prosodic events can be economically represented. Rather, decisions on DOF of a model should be ecologically valid, i.e., with an eye on what human speakers do. We advocate for the position that every degree of freedom (DOF) needs to be independently justified rather than based only on adequacy of curve fitting. In the following discussion we will examine DOF from three critical aspects of speech: motor control of articulatory movements, acquisition of speech production skills, and communicative functions conveyed by prosody.

2.2 The Articulatory Perspective

Prosody, just like segments, is articulatorily generated, and the articulatory process imposes various constraints on the production of prosodic patterns. These constraints inevitably introduce complexity into surface prosody. As a result, if not properly understood, the surface prosodic patterns due to articulatory constraints may either unnecessarily increase the modeling DOF or hide important DOF. Take F_0 for example. We know that both local contours and global shapes of intonation are carried by voiced consonants and vowels. Because F_0 is frequently in movement, either up or down, the F_0 trajectory within a segment is often rising or falling, or with an even more complex shape. A critical question from an articulatory perspective is, how does a voiced segment get its F_0 trajectory with all the fine details? One possibility is that all F_0 contours are generated separately from the segmental string of speech, as assumed in many models and theories, either explicitly or implicitly, and especially

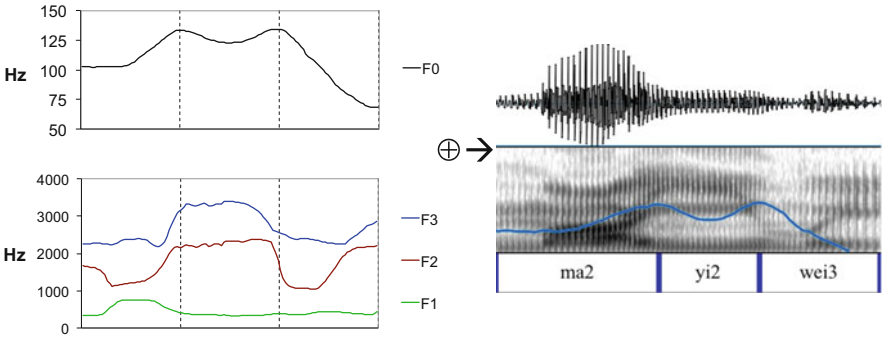


Fig. 2.1 *Left*: Continuous F_0 (*top*) and formant (*bottom*) tracks of the Mandarin utterance "(bi3) *ma2 yi2 wei3* (shan4)" [More hypocritical than Aunt Ma]. *Right*: Waveform, spectrogram and F_0 track of the same utterance. Raw data from Xu (2007)

in those that assume sparse tonal specifications (Pierrehumbert 1980; Taylor 2000; 't Hart et al. 1990). This scenario is illustrated in Fig. 2.1, where continuous F_0 and formant contours of a trisyllabic sequence are first separately generated with all the trajectory details (1a), and then merged together to form the final acoustic output consisting of both formant and F_0 trajectories. The critical yet rarely asked question is, is such an *articulate-and-merge* process biomechanically possible?

As is found in a number of studies, the change of F_0 takes a significant amount of time even if the speaker has used maximum speed of pitch change (Sundberg 1979; Xu and Sun 2002). As found in Xu and Sun (2002), it takes an average speaker around 100 ms to make an F_0 movement of even the smallest magnitude. In Fig. 2.1, for example, the seemingly slow F_0 fall in the first half of syllable 2 is largely due to a necessary transition from the high offset F_0 due to the preceding rising tone to the required low F_0 onset of the current rising tone, and such movements are likely executed at maximum speed of pitch change (Kuo et al. 2007; Xu and Sun 2002). In other words, these transitional movements are mainly due to articulatory inertia. Likewise, there is also evidence that many of the formant transitions in the bottom left panel of Fig. 2.1 are also due to articulatory inertia (Cheng and Xu 2013).

Given that F_0 and formant transitions are mostly due to inertia, and are therefore by-products of a biomechanical system, if the control signals (from the central nervous system (CNS)) sent to this system also contained all the inertia-based transitions, as shown on the left of Fig. 2.1, *the effect of inertia would be applied twice*. This consideration makes the *articulate-and-merge* account of speech production highly improbable. That is, it is unlikely that continuous surface F_0 contours are generated (with either a dense or sparse tonal specification) independently of the segmental events, and are then added to the segmental string during articulation.

But how, then, can F_0 contours and segmental strings be articulated together? One hypothesis, as proposed by Xu and Liu (2006), is that they are coproduced under the coordination of the syllable. That is, at the control level, each syllable is

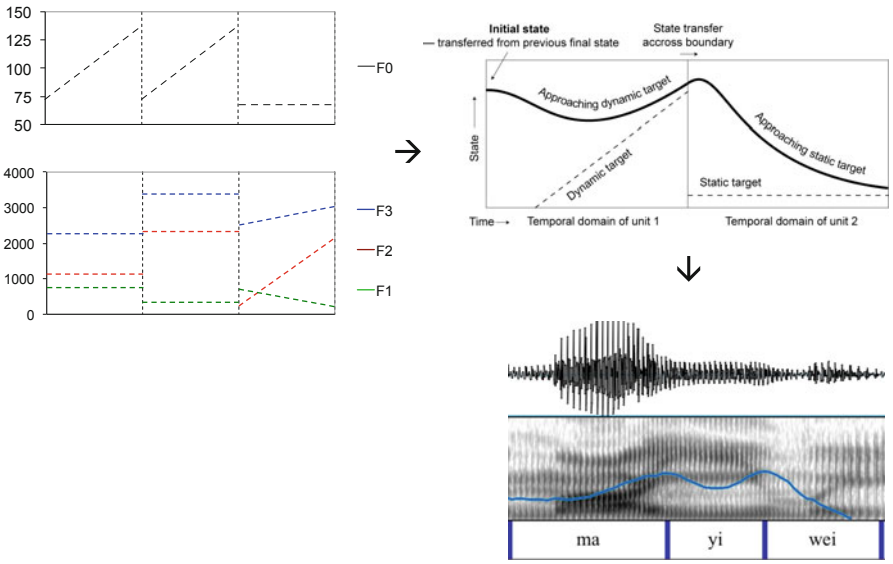


Fig. 2.2 *Left*: Hypothetical underlying pitch (*upper*) and formant (*lower*) targets for the Mandarin utterance shown in Fig. 2.1. *Top right*: The target approximation (TA) model (Xu and Wang 2001). *Bottom right*: Waveform, spectrogram and F_0 track of the same utterance. Raw data from Xu (2007)

specified with all the *underlying articulatory targets* associated with it, including segmental targets, pitch targets, and even phonation (i.e., voice quality) targets. This is illustrated in the top left block of Fig. 2.2 for pitch and formants. Here the formant patterns are representations of the corresponding vocal tract shapes, which are presumably the actual targets. The articulation process then concurrently approaches all the targets, respectively, through target approximation (top right). The target approximation process ultimately generates the continuous surface F_0 and formant trajectories (bottom right), which consist of mostly transitions toward the respective targets. Thus, every syllable, before its articulation, would have been assigned both segmental and suprasegmental targets as control signals for the articulatory system. And importantly, the effects of inertia are applied only once, during the final stage of articulatory execution.

Pitch target specification for each and every syllable may mean greater DOF than the sparse tonal specification models, of course, which is probably one of the reasons why it is not widely adopted in prosody modeling. But what may not have been apparent is that it actually reduces a particular type of DOF, namely, the F_0-segment alignment. For the sparse tonal specification models, because F_0 events are not attached to segments or syllables, the relative alignment of the two becomes a free variable, which constitutes at least one DOF (two if onset and offset of an F_0 event both have to be specified, as in the Fujisaki model). Thus for each tonal event, not only its height, but also its position relative to a segmental event, need to be specified. This complexity is further increased by the assumption of most of the sparse-tonal

specification models that the number of tonal and phrasal units is also a free variable and has to be learned or specified. For the Fujisaki model, for example, either human judgments have to be made based on visual inspection (Fujisaki et al. 2005), or filters of different frequencies are applied first to separately determine the number of phrase and accent commands, respectively (Mixdorff 2000). Even for cases where pitch specifications are obligatory for each syllable, e.g., in a tone language, there is a further question of whether there is freedom of micro adjustments of F_0-segment alignment. Allowance for micro alignment adjustments is assumed in a number of tonal models (Gao 2009; Gu et al. 2007; Shih 1986).

There has been accumulating evidence against free temporal alignment, however. The first line of evidence is the lack of micro alignment adjustment in the production of lexical tones. That is, the unidirectional F_0 movement toward each canonical tonal form starts at the syllable onset and ends at syllable offset (Xu 1999). Also the F_0-syllable alignment is not affected by whether the syllable has a coda consonant (Xu 1998) or whether the syllable–initial consonant is voiced or voiceless (Xu and Xu 2003). Furthermore, the F_0-syllable alignment does not change under time pressure, even if tonal undershoot occurs as a result (Xu 2004). The second line of evidence is from motor control research. A strong tendency has been found for related motor movements to be synchronized with each other, especially when the execution is at a high speed. This is observed in studies of finger tapping, finger oscillation, or even leg swinging by two people monitoring each other's movements (Kelso et al. 1981; Kelso 1984; Kelso et al. 1979; Mechsner et al. 2001; Schmidt et al. 1990). Even non-cyclic simple reaching movements conducted together are found to be fully synchronized with each other (Kelso et al. 1979).

The synchrony constraints could be further related to a general problem in motor control. That is, the high dimensionality of the human motor system (which is in fact true of animal motor systems in general) makes the control of any motor action extremely challenging, and this has been considered as one of the central problems in the motor control literature (Bernstein 1967; Latash 2012). An influential hypothesis is that the CNS is capable of functionally freezing degrees of freedom to simplify the task of motor control as well as motor learning (Bernstein 1967). The freezing of DOF is analogous to allowing the wheels of a car to rotate only around certain shared axes, under the control of a single steering wheel. Thus the movements of the wheels are fully synchronized, and their degrees of freedom merged. Note that such synchronization also freezes the relative timing of the related movements, hence eliminating it as a DOF. This suggests that the strong synchrony tendency found in many studies (Kelso et al. 1979; Kelso et al. 1981; Mechsner et al. 2001; Schmidt et al. 1990) could have been due to the huge benefits brought by the reduction of temporal degrees of freedom.

The benefit of reducing temporal DOF could also account for the tone-syllable synchrony in speech found in the studies discussed above. Since articulatory approximations of tonal and segmental targets are separate movements that need to be produced together, they are likely to be forced to synchronize with each other, just like in the cases of concurrent nonspeech motor movements. In fact, it is possible that

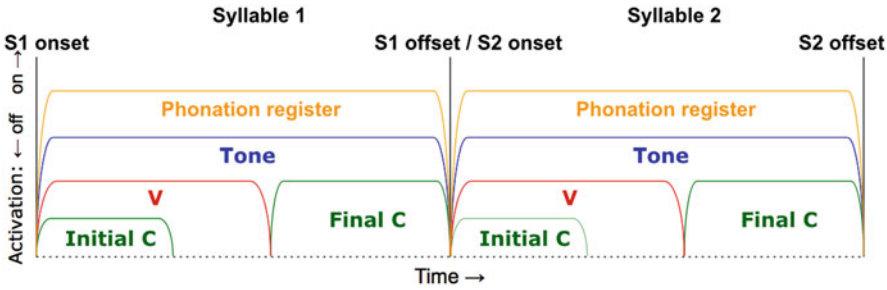


Fig. 2.3 The time structure model of the syllable (Xu and Liu 2006). The syllable is a *time structure* that assigns temporal intervals to consonants, vowels, tones, and phonation registers (each constituting a phone). The alignment of the temporal intervals follows three principles: **a** *Co-onset* of the initial consonant, the first vowel, the tone, and the phonation register at the beginning of the syllable; **b** *Sequential offset* of all noninitial segments, especially coda C; and **c** *Synchrony* of laryngeal units (tone and phonation register) with the entire syllable. In each case, the temporal interval of a phone is defined as the time period during which its target is approached

the syllable is a mechanism that has evolved to achieve synchrony of multiple articulatory activities, including segmental, tonal, and phonational target approximations. As hypothesized by the *time structure model of the syllable* (Xu and Liu 2006), the syllable is a temporal structure that controls the timing of all its components, including consonant, vowel, tone, and phonation register (Xu and Liu 2006), as shown in Fig. 2.3. The model posits that the production of each of these components is to articulatorily approach its ideal target, and the beginning of the syllable is the onset of the target approximation movements of most of the syllabic components, including the initial consonant, the first vowel, the lexical tone, and the phonation register (for languages that use it lexically). Likewise, the end of the syllable is the offset of all the remaining movements. In this model, therefore, there is always full synchrony at the onset and offset of the syllable. Within the syllable, there may be free timing at two places, the offset of the initial consonant, and the boundary between the nuclear vowel and the coda consonant. In the case of lexical tone, it is also possible to have two tonal targets within one syllable, as in the case of the L tone in Mandarin, which may consist of two consecutive targets when said in isolation. The boundary between the two targets is probably partially free, as it does not affect synchrony at the syllable edges.

2.3 The Learning Perspective

Despite the difficulty of motor control just discussed, a human child is able to acquire the ability to speak in the first few years of life, without formal instructions, and without direct observation of the articulators of skilled speakers other than the visible ones such as the jaw and lips. The only sure input that can inform the child of the articulatory details is the acoustics of speech utterances. How, then, can the child

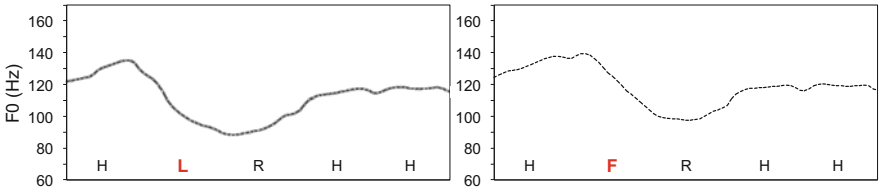


Fig. 2.4 Mean time-normalized F_0 contours of five-syllable Mandarin sentences with syllable two carrying the low tone on the *left* and the falling tone on the *right*. Data from Xu (1999)

learn to control her own articulators to produce speech in largely the same way as the model speakers? One possibility is that the acquisition is done through analysis-by-synthesis with acoustics as the input. The strategy is also known as distal learning (Jordon and Rumelhart 1992). To be able to do it, however, the child has to face the problem of multiplicity of DOF. As discussed earlier, adult speech contains extensive regions of transitions due to inertia. Given this fact, how can the child know which parts of the surface contour are mostly transitions, and which parts best reflect the targets? In Fig. 2.4, for example, how can a child tell that the Mandarin utterance on the left contains a low tone, while the one on the right contains a falling tone at roughly the same location? One solution for the child is to confine the exploration of each tonal target to the temporal domain of the syllable. That way, the task of finding the underlying target is relatively simple. This strategy is implemented in our computational modeling of tone as well as intonation (Liu et al. 2013; Prom-on et al. 2009; Xu and Prom-on 2014). Our general finding is that, when the syllable is used as the tone-learning domain, their underlying targets are easily and accurately extracted computationally, judging from the quality of synthesis with the extracted tonal targets in all these studies.

The ease of extracting tonal targets within the confine of the syllable, however, does not necessarily mean that it is the best strategy. In particular, what if the synchronization assumption is relaxed so that the learning process is given some freedom in finding the optimal target-syllable alignment? In the following we will report the results of a modeling experiment on the effect of flexibility of timing in pitch target learning.

2.3.1 Effect of Freedom of Tone–Syllable Alignment on Target Extraction—An Experiment

The goal of this experiment is to test if relaxing strict target-syllable synchrony improves or reduces F_0 modeling accuracy and efficiency with an articulatory-based model. If there is real timing freedom either in production or in learning, modeling

Fig. 2.5 Illustration of *onset timing shifts* used in the experiment and their impacts on the timing of adjacent syllables

accuracy should improve with increased timing flexibility during training. Also assuming that there is regularity in the target alignment in mature adults' production, the process should be able to learn the alignment pattern if given the opportunity.

2.3.1.1 Method

To allow for flexibility in target alignment, a revised version of PENTAtrainer1 (Xu and Prom-on 2010–2014) was written. The amount of timing freedom allowed was limited, however, as shown in Fig. 2.5. Only onset alignment relative to the original was made flexible. For each syllable, the onset of a pitch target is either set to be always at the syllable onset (fixed alignment), or given a 50 or 100 ms search range (flexible alignment). In the case of flexible alignments, if the hypothetic onset is earlier than the syllable onset, as shown in row two, the synthetic target approximation domain becomes longer than that of the syllable, and the preceding target domain is shortened; if the hypothetic onset is later than the syllable onset, as shown in row three, the synthetic target approximation domain is shortened, and the preceding domain is lengthened. Other, more complex adjustment patterns would also be possible, of course, but even from this simple design, we can already see that the adjustment of any particular alignment has various implications not only for the current target domain, but also for adjacent ones.

The training data are from Xu (1999), which have been used in Prom-on et al. (2009). The dataset consists of 3840 five-syllable utterances recorded by four male and four female Mandarin speakers. In each utterance, the first two and last two syllables are disyllabic words while the third syllable is a monosyllabic word. The first and last syllables in each sentence always have the H tone while the tones of the other syllables vary depending on the position: H, R, L, or F in the second syllable, H, R, or F in the third syllable, and H or L in the fourth syllable. In addition to tonal variations, each sentence has four focus conditions: no focus, initial focus, medial focus, and final focus. Thus, there are 96 total variations in tone and focus.

2.3.1.2 Results

Figure 2.6 displays bar graphs of RMSE and correlation values of resynthesis performed by the modified PENTAtrainer1 using the three onset time ranges. As can be seen, the 0 ms condition produced lower RMSE and higher correlation than

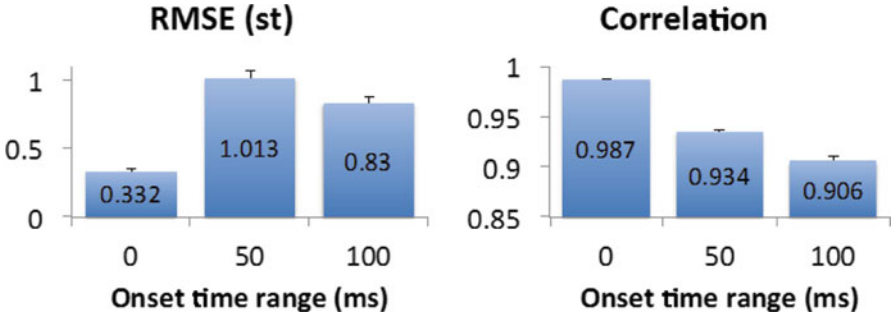


Fig. 2.6 Root mean square error (*RMSE*) and Pearson's correlation in resynthesis of F_0 contours of Mandarin tones in connected speech (data from Xu 1999) using targets obtained with three onset time ranges: 0, 50, and 100 ms

both the 50 and 100 ms conditions. Two-way repeated measures ANOVAs showed a highly significant effect of onset time range on both RMSE ($F\ [2,\ 7] = 387.4$, $p < 0.0001$) and correlation ($F\ [2,\ 7] = 320.1$, $p < 0.0001$). Bonferroni/Dunn post-hoc tests showed significant differences between all onset time ranges for both RMSE and correlation. More interestingly, on average, the learned alignments in the flexible conditions are still centered around syllable onset. The average deviation from the actual syllable boundaries is -2.3 ms in the 50 ms onset range condition and -5.1 ms in the 100 ms onset range condition (where the negative values mean that the optimized onset is earlier than the syllable boundary). A similarly close alignment to the early part of the syllable has also been found in Cantonese for the accent commands in the Fujisaki model, despite lack of modeling restrictions on the command–syllable alignment (Gu et al. 2007).

Figure 2.7 shows an example of curve fitting with 0 and 100 ms onset shift ranges. As can be seen, pitch targets learned with the 0 onset shift range produced a much tighter curve fitting than those learned with free timing. More importantly, we can see why the increased onset time range created problems. For the third syllable, the learned optimal onset alignment is later than the syllable onset. As a result, the temporal interval for realizing the preceding target is increased, given the alignment adjustment scheme shown in Fig. 2.6. As a result, the original optimal F_0 offset is no longer optimal, which leads to the sizeable discrepancy in Fig. 2.7b. Note that it is possible for us to modify the learning algorithm so that once an optimized onset alignment deviates from the syllable boundary, the preceding target is reoptimized. However, that would lead to a number of other issues. Should this reoptimization use fixed or flexible target onset? If the latter, shouldn't the further preceding target also be reoptimized? If so, the cycles will never end. Note also, that having a flexible search range at each target onset already increases the number of searches by many folds; having to reapply such searches to earlier targets would mean many more folds of increase. Most importantly, these issues are highly critical not just for modeling, but also for human learners, because they, too, have to find the optimal targets during their vocal learning.

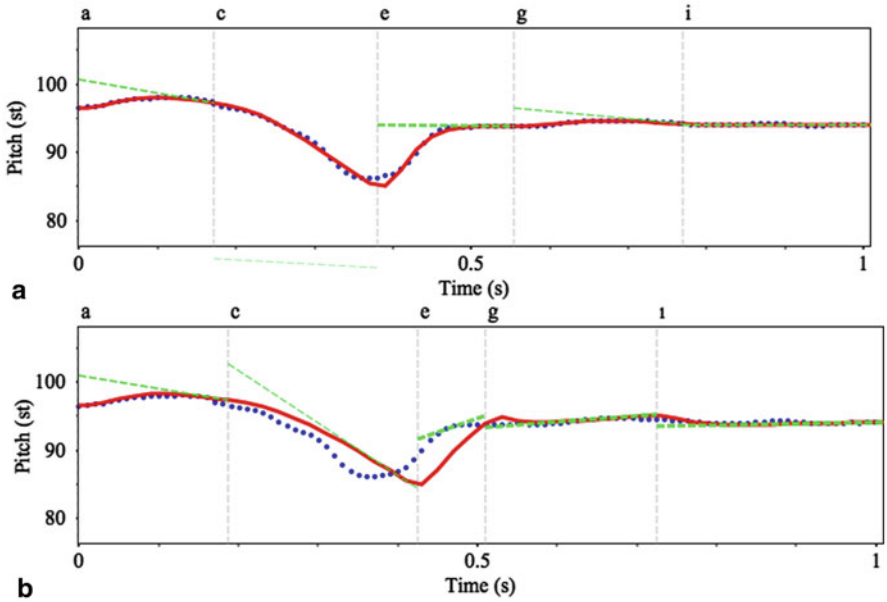


Fig. 2.7 Examples of curve fitting with targets learned with 0 ms onset timing shift (**a**), and 100 ms shift (**b**). The *blue dotted lines* are the original contours and the *red solid lines* the synthetic ones

In summary, the results of this simple modeling experiment demonstrate the benefit of fixing the temporal domain of the tonal target to that of the syllable in tone learning. Fully synchronizing the two temporal domains reduces DOF, simplifies the learning task, shortens the learning time, and also produces better learning results.

2.4 The Functional Perspective

Given that prosody is a means to convey communicative meanings (Bailly and Holm 2005; Bolinger 1989; Hirst 2005), the free parameters in a prosody model should be determined not only by knowledge of articulatory mechanisms, but also by consideration of the communicative functions that need to be encoded. Empirical research so far has established many functions that are conveyed by prosody, including lexical contrast, focus, sentence type (statement versus question), turn taking, boundary marking, etc. (Xu 2011). Each of these functions, therefore, needs to be encoded by specific parameters, and all these parameters would constitute separate degrees of freedom. In this respect, a long-standing debate over whether prosody models should be linear or superpositional is highly relevant. The linear approach, as represented by the autosegmental–metrical (AM) theory (Ladd 2008; Pierrehumbert 1980; Pierrehumbert and Beckman 1988), is based on the observation that speech intonations manifest clearly visible F_0 peaks, valleys, and plateaus. It is therefore assumed that

prosody consists of strings of discrete prosodic units, each exclusively occupying a temporal location. Such a linear approach naturally requires rather limited degrees of freedom.

The superpositional models, on the other hand, assume that surface F_0 contours are decomposable into *layers*, each consisting of a string of F_0 shapes, and the shapes of all the layers are added together to form surface F_0 contours (Bailly and Holms 2005; Fujisaki 1983; Thorsen 1980; van Santen et al. 2005). Take the Fujisaki model, for example, two layers are used to represent local shapes generated by accent commands, and global shapes generated by phrase commands, respectively. The output of the two layers is added together on a logarithmic scale to form a smooth global surface F_0 contour. Thus, superpositional models allow more than one unit to occur at any particular temporal location. This means more DOF than the linear models. In terms of economy of DOF, therefore, superpositional models may seem less optimal than linear models.

However, economy of DOF should not be the ultimate goal of prosody modeling. Instead, a model should be able to represent as many meanings conveyed by prosody as possible, while minimizing redundancy of representation. From this perspective, superpositional models with more than one layer of potential prosodic unit are aiming to provide sufficient DOF for encoding rich prosodic meanings (e.g., Bailly and Holm 2005), which makes them compare favorably to linear models. Meanwhile, however, as shown in our earlier discussion on articulatory mechanisms, each and every DOF should be articulatorily plausible. In this regard, superposition models still share with linear models the problematic *articulate-and-merge* assumption. Furthermore, from a modeling perspective, a superposition model has to first separate the surface contours into different layers, each corresponding to a particular communicative function. But this task is by no means easy. In Mixdorff (2000), filters of different frequencies were used to first separate surface F_0 contours into large global waves and small local ripples. Phrase commands are then sought from the global waves and accent commands from the local ripples. But the results are not satisfactory and manual intervention is often still needed (Mixdorff 2012).

The parallel encoding and target approximation model (PENTA) takes both articulatory mechanisms and communicative functions into consideration (Xu 2005). As shown in Fig. 2.8, the articulatory mechanism that generates surface F_0 is syllable-synchronized sequential target approximation, shown in the lower panel of the figure. In this mechanism, each syllable is assigned an underlying pitch target, and surface F_0 is the result of continuous articulatory approximation of successive pitch targets. This process is controlled by four free parameters: target height, target slope, target approximation rate, and duration. Each of these parameters therefore constitutes a DOF controllable by the encoding schemes. But there is no temporal DOF, as each target approximation is fully synchronized with the associated syllable. The encoding schemes each correspond to a specific communicative function, and the communicative functions are assumed to be parallel to (rather than dominating) each other, hence "parallel" in the name of the model. Like superposition, parallel encoding allows more than one prosodic element at any temporal location. But unlike superposition, in which streams of surface output are generated separately and then

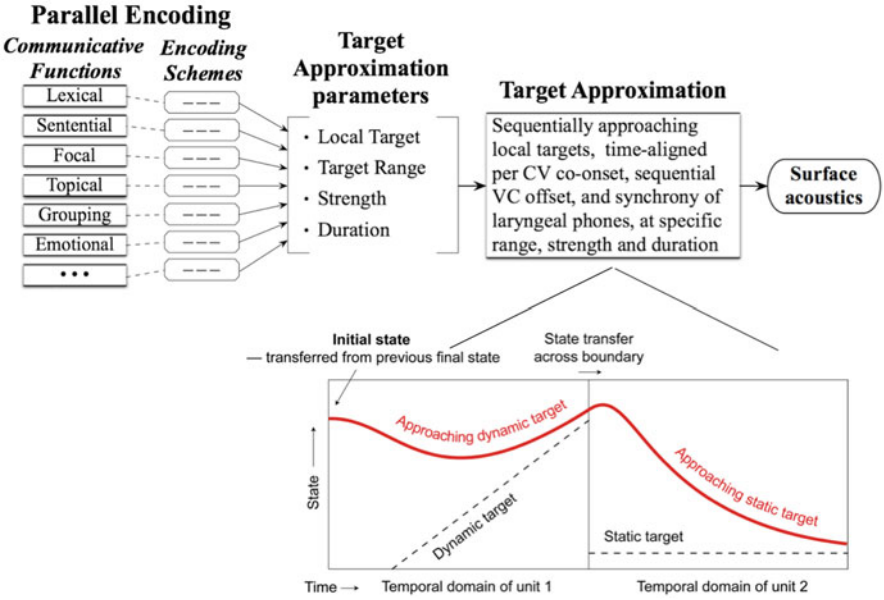


Fig. 2.8 *Upper panel*: A schematic sketch of the PENTA model (Xu 2005). *Lower panel*: The target approximation model of the articulation process (Xu and Wang 2001)

combined by summation, the encoding schemes in PENTA all influence a common sequence of underlying targets. The final sequence of targets carrying the combined influences from all communicative functions then generate surface output in a single articulatory process.

This single-target-sequence assumption of PENTA not only makes the generation of surface F_0 contours a rather straightforward process, it also makes it easy to account for a particular type of prosodic phenomenon, namely, target shift under the interaction of different communicative functions. Target shift is most vividly seen in the case of tone sandhi, whereby the underlying pitch target of a lexical tone is changed from the canonical one to a form that is very different in certain contextual conditions. In Mandarin, for example, the Low tone changes its target from low-level to rising when followed by another Low tone, and the resulting surface F_0 closely resembles that of the Rising tone. In American English, the pitch target of a word–final stressed syllable is a simple high in a pre-focus position; it changes to a steep fall when under focus in a statement, but to a steep rise in a question (Liu et al. 2013; Xu and Prom-on 2014; Xu and Xu 2005). In the PENTA approach, such a shift is modeled without taking any special steps, since for each functional combination a unique target has to be learned whether or not a sizable target shift occurs. Therefore, to the extent the PENTA approach is ecologically realistic, the way it models target shift may suggest why target shifts occur in languages in the first place. That is, because each multifunctional target needs to be learned as a whole, it is possible for

them to evolve properties that deviate significantly from their canonical forms. This issue is worth further exploration in future studies.

In terms of the specific target parameters in a model, the justification for each should come from empirical evidence of their usage in a specific function, and this principle is followed in the development of the PENTA model. For example, lambda, the rate of target approximation, could be fixed just like the time constant in the Fujisaki model (Fujisaki 1983). However, empirical research has provided evidence that the neutral tone in Mandarin and unstressed syllable in English approach their targets at much slower rates than normal tones and stressed syllables (Chen and Xu 2006; Xu and Xu 2005). Modeling studies have also shown that much lower lambda values are learned for the neutral tone and unstressed syllables (Prom-on et al. 2012; Xu and Prom-on 2014). Thus there is sufficient justification to keep lambda as a free parameter. Likewise, there is both analytical and modeling evidence for Rising and Falling tones to have unitary dynamic targets rather than successive static targets (Xu 1998; Xu and Wang 2001). Target duration is found to be used mainly in boundary marking, lexical contrast, and focus (Xu and Wang 2009; Wagner 2005). Target slope is found to be critical for tonal contrast.

In the PENTA approach, therefore, although there are only four free parameters at the target level, at the functional level, there can be as many degrees of freedom as required by the number of functions and the number of function-internal categories that need to be encoded. For English, for example, the communicative functions that need to be prosodically encoded include lexical stress, focus, sentence type, boundary marking, etc. (Liu et al. 2013; Xu and Xu 2005). For focus, it is necessary to control separate target attributes in pre-focus, on-focus, and post-focus regions. For sentence type, target attributes need to be controlled throughout the sentence, especially from the focused location onward (Liu et al. 2013). Also, focus, sentence type, and lexical stress have three-way interactions that determine the final attributes of all pitch targets in a sentence, which often result in rich diversities in local pitch targets (Liu et al. 2013).

2.5 Summary

We have examined the issue of degrees of freedom in speech prosody modeling from three different perspectives: the motor control of articulatory movements, the acquisition of speech production skills, and the communicative functions conveyed by prosody. From the articulatory perspective, we have shown that it is unlikely for the CNS to first generate separate continuous laryngeal and superlaryngeal trajectories and then merge them together when producing the whole utterance. Rather, it is more likely that individual syllables are assigned underlying laryngeal and supralaryngeal targets before their execution; and during articulation, multiple target approximation movements are executed in synchrony, under the time structure provided by the syllable. From the learning perspective, a new modeling experiment demonstrated the benefit of having minimum temporal DOF when learning pitch targets from

continuous speech, i.e., by confining the target search strictly within the temporal domain of the syllable. From the functional perspective, we have demonstrated how the PENTA approach allows multiple encoding schemes of prosodic functions to influence a common string of underlying targets, and then generate surface output in a single articulatory process of syllable synchronized sequential target approximation. We have further argued that DOF at the functional level should be based on the number of functions and number of function-internal categories that need to be encoded.

Overall, we have shown that DOF is a critical issue not only for the computational modeling of prosody, but also for the theoretical understanding of how speech prosody, and probably speech in general, can be learned in acquisition and articulated in skilled production.

References

Arvaniti, A., and D. R. Ladd. 2009. Greek wh-questions and the phonology of intonation. *Phonology* 26 (01): 43–74.

Bailly, G., and B. Holm. 2005. SFC: A trainable prosodic model. *Speech Communication* 46:348–364.

Bernstein, N. A. 1967. *The co-ordination and regulation of movements*. Oxford: Pergamon.

Bolinger, D. 1989. *Intonation and its uses—melody in grammar and discourse*. California: Stanford University Press.

Chen, Y., and Y. Xu. 2006. Production of weak elements in speech—evidence from f0 patterns of neutral tone in standard chinese. *Phonetica* 63:47–75.

Cheng, C., and Y. Xu 2013. Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *Journal of the Acoustical Society of America* 134 (6):4481–4495.

Fujisaki, H. 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In *The production of speech,* ed. P. F. MacNeilage, 39–55. New York: Springer-Verlag.

Fujisaki, H., C. Wang. Ohno, S., and Gu, W. 2005. Analysis and synthesis of fundamental frequency contours of standard Chinese using the command–response model. *Speech communication* 47:59–70.

Gao, M. 2009. Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese. *Chinese Journal of Phonetics* 2:43–50.

Gu, W., K. Hirose, and H. Fujisaki. 2007. Analysis of tones in Cantonese speech based on the command-response model. *Phonetica* 64:29–62.

Hirst, D. J. 2005. Form and function in the representation of speech prosody. *Speech Communication* 46:334–347.

Jordan, M. I., and D. E. Rumelhart. 1992. Forward models: Supervised learning with a distal teacher. *Cognitive Science* 16:316–354.

Kelso, J. A. S. 1984. Phase transitions and critical behavior in human bimanual coordination. *American Journal of Physiology: Regulatory, Integrative and Comparative* 246:R1000–R1004.

Kelso, J. A. S., D. L. Southard, and D. Goodman. 1979. On the nature of human interlimb coordination. *Science* 203:1029–1031.

Kelso, J. A. S., K. G. Holt, P. Rubin, and P. N. Kugler. 1981. Patterns of human interlimb coordination emerge from the properties of non-linear, limit cycle oscillatory processes: Theory and data. *Journal of Motor Behavior* 13:226–261.

Kuo, Y.-C., Y. Xu, and M. Yip. 2007. The phonetics and phonology of apparent cases of iterative tonal change in standard chinese. In *Tones and tunes Vol. 2: Experimental studies in word and sentence prosody*, ed. C. Gussenhoven and T. Riad, 211–237. Berlin: Mouton de Gruyter.

Ladd, D. R. 2008. *Intonational phonology*. Cambridge: Cambridge University Press.

Latash, M. L. 2012. Fundamentals of motor control. London: Academic Press.
Liu, F., Y. Xu, S, Prom-on, and A. C. L. Yu. 2013. Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling. *Journal of Speech Sciences* 3 (1): 85–140.
Mechsner, F., D. Kerzel, G. Knoblich, and W. Prinz. 2001. Perceptual basis of bimanual coordination. *Nature* 414:69–73.
Mixdorff, H. 2000. A novel approach to the fully automatic extraction of fujisaki model parameters. In *Proceedings of ICASSP 2000*, Istanbul, Turkey, 1281–1284.
Mixdorff, H. 2012. The application of the Fujisaki model in quantitative prosody research. In *Understanding prosody—The role of context, function, and communication,* ed. O. Niebuhr, 55–74. New York: Walter de Gruyter.
Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. PhD. Diss., MIT, Cambridge, MA. (Published in 1987 by Indiana University Linguistics Club, Bloomington).
Pierrehumbert, J., and M. Beckman. 1988. *Japanese tone structure*. Cambridge: The MIT Press.
Prom-on, S., Y. Xu, and B. Thipakorn. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America* 125:405–424.
Prom-on, S., F. Liu, and Y. Xu 2012. Post-low bouncing in Mandarin chinese: Acoustic analysis and computational modeling. *Journal of the Acoustical Society of America*. 132:421–432.
Schmidt, R. C., C. Carello, and M. T. Turvey. 1990. Phase transitions and critical fluctuations in the visual coordination of rhythmic movements between people. *Journal of Experimental Psychology: Human Perception and Performance* 16:227–247.
Shih, C. 1986. The prosodic domain of tone sandhi in Chinese. PhD. Diss., University of California, San Diego.
Sundberg, J. 1979. Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics* 7:71–79.
't Hart, J., R. Collier, and A. Cohen. 1990. *A perceptual study of intonation—An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.
Taylor, P. 2000. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America* 107:1697–1714.
Thorsen, N. G. 1980. A study of the perception of sentence intonation—Evidence from Danish. *Journal of the Acoustical Society of America* 67:1014–1030.
van Santen, J., A. Kain, E. Klabbers, and T. Mishra. 2005. Synthesis of prosody using multi-level unit sequences. *Speech Communication* 46:365–375.
Wagner, M. 2005. Prosody and recursion. PhD. Diss., Massachusetts Institute of Technology.
Xu, Y. 1998. Consistency of tone-syllable alignment across different syllable structures and speaking rates. *Phonetica* 55:179–203.
Xu, Y. 1999. Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics* 27:55–105.
Xu, Y. 2004. Understanding tone from the perspective of production and perception. *Language and Linguistics* 5:757–797.
Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Communication* 46:220–251.
Xu, Y. 2007. Speech as articulatory encoding of communicative functions. Proceedings of the 16th international congress of phonetic sciences, Saarbrucken, 25–30.
Xu, Y. 2011. Speech prosody: A methodological review. *Journal of Speech Sciences* 1:85–115.
Xu, Y., and F. Liu. 2006. Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics* 18:125–159.
Xu, Y., and S. Prom-on. 2010–2014. PENTAtrainer1.praat. http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer1/. Accessed 24 Nov 2013.
Xu, Y., and S. Prom-on 2014. Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication* 57:181–208.
Xu, Y., and X. Sun. 2002. Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America* 111:1399–1413.

Xu, Y., and M. Wang. 2009. Organizing syllables into groups—Evidence from F0 and duration patterns in Mandarin. *Journal of Phonetics* 37:502–520.

Xu, Y., and Q. E. Wang. 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication* 33:319–337.

Xu, C. X., and Y. Xu 2003. Effects of consonant aspiration on Mandarin tones. *Journal of the International Phonetic Association* 33:165–181.

Xu, Y., and C. X. Xu. 2005. Phonetic realization of focus in English declarative intonation. *Journal of Phonetics* 33:159–197.

Chapter 3
Extraction, Analysis and Synthesis of Fujisaki model Parameters

Hansjörg Mixdorff

Abstract The Fujisaki model provides a parsimonious representation of naturally observed fundamental frequency contours. It decomposes these contours into a constant base frequency onto which are superimposed global phrase and local accent gestures whose amplitudes and timings are given by underlying phrase and accent commands. These commands are estimated in a process of Analysis-by-Synthesis. The current chapter describes methods for extraction of Fujisaki model parameters, and then presents applications of the model parameters in several fields of research, especially speech synthesis.

3.1 Introduction

First versions of the Fujisaki model were published as early as 1969 (Fujisaki and Nagashima 1969), inspired by the works of Sven Öhman (1967). The current version dates back to 1984 and is displayed in Fig. 3.1 (Fujisaki and Hirose 1984). The model reproduces a given *F0* contour in the log *F0* domain by superimposing three components: A speaker–individual base frequency *Fb*, a phrase component, and an accent component (Eq. 3.1). The phrase component results from impulse responses to impulse-wise phrase commands typically associated with prosodic breaks. Phrase commands are described by their onset time *T0*, magnitude *Ap,* and time constant, *alpha*. The accent component results from rectangular accent commands associated with accented syllables. Accent commands are described by on- and offset times *T1* and *T2*, amplitude *Aa,* and time constant, beta.

$$\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi} G_{pi}(t - T_{0i}) + \sum_{j=1}^{J} A_{aj} \left\{ G_{aj}(t - T_{1j}) - G_{aj}(t - T_{2j}) \right\} \tag{3.1}$$

H. Mixdorff (✉)
Computer Science and Media, Beuth University Berlin, Berlin, Germany
e-mail: mixdorff@beuth-hochschule.de

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_3

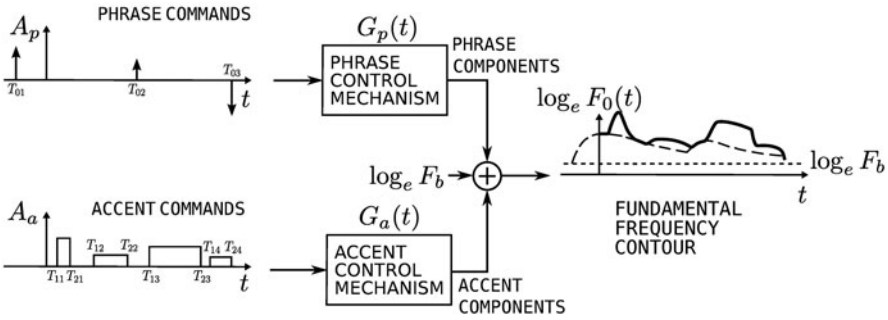


Fig. 3.1 Block diagram of the Fujisaki intonation model. (Adopted from Fujisaki and Hirose 1984, p. 235)

The i-th phrase component $G_{pi}(t)$ of the F0 model (Eq. 3.2) is generated by a second-order, critically-damped linear filter in response to the impulse-like phrase command, while the j-th accent component $G_{aj}(t)$ (Eq. 3.3) is generated by another second-order, critically-damped linear filter in response to a rectangular accent command:

$$G_{pi}(t) = \begin{cases} \alpha_i^2 t e^{-\alpha_i t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{3.2}$$

$$G_{aj}(t) = \begin{cases} \min\left[1 - (1 + \beta_j t)e^{-\beta_j t}, \gamma\right] & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{3.3}$$

Although the model was originally developed for Japanese, it has been successfully applied to many other languages including English and German. Fujisaki has attempted to provide an interpretation linking the model parameters to the activity of the crycothyroid muscle of the larynx, however, physiological data supporting this idea has yet to be provided (Fujisaki 1988).

3.2 Model Parameter Extraction

Several approaches have been developed for extracting Fujisaki model parameters from natural *F0* contours. These are usually based on interpolated and sometimes smoothed natural *F0* contours. By differentiating these contours, turning points can be determined, serving as anchoring points for phrase and accent commands which are subsequently optimized locally. Often, the phrase component is subtracted from the interpolated smooth contour at some point during the analysis and the remainder analyzed for determining the accent component (see, for instance, Pätzold 1991; Narusawa et al. 2000).

In this chapter, I will discuss and document in more detail the method developed by myself (Mixdorff 2000). This approach to my knowledge is the only one to date

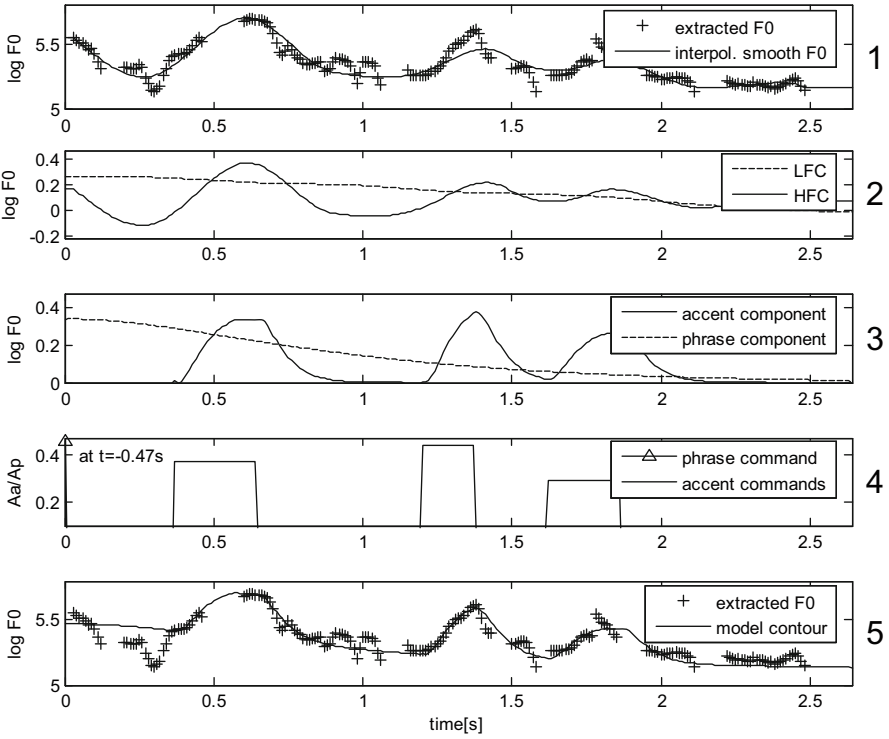


Fig. 3.2 Steps of Fujisaki model parameter extraction after Mixdorff. (2000)

that is publicly available as software (Mixdorff 2009). It is important to state that like most other methods, it exclusively operates on the observed *F0* data, hence disregarding linguistic knowledge of the underlying utterances. In contrast, a more recent modification to the method considers potentially accented words as candidate locations for accent commands (Torres et al. 2010). I will also present the approach by Kruschke (2001) for some of the original ideas it puts forward, and since it was evaluated in one of the few studies comparing Fujisaki model parameter extractors (Pfitzinger et al. 2009).

3.2.1 Mixdorff (2000)

Figure 3.1 illustrates the steps of analysis of this method on a short utterance of German uttered by a female speaker (Fig. 3.2).

The extracted F0 contour (panel 1, $+ + +$) is interpolated and smoothed using a cubic spline (panel 1, solid line). The resulting spline contour is passed through a high-pass filter with a stop frequency at 0.5 Hz, similar to Strom (1995). The

output of the high-pass (henceforth called "high frequency contour" or HFC, panel 2, solid line) is subtracted from the spline contour yielding a "low frequency contour" (LFC, panel 2, dashed line), containing the sum of phrase components and *Fb*. The latter is initially set to the overall minimum of the LFC. The HFC is searched for consecutive minima delimiting potential accent commands whose *Aa* is initialized to reach the maximum of F0 between the two minima. Since the onset of a new phrase command is characterized by a local minimum in the phrase component, the LFC is searched for local minima, applying a minimum distance threshold of 1 s between consecutive phrase commands. For initializing the magnitude value *Ap* assigned to each phrase command, the part of the LFC after the potential onset time *T0* is searched for the next local maximum. *Ap* is then calculated in proportion to the F0 at this point, considering contributions of preceding commands. The Analysis-by-Synthesis procedure is performed in three steps, optimizing the initial parameter set iteratively by applying a hill-climb search for reducing the overall mean-square-error in the *log F0* domain.

At the first step, phrase and accent components are optimized separately, using LFC and HFC, respectively, as the targets. Next, phrase component, accent component, and *Fb* are optimized jointly, with the spline contour as the target. In the final step, the parameters are finetuned by making use of a weighted representation of the extracted original F0 contour. The weighting factor applied is the product of degree of voicing and frame energy for every F0 value, which favors "reliable" portions of the contour. Panel 3 shows the resulting optimized phrase (dashed line) and accent components (solid line), and panel 4 the underlying impulse-wise phrase command (for simplicity indicated at $t = 0$, but actually located before the onset of the utterance at $t = -0.47$) and the accent commands. Panel 5 shows the superposition (solid line) of *Fb*, the phrase and accent component, as well as the original *F0* contour ($+++$).

Alpha and beta are inherently treated as constants for all phrase and accent commands, respectively. Whereas beta is set to a fixed value of 20/s, alpha can optionally be selected to differ from the default value of 2/s.

The base frequency, *Fb,* can be either determined automatically or set to a fixed value. Although *Fb* can be automatically optimized with respect to each individual utterance it appears more plausible to treat it as a speaker-inherent constant which should not vary greatly during the same recording session. Furthermore, depending on the sentence type, a speaker will not always reach the lower limit of his/her *F0* range. Figure 3.3 shows such an example. In the topmost panel, we see the example of a declarative sentence where the speaker reaches her *F0* floor at an *Fb* of 170 Hz. The middle panel shows the following utterance in the discourse which ends in a high, nonterminal mode. As a consequence *Fb* climbs to 196 Hz. Given the preceding utterance, this value is too high and will impede comparisons of *Ap*, for instance. Therefore the utterance is reanalyzed using a fixed *Fb* of 170 Hz, and the result is shown in the bottom panel.

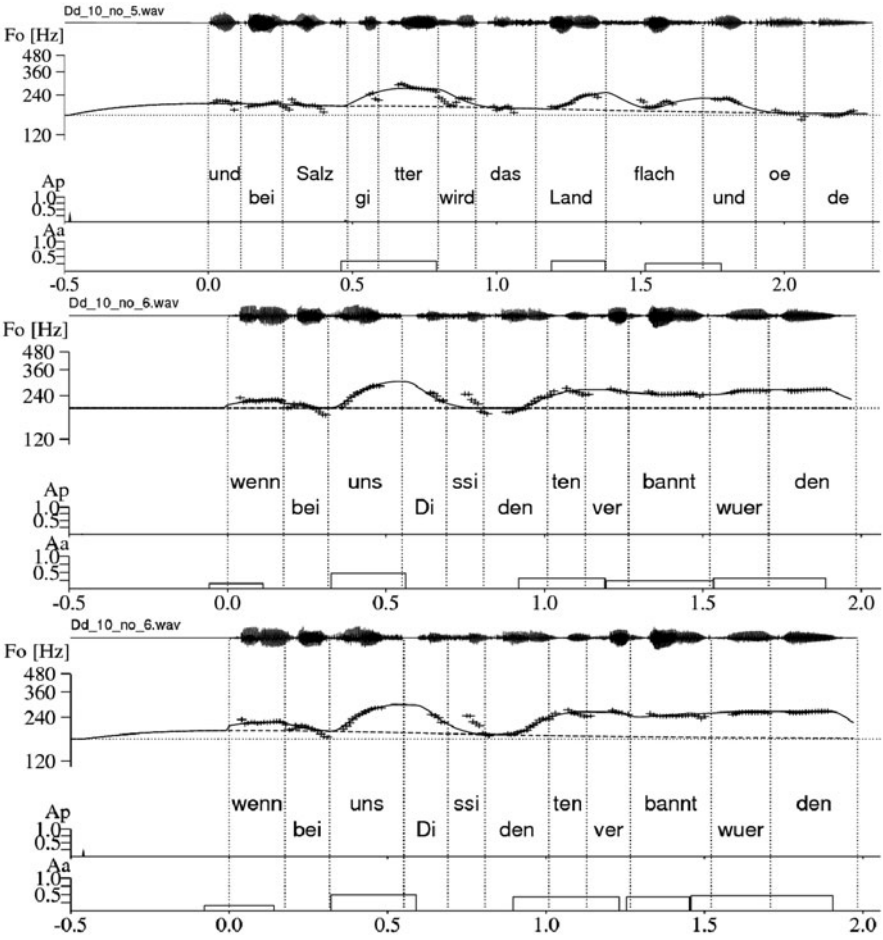


Fig. 3.3 Selecting *Fb* (see discussion in the text)

3.2.2 Kruschke (2001)

The algorithm developed by Kruschke differs from other approaches quite substantially in the way accent commands are located in the *F0* contour, as well as the optimization of the initial parameter set. Therefore it is discussed here in more detail.

After piecewise polynomial interpolation and smoothing the lowest, $F0 > 0$ is selected as a first approximation of *Fb*, and subtracted from the logarithmic *F0* contour. Then a Wavelet Transform using a Mexican hat wavelet is applied to the residual signal *F0res*(t). From the left to the right, the first marked maximum in the resulting scalogram is searched and picked as the maximum of a detected accent. The preceding marked minimum is selected as a starting value for *T1*. *T2*, the point where the smoothed accent command reaches 0 is set to the next *F0* minimum. The initial

values of the parameters *Aa*, beta, and *T2* are obtained in a pattern comparison, i.e., within specific ranges. *Aa*, beta, and *T2* are subsequently incremented to match the local *F0* contour around the accent. The parameter set with the smallest RMSE is taken as a first approximation of the parameters *Aa*, beta, and *T2*. Accent detection continues by searching the next marked maximum after T2. Then the resulting parameters are optimized in an Analysis-by-Synthesis procedure, which is controlled by an evolutionary strategy. An *F0* contour is generated from the accent commands and subtracted from the contour *F0res1(t)*. The resulting residual contour *F0res2(t)* is smoothed and used for detecting the phrase commands, again by Wavelet Transform using the Mexican hat wavelet. Each marked maximum in the scalogram is assigned to a phrase. The point in time 200 ms before a maximum at the beginning of the *F0* contour is chosen as a first approximation of *T0* and the lowest *F0* value between two extremes is selected as a starting value of T0. *Ap*, *alpha,* and *T0* are estimated by a procedure similar to that for accents. The algorithm continues until the parameters of the last phrase have been estimated. Finally, the parameters of all phrase and accent commands are optimized jointly.

3.2.3 Comparison of Model Parameter Extraction Methods

Probably the only study comparing several methods for extracting Fujisaki model extractors was published in 2009 (Pfitzinger et al. 2009). The automatic results from the approaches by Mixdorff, Kruschke, and two others (Pfitzinger and Schwarz, exclusively developed for this study) were compared against a manually post-processed version of Mixdorff on the IMS Radio Corpus (Rapp 1998) serving as the gold standard. Visual inspection confirmed that all methods captured the essential F0 movements adequately and yielded very similar smoothed versions of the original. Of all Fujisaki model extractors, the one by Kruschke obtained the smallest standard deviation and hence was the best overall fit. It was followed by the automatic and the manually corrected versions of Mixdorff. It was also found that 90 % of the deviations of three Fujisaki extractors were below 2.5 semitones. In order to interpret these results we have to bear in mind that the number of accent and phrase commands as well as the variability of the (theoretical) model constants *alpha, beta,* and *Fb* have a direct influence on the accuracy of approximation. The more commands are employed, the better the fitting of an observed *F0* contour becomes. As a consequence, however, the resulting parameters will become more and more difficult to interpret, since they will ultimately model microprosodic fluctuations and not accented syllables, boundary tones, or phrasal declination. Hence, moving from the automatic to the manually post-processed version of Mixdorff, the fitting accuracy decreases, because only those commands remain that can be motivated by accented syllables and prosodic phrase onsets. As an additional restriction, the manually post-processed version employs constant *Fb, alpha,* and *beta* for the same speaker, whereas *Fb* is adjusted in the other approaches depending on the particular sentence. In Kruschke,

Table 3.1 Some properties of approaches compared in Pfitzinger et al. (2009)

Method	Accents/s	Phrases/s	*Fb*	*Alpha beta*	RMSE	Algorithmic complexity
Pfitzinger	0.66	0.32	Const.	Const.	1.99	Low
Schwarz	1.10	0.56	Var.	Const.	1.61	Very high
Kruschke	1.43	0.46	Var.	Var.	1.23	Very high
Mixdorff	1.06	0.42	Const.	Const.	1.48	High

besides *Fb*, *alpha* and *beta* are also varied for each phrase and accent command, respectively, and therefore lead to a smaller error. Since, however, *Fb*, *Ap,* and *alpha*, as well as *Aa* and *beta* are related through the model formulation, *Ap* and *Aa* become more difficult to compare when *Fb*, *alpha,* and *beta* are treated as variables. The following table summarizes the main properties of the four extractors:

With respect to the evaluation of the approaches compared, we are aware that objective differences such as RMSE cannot replace psychoacoustic experiments regarding either the perceptual or—as a somewhat relaxed criterion—the functional–semantic equivalence of original, stylized, and modeled *F0* contours.

Obviously, the best way of ensuring that the Fujisaki model parameters reflect the underlying linguistic units and structures of an utterance will be by introducing such knowledge at the stage of parameter extraction (Table 3.1).

Applying these restrictions might lead to poorer approximations. However, from the standpoint of intonation research, we are not so interested in just noticeable differences between *F0* contours, but rather in the functional differences. Therefore, the ultimate goal should not be the closest approximation to automatically extracted *F0* values, which by nature is an unreliable reference, but rather the derivation of an interpretable set of parameters that can be related to the meaning conveyed by an utterance.

3.3 Linguistic Interpretation

Different from the parameters of the qTA model (Prom-on and Xu 2010) presented in Chap. 5, the extracted Fujisaki model parameters are not automatically anchored to underlying units or structures of the utterance being analyzed. Timing values of phrase and accent commands are yielded with respect to the time axis of the recording. Therefore, the alignment has to be performed post hoc, by considering the timing of the underlying utterance. In my own work, I decided to anchor phrase and accent commands to the syllables of the utterance analyzed. Other options include inter-perceptual-center groups or vowel nuclei. In my mind, the syllable as a basic rhythmic unit lends itself relatively well to the alignment of intonational events since definitions of stress and accent are usually formulated with respect to the syllables of a word. Furthermore, this approach facilitates comparisons between stress-timed and mora- or syllable-timed languages, respectively.

Therefore, onsets and offsets of accent commands are typically related to onset and offset times of the underlying syllables. This approach can be further refined if we take into account the importance of accented syllables for the prosodic structure of utterances in most European languages. In my approach for German, following the works of Isačenko and Schädlich (1964), a given *F0* contour is mainly described as a sequence of communicatively motivated tone switches, major transitions of the *F0* contour aligned with accented syllables. Tone switches can be thought of as the phonetic realization of phonologically distinct intonational elements, the so-called "intonemes". In the original formulation by Stock and Zacharias (1982), depending on their communicative function, three classes of intonemes are distinguished, namely the N↑intoneme ("non-terminal intoneme", signaling incompleteness and continuation, rising tone switch), I↓intoneme ("information intoneme" at declarative-final accents, falling tone switch, conveying information), and the C↑ intoneme ("contact intoneme" associated, for instance, with question-final accents, rising tone switch, establishing contact). Hence, intonemes in the original sense mainly distinguish sentence modality, although there exists a variant of the I↓ intoneme, I(E)↓ which denotes emphatic accentuation and occurs in contrastive, narrowly focused environments. Intonemes for reading style speech are predictable by applying a set of phonological rules to a string of text as to word accentability and accent group formation.

In a perception study (Mixdorff and Fujisaki 1995) employing synthetic stimuli of identical wording but varying F0 contours created with the Fujisaki model, it was shown that information intonemes are characterized by an accent command ending before or early in the accented syllable, creating a falling contour. N↑ intonemes were connected with rising tone switches to the midrange of the subject connected with an accent command beginning early in the accented syllable and a plateau-like continuation up to the phrase boundary, whereas C↑ intonemes required F0 transitions to span a total interval of more than 10 semitones and generally starting later in the accented syllable, although the *F0* interval was a more important factor than the precise alignment.

Hence, a rising tone switch is invariably connected with the onset of an accent command and a falling tone switch with an offset of an accent command. Statistical analysis of tone switch alignment indicated that rising tone switches are most closely linked to syllable onsets, whereas falling tone switches are more closely aligned with syllable offsets.

With respect to phrase command locations, the first command in an utterance will always be associated with the first syllable of that utterance. Due to the rise–fall characteristics of the phrase component, however, the phrase command usually occurs several hundred milliseconds before utterance onset and the maximum of the phrase component ideally coincides with the segmental onset (see example in Fig. 3.3). Subsequent phrase commands can be linked to following phrases, especially when these are preceded by short pauses. In many cases of smaller phrase commands, however, an exact assignment between onsets of prosodic phrases and phrase commands will be difficult and will have to be performed by rule. By default, alpha is set to 2/s, a value found appropriate for most speech materials in German. However, selecting a lower value of alpha around 1.0/s will produce fewer phrase commands

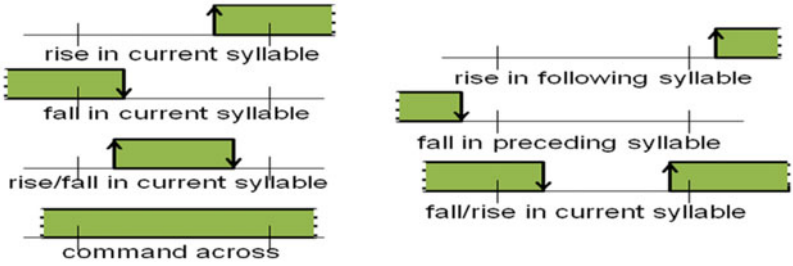


Fig. 3.4 Most important alignment options for linking accent commands with underlying syllables

coinciding with deep phrase boundaries, whereas values above 2.0/s will cause the phrase component to decay faster, requiring a greater number of phrase commands to model the declination line.

In recent studies, I employed knowledge about stressed syllables provided by a text-to-speech (TTS) frontend for automating the alignment between syllables and accent commands (Mixdorff 2012). This approach is based on the observation that falling or rising tone switches related to accented syllables do not necessarily occur during those syllables but before or after, respectively. Therefore the search for the best alignment option has to include the neighboring syllables. Once the locations of stressed syllables have been scanned for accent commands nearby, the rest of the commands are aligned with the closest syllable based on a criterion of maximum overlap. Figure 3.4 shows the most important alignment options taken into account.

3.4 Application of the Fujisaki Model in Speech Analysis and Perceptual Research

Following the rationale outlined in the preceding sections, over recent years, the Fujisaki model was applied in several areas of research, including speech analysis and perception of prosody. Experiments requiring the modification of the F0 benefit from the fact that the Fujisaki model creates a smooth and continuous F0 contour controlled by a relatively small set of parameters.

Mixdorff and Fujisaki (2000) compared German ToBI labels with Fujisaki model parameters on the IMS Radio Corpus (Rapp 1998). They found that tone labels were strongly correlated with accent commands, and the type of label (typically H*L and L*H) was clearly reflected by the onset and offset times of these accent commands. These main label types once again correspond to the I↓ and N↑ intonemes in Stock's formulation, respectively.

The contribution of *F0* to the perceived syllabic prominence was examined in Mixdorff and Widera (2001). It was shown that *F0* transitions only increased syllable prominence when they occurred in the vicinity of accented syllables. In these cases,

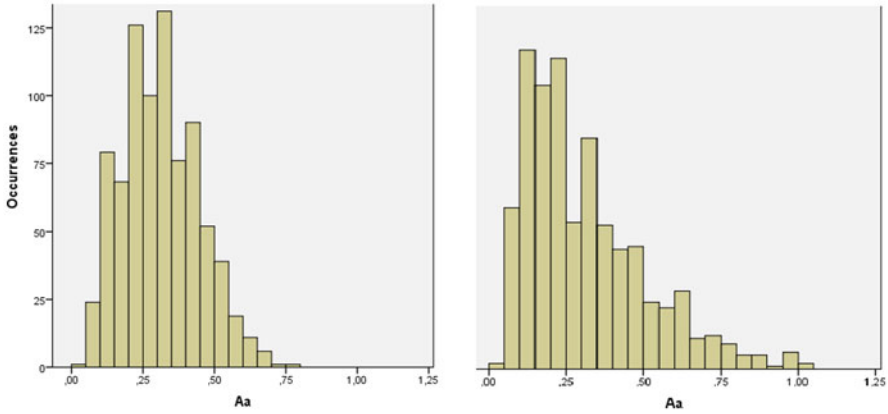


Fig. 3.5 Histograms of accent command amplitude *Aa* for German, left reading, right storytelling from (Mixdorff and Barbosa 2012)

the amplitudes *Aa* of the underlying accent commands correlated well with ratings of perceived prominence.

Pfitzinger's method for determining the perceived speech rate (1998) has been shown to be a suitable way of capturing the rhythmical properties of speech. It was consequently shown in Mixdorff and Pfitzinger (2005) that this framework can be successfully applied to the evaluation of prosodic differences between speaking styles. The more recent work by Mixdorff and Barbosa (2012) showed that differences in speaking style and language are reflected in the distributions and alignments of accent commands. Figure 3.5 displays histograms of accent command amplitudes *Aa* for reading style (left) and storytelling (right). Although in both cases the mean of *Aa* is.31, the standard deviation of.20 is much larger in storytelling ($N = 790$) than in readings ($s = .14$, $N = 824$). The strong left skew of the distribution indicates that in storytelling, more accents are weaker than in reading style, but some exhibit rather high amplitudes.

Prosodic L2 deviations were the focus of Mixdorff and Ingram (2009) and Mixdorff and Munro (2013). It was shown that F0 contours parameterized with the Fujisaki model can be readily transplanted between L1 and L2 utterances of the same sentence and employed for examining the perceptual impact of L2 prosody.

The impact of emotional expression on the F0 contour was examined in Amir et al. (2010) where ratings on Activation, Dominance, and Valence were related to the underlying Fujisaki model parameters whose predictive power was compared with that of raw F0 features such as F0 mean and range. Activation was mostly associated with stronger F0 resets (*mean Ap*) and stronger (*mean Aa*) and more frequent accents (expressed as *accent command distance*), whereas Valence was negatively correlated with the latter two factors, but to a lesser degree. A negative correlation was found between Dominance and *Fb*. Each of the raw F0 features showed much higher correlations with Activation and Valence, respectively, than the Fujisaki model-based

Table 3.2 Predictor variables for prosodic parameters in Mixdorff and Jokisch (2001)

Syllable level parameters	
Sum of mean durations of phones in syllable	Sum of mean durations of phones in onset
Sum of mean durations of phones in rhyme	Nuclear vowel schwa/non-schwa
Number of phones in onset	
Word level parameters	
Index of syllable in word	Part-of-speech of word (32 duration classes)
Number of syllables in word	Lexical word accent (0/1)
Features on the phrase level and above	
Syllables in preceding phrase	Boundary tone (0/1, before phrase boundaries)
Break index to the left (0–4)	Break index to the right (0–4)
Index of phrase in sentence	Index of sentence in paragraph
Start of phrase (0/1)	Start of paragraph (0/1)
Start of sentence (0/1)	Type of accent ("intoneme", three classes)
Syllable strength (0–2)	

parameters. However, the Fujisaki model parameters are based on a parsimonious decomposition of the same *F0* contours that the raw features were extracted from. Furthermore, the raw features were strongly correlated with each other (mean *F0* vs. SD of *F0*, $r = .551$; mean *F0* vs. F0 range, $r = .438$). When the four Fujisaki model-based parameters were introduced to a multiple linear regression (MLR) model of the judgments of Activation, all parameters added significant contributions and explained 31.1 % of the variance ($r = .558$), whereas a similar model based on mean *F0* and *F0* range (which give the best prediction provided that all factors be significant) explained 29.8 % ($r = .546$). Hence, the *F0* information was still captured by the model parameters, albeit on a higher level of abstraction and decomposed into the contributing factors at the utterance (*Fb*), phrase (*Ap*) and word levels (*Aa*).

3.5 Application in Speech Synthesis

Based on the considerations in the preceding sections, Mixdorff and Jokisch (2001) developed a model of German prosody anchoring prosodic features such as F0, duration, and intensity to the syllable as a basic unit of speech rhythm for prosody prediction in TTS synthesis. The *F0* contour is represented by syllable-aligned Fujisaki model parameters. Based on the analysis of one hour of speech from the IMS Radio Corpus, a neural network was trained predicting a set of syllable-aligned prosodic parameters from a vector of symbolic and quantitative parameters shown in Table 3.2. The structure of the feed-forward neural network is shown in Fig. 3.6.

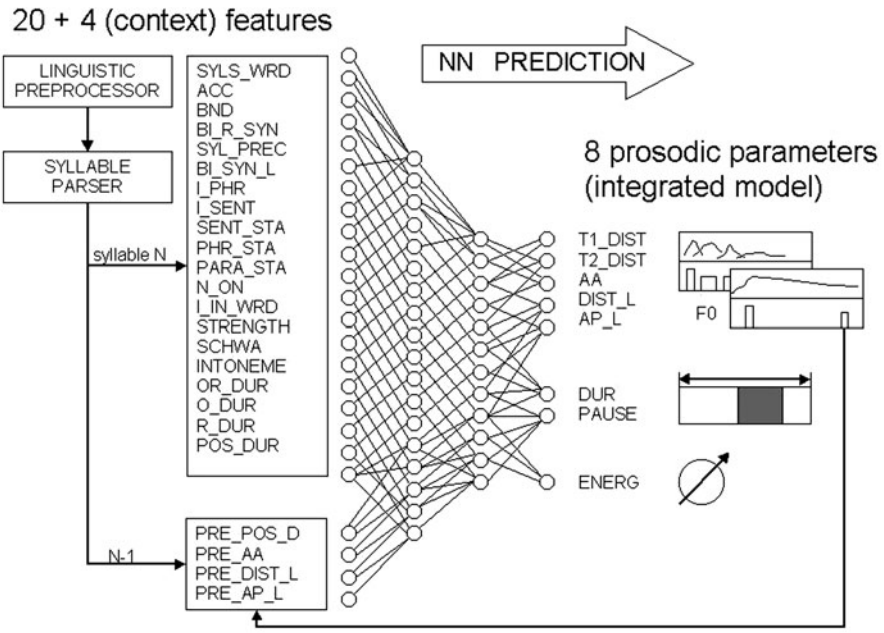


Fig. 3.6 Structure of feed-forward neural network for jointly predicting prosodic parameters for text-to-speech synthesis

References

Amir Noam et al. 2010. Unresolved anger: Prosodic analysis and classification of speech from a therapeutical setting. In Proceedings of Speech Prosody 2010, Chicago, USA.

Fujisaki, H. 1988. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In *Vocal Physiology: Voice production, mechanisms and functions,* ed. Fujimura, Osamu, 347–355. New York: Vocal Fold Physiology; 2.

Fujisaki, H., and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustic Society of Japan* 5 (4): 233–242.

Fujisaki, H., and S. Nagashima. 1969. A model for the synthesis of pitch contours of connected speech. In annual report of the engineering research institute, vol. 28, 53–60. Faculty of Engineering, University of Tokyo.

Isačenko, A. V., and H. J. Schädlich. 1964. *Untersuchungen über die deutsche Satzintonation*. Berlin: Akademie-Verlag.

Kruschke, H. 2001. Advances in the parameter extraction of a command-response intonation model. In Proceedings of IEEE International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS. Nashville Tennessee, USA.

Mixdorff, H. 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), vol. 3, 1281–1284. Istanbul.

Mixdorff, H. 2009. http://public.beuth-hochschule.de/∼mixdorff/thesis/fujisaki.html.

Mixdorff, H. 2012. *Form versus function—prosodic annotation and modeling go hand in hand, invited talk at speech prosody 2012*. China: Shanghai.

Mixdorff, H., and P. A. Barbosa. 2012. Alignment of intonational events in German and Brazilian Portuguese—a quantitative study. In Proceedings of Speech Prosody 2012. Shanghai, China.

Mixdorff, H., and H. Fujisaki. 1995. Production and perception of statement, question and non-terminal intonation in German. In Proceedings of the ICPhS '95, vol. 2, 410–413. Stockholm, Schweden.

Mixdorff, H., and H. Fujisaki. 2000. A quantitative description of German prosody offering symbolic labels as a by-product. In Proceedings of the ICSLP 2000, vol. 2, 98–101. Beijing, China.

Mixdorff, H., and J. Ingram. 2009. Prosodic analysis of foreign-accented English. In Proceedings of Interspeech 2009. Brighton, England.

Mixdorff, H., and O. Jokisch. 2001. Implementing and evaluating an integrated approach to modeling German prosody. In Proceedings of the 4th ISCA Workshop on Speech Synthesis, Perth Atholl Palace Hotel, Scotland.

Mixdorff, H., and M. Munro. 2013. Quantifying and evaluating the impact of prosodic differences of foreign-accented English. In Proceedings of Slate 2013. Grenoble, France.

Mixdorff, H., and H. R. Pfitzinger. 2005. *Analysing fundamental frequency contours and local speech rate in map task dialogs*, *in speech communication*, vol. 46, 310–325. Amsterdam: Elsevier.

Mixdorff, H., and C. Widera. 2001. Perceived prominence in terms of a linguistically motivated quantitative intonation model. In Proceedings of Eurospeech 2001, vol. 1, 403–406. Aalborg, Denmark.

Narusawa, S., H. Fujisaki, and H. S. Ohno. 2000. A method for automatic extraction of parameters of the fundamental frequency contour. In Proceedings of ICSLP 2000, vol. 1, 649–652. Beijing.

Öhman, S. E. G. 1967. Word and sentence intonation: A quantitative model. In STL-QPSR 2–3, 20–54. Royal Institute of Technology, Stockholm.

Pätzold, M. 1991. Nachbildung von Intonationskonturen mit dem Modell von Fujisaki. Implementierung des Algorithmus und erste Experimente mit ein- und zweiphrasigen Aussagesätzen, Master's Thesis, University of Bonn.

Pfitzinger, H. R. 1998. Local speech rate as a combination of syllable and phone rate. Proc. ICSLP 1998, vol. 3, 1087–1090. Sydney, Australia.

Pfitzinger, H., H. Mixdorff, and J. Schwarz. 2009. Comparison of Fujisaki-model extractors and F0 stylizers. In Proceedings of Interspeech 2009. Brighton, England.

Prom-on, S., and Y. Xu. 2010. The qTA toolkit for prosody: Learning underlying parameters of communicative functions through modeling. In Speech Prosody 2010. Chicago.

Rapp, S. 1998. Automatisierte Erstellung von Korpora für die Prosodieforschung, Arbeitspapiere (phonetikAIMS) 4(1), 1.167. Institut für Maschinelle Sprachverarbeitung, Lehrstuhl für Experimentelle Phonetik der Universität, Stuttgart.

Stock, E., and C. Zacharias. 1982. *Deutsche Satzintonation*. Leipzig: VEB Verlag Enzyklopädie.

Strom, V. 1995. Detection of accents, phrase boundaries and sentence modality in German with prosodic features. In Proceedings of EUROSPEECH '95, vol. 3, 2039–2041. Madrid.

Torres, H., H. Mixdorff, J. Gurlekian, and H. R. Pfitzinger. 2010. Two new estimation methods for a superpositional intonation model. Interspeech 2010. Makuhari, Japan.

Chapter 4
Probabilistic Modeling of Pitch Contours Toward Prosody Synthesis and Conversion

Hirokazu Kameoka

Abstract Since the voice fundamental frequency (F_0) contour is an important acoustic correlate of many prosodic constructs, modeling and analyzing F_0 contours can potentially be useful for many speech applications such as speech synthesis, speaker identification, speech conversion, and dialogue systems, in which prosodic information plays a significant role. In this chapter, we formulate a statistical model of F_0 contours by translating the "Fujisaki model," a well-founded mathematical model representing the control mechanism of vocal fold vibration, into a probabilistic model described as a discrete-time stochastic process. There are two motivations behind this formulation. One is to derive a general parameter estimation framework for the Fujisaki model, allowing for the introduction of powerful statistical methods. The other is to construct an automatically trainable version of the Fujisaki model so that it can be naturally incorporated into statistical speech synthesis and conversion frameworks.

4.1 Introduction

Prosody aids the listener in interpreting an utterance by grouping words into larger information units and drawing attention to specific words. It also plays an important role in conveying various types of nonlinguistic information such as the identity, intention, attitude, and mood of the speaker. Since the voice fundamental frequency (F_0) contour is an important acoustic correlate of prosody, modeling, and analyzing F_0 contours can potentially be useful for many speech applications such as speech synthesis, speaker identification, speech conversion, and dialogue systems, in which

H. Kameoka (✉)
Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: kameoka@hil.t.u-tokyo.ac.jp

NTT Communication Science Laboratories, Nippon Telegraph
and Telephone Corporation, 3-1 Morinosato Wakamiya, Atsugi, Kanagawa, Japan
e-mail: kameoka.hirokazu@lab.ntt.co.jp

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_4

prosody plays a significant role. It is also important to note that F_0 contours indicate intonation in pitch accent languages.

An F_0 contour consists of long-term pitch variations over the duration of prosodic units and short-term pitch variations in accented syllables. The former usually contribute in phrasing while the latter contribute in accentuation during an utterance. These two types of pitch variations can be interpreted as the manifestations of two independent movements by the thyroid cartilage. The Fujisaki model (Fujisaki 1988) is a physically founded model that describes an F_0 contour as the sum of these two contributions. The notable feature of the Fujisaki model is that it consists of linguistically, physiologically, and physically meaningful parameters (called phrase and accent commands) and is able to fit F_0 contours of real speech well when the values of these parameters are set appropriately. In addition, the validity of the Fujisaki model has been shown for many, typologically diverse languages.

In speech synthesis technology, one important challenge involves synthesizing an F_0 contour that is not only linguistically appropriate but also physically natural as if it is uttered by a human speaker. To make speech synthesizers generate natural-sounding F_0 contours, one promising approach would be to incorporate the Fujisaki model into the generative model of speech feature sequences so that all the parameters can be learned from a speech corpus in a unified manner. However, since the Fujisaki model does not take the form of a statistical (automatically trainable) model, incorporating the Fujisaki model into a statistical learning framework is not straightforward. In addition, estimating (learning) the Fujisaki model parameters from raw F_0 contour observations has been a difficult task. Several techniques for this task have already been developed (Fujisaki and Hirose 1984; Mixdorff 2000; Narusawa et al. 2002 among others), but so far with limited success.

To overcome this hurdle, we have been concerned with translating the Fujisaki model into a probabilistic model. With this formulation, we will shortly show that we can derive a general parameter estimation framework for the Fujisaki model, allowing for the introduction of powerful statistical methods. Furthermore, this formulation naturally allows us to incorporate the Fujisaki model into statistical speech synthesis and conversion frameworks.

4.2 Original Fujisaki Model

The Fujisaki model (Fujisaki 1988), shown in Fig. 4.1, assumes that an F_0 contour on a logarithmic scale, $y(t)$, where t is time, is the superposition of three components: a phrase component $x_{\mathrm{p}}(t)$, an accent component $x_{\mathrm{a}}(t)$, and a base value μ_{b}:

$$y(t) = x_{\mathrm{p}}(t) + x_{\mathrm{a}}(t) + \mu_b. \tag{4.1}$$

The phrase and accent components are considered to be associated with mutually independent types of movement of the thyroid cartilage with different degrees of freedom and muscular reaction times. The phrase component $x_{\mathrm{p}}(t)$ consists of the major-scale pitch variations over the duration of the prosodic units, and the accent component $x_{\mathrm{a}}(t)$ consists of the smaller scale pitch variations in accented syllables.

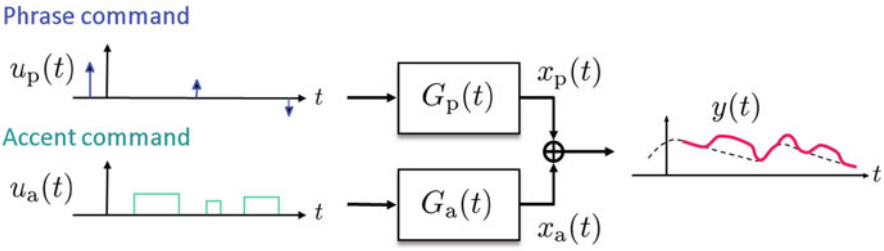


Fig. 4.1 A block diagram of the original Fujisaki model (Fujisaki 1988)

These two components are modeled as the outputs of second-order critically damped filters, one being excited with a command function $u_p(t)$ consisting of Dirac deltas (phrase commands), and the other with $u_a(t)$ consisting of rectangular pulses (accent commands):

$$x_p(t) = G_p(t) * u_p(t), \tag{4.2}$$

$$G_p(t) = \begin{cases} \alpha^2 t e^{-\alpha t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \tag{4.3}$$

$$x_a(t) = G_a(t) * u_a(t), \tag{4.4}$$

$$G_a(t) = \begin{cases} \beta^2 t e^{-\beta t} & (t \geq 0) \\ 0 & (t < 0) \end{cases}, \tag{4.5}$$

where $*$ denotes convolution over time. μ_b is a constant value related to the lower bound of the speaker's F_0, below which no regular vocal fold vibration can be maintained. α and β are natural angular frequencies of the two second-order systems, which are known to be almost constant within an utterance as well as across utterances for a particular speaker. It has been shown that $\alpha = 3$ rad/s and $\beta = 20$ rad/s can be used as default values.

4.3 Discretized Fujisaki Model

In this section, we apply a backward difference s-to-z transform to the phrase and accent control mechanisms described as continuous-time linear systems in order to obtain a discrete-time version of the Fujisaki model. The reason for the discretization will be made apparent later. The transfer function of the impulse response of the phrase control mechanism is given in the Laplace transform domain as

$$\mathcal{G}_p(s) = \mathcal{L}[G_p(t)] = \frac{\alpha^2}{(s+\alpha)^2}. \tag{4.6}$$

The backward difference transform approximates the time differential operator s by the backward difference operator in the z-domain such that

$$s \simeq \frac{1 - z^{-1}}{t_0}, \tag{4.7}$$

where t_0 is the sampling period of the discrete-time representation. By undertaking this transform, the transfer function of the inverse system $\mathcal{G}_{\mathrm{p}}^{-1}(s)$ can be written in the z-domain as

$$\mathcal{G}_{\mathrm{p}}^{-1}(z) = f_2^{(\mathrm{p})} z^{-2} + f_1^{(\mathrm{p})} z^{-1} + f_0^{(\mathrm{p})}, \tag{4.8}$$

where

$$f_2^{(\mathrm{p})} = (\psi - 1)^2, \tag{4.9}$$

$$f_1^{(\mathrm{p})} = -2\psi(\psi - 1), \tag{4.10}$$

$$f_0^{(\mathrm{p})} = \psi^2, \tag{4.11}$$

$$\psi = 1 + 1/(\alpha t_0). \tag{4.12}$$

Let us use $u_{\mathrm{p}}[k]$ and $x_{\mathrm{p}}[k]$ to denote the discrete-time version of the phrase command function and phrase component, respectively, where k is the discrete-time index. $x_{\mathrm{p}}[k]$ can thus be regarded as the output of a constrained all-pole system whose characteristics are governed by a single parameter ψ (or α), such that

$$u_{\mathrm{p}}[k] = f_0^{(\mathrm{p})} x_{\mathrm{p}}[k] + f_1^{(\mathrm{p})} x_{\mathrm{p}}[k-1] + f_2^{(\mathrm{p})} x_{\mathrm{p}}[k-2]. \tag{4.13}$$

In the same way, the relationship between the accent command function $u_{\mathrm{a}}[k]$ and the accent component $x_{\mathrm{a}}[k]$ is described as

$$u_{\mathrm{a}}[k] = f_0^{(\mathrm{a})} x_{\mathrm{a}}[k] + f_1^{(\mathrm{a})} x_{\mathrm{a}}[k-1] + f_2^{(\mathrm{a})} x_{\mathrm{a}}[k-2], \tag{4.14}$$

where

$$f_2^{(\mathrm{a})} = (\varphi - 1)^2, \tag{4.15}$$

$$f_1^{(\mathrm{a})} = -2\varphi(\varphi - 1), \tag{4.16}$$

$$f_0^{(\mathrm{a})} = \varphi^2, \tag{4.17}$$

$$\varphi = 1 + 1/(\beta t_0). \tag{4.18}$$

As described in (4.1), an F_0 contour $y[k]$ is expressed as $x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k] + \mu_{\mathrm{b}}$.

4.4 Generative Model of Speech F_0 Contours

Here, we model the probabilistic generative process of a speech F_0 contour based on the discrete-time verison of the Fujisaki model.

4.4.1 Modeling Phrase and Accent Command Pair

We first describe the process for generating the phrase and accent command functions, $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$. In the original Fujisaki model, it is required that they satisfy the following requirements:

1. Phrase commands are a set of impulses and accent commands are a set of stepwise functions.
2. A phrase command occurs at the start of an utterance or after the offset of an accent command in the preceding phrase, and is followed by the onset of the next accent command. This means that a phrase command will not occur while an accent command is being activated.
3. The onset of an accent command is followed by its offset. This means that neighboring accent commands will not overlap each other.

According to assumption 2, $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ are reciprocally constrained and so they should not simply be modeled separately. One challenge in the estimation of the Fujisaki model parameters has been how to deal with the optimization problem under these constraints. As a convenient way of incorporating these requirements into the command functions, we propose modeling the $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ pair using a hidden Markov model (HMM).

We denote the $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ pair by $\boldsymbol{o}[k]$ and assume that it is normally distributed:

$$\boldsymbol{o}[k] \sim \mathcal{N}\left(\boldsymbol{o}[k]; \boldsymbol{\rho}[k], \boldsymbol{\Upsilon}\right), \tag{4.19}$$

where

$$\boldsymbol{o}[k] = \begin{bmatrix} u_{\mathrm{p}}[k] \\ u_{\mathrm{a}}[k] \end{bmatrix}, \ \boldsymbol{\rho}[k] = \begin{bmatrix} \mu_{\mathrm{p}}[k] \\ \mu_{\mathrm{a}}[k] \end{bmatrix}, \ \boldsymbol{\Upsilon} = \begin{bmatrix} \sigma_{\mathrm{p}}^2 & 0 \\ 0 & \sigma_{\mathrm{a}}^2 \end{bmatrix}.$$

Equation (4.19) can be viewed as an HMM in which the output distribution of each state is a Gaussian distribution and the mean vector $\boldsymbol{\rho}[k]$ evolves in time as a result of the state transition. The mean vector $\boldsymbol{\rho}[k]$ consists of the means of the phrase and accent command functions, $\mu_{\mathrm{p}}[k]$ and $\mu_{\mathrm{a}}[k]$. This interpretation allows us to incorporate assumptions 1–3 into $\mu_{\mathrm{p}}[k]$ and $\mu_{\mathrm{a}}[k]$ by constraining the path of the state transitions as illustrated in Fig. 4.2.

The present HMM consists of $N + 3$ distinct states, r_0, p_0, r_1, and $\mathrm{a}_0, \ldots, \mathrm{a}_{N-1}$. In state r_0, $\mu_{\mathrm{p}}[k]$ and $\mu_{\mathrm{a}}[k]$ are both constrained to be zero. In state p_0, $\mu_{\mathrm{p}}[k]$ can take a nonzero value, $C^{(\mathrm{p})}[k]$, whereas $\mu_{\mathrm{a}}[k]$ is still restricted to zero. In state p_0, no self-transitions are allowed. In state r_1, $\mu_{\mathrm{p}}[k]$ and $\mu_{\mathrm{a}}[k]$ become zero again. This specific constraint restricts $\mu_{\mathrm{p}}[k]$ to consisting of isolated deltas. State r_1 leads to states $\mathrm{a}_0, \ldots, \mathrm{a}_{N-1}$, in each of which $\mu_{\mathrm{a}}[k]$ can take a different nonzero value $C_n^{(\mathrm{a})}$, whereas, $\mu_{\mathrm{p}}[k]$ is forced to be zero. Direct state transitions from state a_n to state $\mathrm{a}_{n'}$ $(n \neq n')$ without passing through state r_1 are not allowed. This constraint restricts $\mu_{\mathrm{a}}[k]$ to consisting of rectangular pulses. It should also be noted that the this HMM

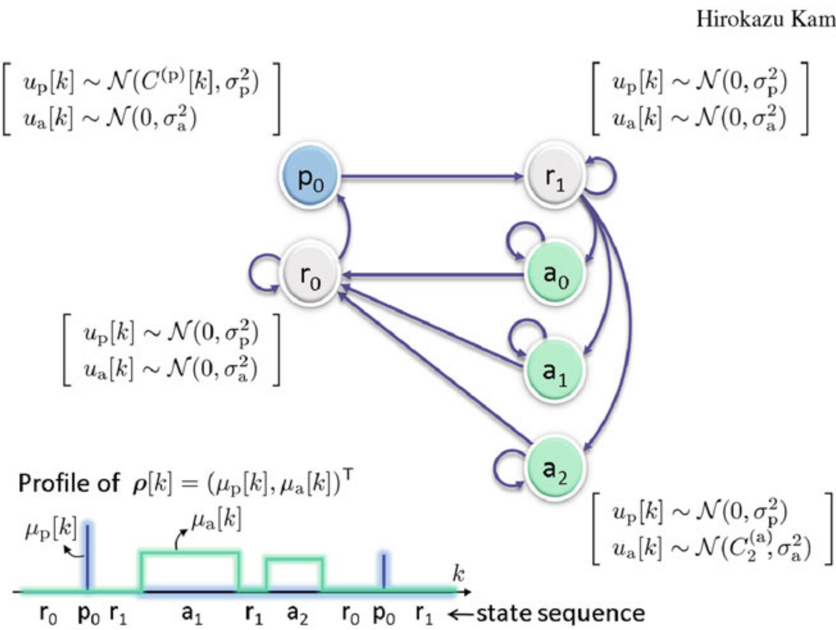


Fig. 4.2 Command function modeling with a path-restricted HMM

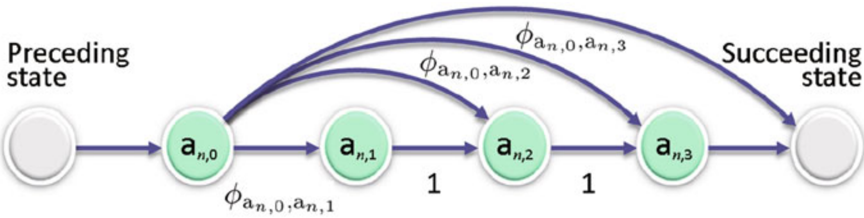


Fig. 4.3 A duration-explicit representation of the hidden states. The splitting of state a_n into substates $a_{n,0}$, $a_{n,1}$, $a_{n,2}$, and $a_{n,3}$ allows us to parameterize the duration of each hidden state. For example, $\phi_{a_{n,0},a_{n,1}}$ corresponds to the probability of staying at state a_n with four consecutive times

ensures that not more than one command will be active at each point in time. The use of the HMM described above for modeling the command functions has been our primary reason for translating the Fujisaki model into the discrete-time counterpart.

To parameterize the durations of the self transitions, we propose to split each state into a certain number of substates such that they all have exactly the same emission densities. Figure 4.3 shows an example of the splitting of state a_n. The number of substates is set at a sufficiently large value and the transition probability from substate $a_{n,m}$ to substate $a_{n,m+1}$ is set at 1 for $m \neq 0$. This state splitting allows us to flexibly control the expectations of the durations for which the process stays in state a_n through the settings of the transition probability. The transition probability from

substate $\mathrm{a}_{n,0}$ to substate $\mathrm{a}_{n,m}$ ($m \geq 1$) corresponds to the probability of the present HMM, generating a rectangular pulse that has a particular duration. In the same way, we split states r_0 and r_1 to parameterize the probability of the spacing between phrase and accent commands. Note that this is equivalent to the explicit-duration HMM proposed by Ferguson (1980). Alternatively, the use of a hidden semi-Markov model (Russell and Moore 1985; Levinson 1986) would also be appropriate for the same purpose. Henceforth, we use the notation $\mathrm{r}_0 = \{\mathrm{r}_{0,0}, \mathrm{r}_{0,1}, \dots\}$, $\mathrm{r}_1 = \{\mathrm{r}_{1,0}, \mathrm{r}_{1,1}, \dots\}$, and $\mathrm{a}_n = \{\mathrm{a}_{n,0}, \mathrm{a}_{n,1}, \dots\}$. Let $\phi_{i',i}$ be the logarithm of the transition probability from state i' and i. To sum up, the present HMM is defined as follows:

$$\text{Output sequence:} \{\boldsymbol{o}[k]\}_{k=0}^{K-1}$$

$$\text{Set of states:} \mathcal{S} = \{\mathrm{r}_0, \mathrm{p}_0, \mathrm{r}_1, \mathrm{a}_0, \dots, \mathrm{a}_{N-1}\}$$

$$\text{State sequence:} \{s_k\}_{k=0}^{K-1}$$

$$\text{Output distribution:} P(\boldsymbol{o}[k]|s_k = i) = \mathcal{N}(\boldsymbol{c}_i[k], \boldsymbol{\Upsilon})$$

$$\boldsymbol{c}_i[k] = \begin{cases} (0,0)^{\mathrm{T}} & (i \in \mathrm{r}_0) \\ (C^{(\mathrm{p})}[k], 0)^{\mathrm{T}} & (i = \mathrm{p}_1) \\ (0,0)^{\mathrm{T}} & (i \in \mathrm{r}_1) \\ (0, C_n^{(\mathrm{a})})^{\mathrm{T}} & (i \in \mathrm{a}_n) \end{cases}$$

$$\boldsymbol{\Upsilon} = \begin{bmatrix} \sigma_{\mathrm{p}}^2 & 0 \\ 0 & \sigma_{\mathrm{a}}^2 \end{bmatrix}$$

$$\text{Transition probability:} \phi_{i',i} = \log P(s_k = i | s_{k-1} = i')$$

The free parameters to be estimated in our command function model consist of the state sequence, $\{s_k\}_{k=0}^{K-1}$, and the mean and variance of the output distribution of each state, $\{C^{(\mathrm{p})}[k]\}_{k=0}^{K-1}$, $\{C_n^{(\mathrm{a})}\}_{n=0}^{N-1}$, $\{\sigma_{\mathrm{p}}^2, \sigma_{\mathrm{a}}^2\}$. Hereafter, we use $\boldsymbol{s}$ to denote $\{s_k\}_{k=0}^{K-1}$ and θ to denote the rest of the parameters:

$$\boldsymbol{s} := \{s_k\}_{k=1}^{K},$$

$$\theta := \{\{C^{(\mathrm{p})}[k]\}_{k=0}^{K-1}, \{C_n^{(\mathrm{a})}\}_{n=0}^{N-1}, \sigma_{\mathrm{p}}^2, \sigma_{\mathrm{a}}^2\}.$$

The generating process for the phrase and accent components is summarized as follows: The state sequence $\{s_k\}_{k=0}^{K-1}$ is first generated according to a Markov chain. Given the state sequence $\{s_k\}_{k=0}^{K-1}$, the mean sequences $\{\mu_{\mathrm{p}}[k]\}_{k=0}^{K-1}$ and $\{\mu_{\mathrm{a}}[k]\}_{k=0}^{K-1}$ are determined by

$$\begin{bmatrix} \mu_{\mathrm{p}}[k] \\ \mu_{\mathrm{a}}[k] \end{bmatrix} = \boldsymbol{\rho}[k] = \boldsymbol{c}_{s_k}[k]. \tag{4.20}$$

Given $\boldsymbol{\rho}[k]$ and $\boldsymbol{\Upsilon}$, the present HMM generates the $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ pair according to Eq. (4.19). From Eqs. (4.13) and (4.14), $u_{\mathrm{p}}[k]$ and $u_{\mathrm{a}}[k]$ are then fed through different all-pole systems to generate the phrase and accent components, $x_{\mathrm{p}}[k]$ and $x_{\mathrm{a}}[k]$.

4.4.2 *Uncertainty of F_0 Observations*

For real speech F_0 contours, observed F_0s should not always be considered reliable. For example, F_0 estimates obtained with a pitch extractor in unvoiced regions would be totally unreliable. When performing parameter inference, we would want to trust only reliable observations and neglect unreliable ones. To incorporate the degree of uncertainty of F_0 observations, we consider modeling an observed F_0 contour $y[k]$ as a superposition of the "ideal" F_0 contour, i.e., $x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k] + \mu_{\mathrm{b}}$, and a normally distributed noise component

$$x_{\mathrm{n}}[k] \sim \mathcal{N}\left(0, \sigma_{\mathrm{n}}^2[k]\right), \tag{4.21}$$

where $\sigma_{\mathrm{n}}^2[k]$ represents the degree of uncertainty of the F_0 observation at time k, which is assumed to be given. For example, one simple way would be to set $\sigma_{\mathrm{n}}^2[k]$ at a small value near 0 for voiced regions and a sufficiently large value for unvoiced regions. By denoting

$$x_{\mathrm{b}}[k] = \mu_b + x_{\mathrm{n}}[k], \tag{4.22}$$

the entire F_0 contour is given by

$$y[k] = x_{\mathrm{p}}[k] + x_{\mathrm{a}}[k] + x_{\mathrm{b}}[k]. \tag{4.23}$$

4.4.3 *Likelihood Function and Prior Probabilities*

In this section, we derive the probability density function of an observed F_0 contour, $y[0], \ldots, y[K-1]$, based on the probabilistic modeling of the command functions and the reliability modeling presented in the previous sections. From Eq. (4.19),

$$u_{\mathrm{p}}[k]|\theta, s_k \sim \mathcal{N}(\mu_{\mathrm{p}}[k], \sigma_{\mathrm{p}}^2), \tag{4.24}$$

$$u_{\mathrm{a}}[k]|\theta, s_k \sim \mathcal{N}(\mu_{\mathrm{a}}[k], \sigma_{\mathrm{a}}^2). \tag{4.25}$$

By defining $\boldsymbol{u}_{\mathrm{p}}$ and $\boldsymbol{u}_{\mathrm{a}}$ by

$$\boldsymbol{u}_{\mathrm{p}} = (u_{\mathrm{p}}[0], \ldots, u_{\mathrm{p}}[K-1])^{\mathsf{T}}, \tag{4.26}$$

$$\boldsymbol{u}_{\mathrm{a}} = (u_{\mathrm{a}}[0], \ldots, u_{\mathrm{a}}[K-1])^{\mathsf{T}}, \tag{4.27}$$

we can write Eqs. (4.24) and (4.25) in vector notation:

$$\boldsymbol{u}_{\mathrm{p}}|\theta, \boldsymbol{s} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{p}}, \boldsymbol{\Sigma}_{\mathrm{p}}), \tag{4.28}$$

$$\boldsymbol{u}_{\mathrm{a}}|\theta, \boldsymbol{s} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathrm{a}}, \boldsymbol{\Sigma}_{\mathrm{a}}), \tag{4.29}$$

where

$$\boldsymbol{\mu}_{\mathrm{p}} = (\mu_{\mathrm{p}}[0], \ldots, \mu_{\mathrm{p}}[K-1])^{\mathsf{T}}, \quad \boldsymbol{\Sigma}_{\mathrm{p}} = \sigma_{\mathrm{p}}^2 \boldsymbol{I}, \tag{4.30}$$

$$\boldsymbol{\mu}_{\mathrm{a}} = (\mu_{\mathrm{a}}[0], \ldots, \mu_{\mathrm{a}}[K-1])^{\mathsf{T}}, \quad \boldsymbol{\Sigma}_{\mathrm{a}} = \sigma_{\mathrm{a}}^2 \boldsymbol{I}. \tag{4.31}$$

By using the linear equation given by Eqs. (4.13) and (4.14), the vectors consisting of the phrase and accent components

$$\boldsymbol{x}_{\mathrm{p}} = (x_{\mathrm{p}}[0], \ldots, x_{\mathrm{p}}[K-1])^{\mathsf{T}}, \tag{4.32}$$

$$\boldsymbol{x}_{\mathrm{a}} = (x_{\mathrm{a}}[0], \ldots, x_{\mathrm{a}}[K-1])^{\mathsf{T}}, \tag{4.33}$$

can be written in terms of $\boldsymbol{u}_{\mathrm{p}}$ and $\boldsymbol{u}_{\mathrm{a}}$,

$$\boldsymbol{u}_{\mathrm{p}} = \boldsymbol{F}_{\mathrm{p}} \boldsymbol{x}_{\mathrm{p}}, \tag{4.34}$$

$$\boldsymbol{u}_{\mathrm{a}} = \boldsymbol{F}_{\mathrm{a}} \boldsymbol{x}_{\mathrm{a}}, \tag{4.35}$$

where

$$\boldsymbol{F}_{\mathrm{p}} := \begin{bmatrix} f_0^{(\mathrm{p})} & & & & O \\ f_1^{(\mathrm{p})} & f_0^{(\mathrm{p})} & & & \\ f_2^{(\mathrm{p})} & f_1^{(\mathrm{p})} & f_0^{(\mathrm{p})} & & \\ & \ddots & \ddots & \ddots & \\ O & & f_2^{(\mathrm{p})} & f_1^{(\mathrm{p})} & f_0^{(\mathrm{p})} \end{bmatrix}, \quad \boldsymbol{F}_{\mathrm{a}} := \begin{bmatrix} f_0^{(\mathrm{a})} & & & & O \\ f_1^{(\mathrm{a})} & f_0^{(\mathrm{a})} & & & \\ f_2^{(\mathrm{a})} & f_1^{(\mathrm{a})} & f_0^{(\mathrm{a})} & & \\ & \ddots & \ddots & \ddots & \\ O & & f_2^{(\mathrm{a})} & f_1^{(\mathrm{a})} & f_0^{(\mathrm{a})} \end{bmatrix}. \tag{4.36}$$

Hence, it follows from Eqs. (4.28), (4.29), (4.34), and (4.35) that

$$\boldsymbol{x}_{\mathrm{p}}|\theta, \boldsymbol{s}, \psi \sim \mathcal{N}(\boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}}, \boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\Sigma}_{\mathrm{p}}(\boldsymbol{F}_{\mathrm{p}}^{-1})^{\mathsf{T}}), \tag{4.37}$$

$$\boldsymbol{x}_{\mathrm{a}}|\theta, \boldsymbol{s}, \varphi \sim \mathcal{N}(\boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\mu}_{\mathrm{a}}, \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\Sigma}_{\mathrm{a}}(\boldsymbol{F}_{\mathrm{a}}^{-1})^{\mathsf{T}}). \tag{4.38}$$

We refer $x_{\mathrm{b}}[k]$ to as the base component and let $\boldsymbol{x}_{\mathrm{b}}$ be

$$\boldsymbol{x}_{\mathrm{b}} = (x_{\mathrm{b}}[0], \ldots, x_{\mathrm{b}}[K-1])^{\mathsf{T}}. \tag{4.39}$$

It follows from Eqs. (4.21) and (4.22) that $\boldsymbol{x}_{\mathrm{b}}$ is normally distributed

$$\boldsymbol{x}_{\mathrm{b}}|\mu_{\mathrm{b}} \sim \mathcal{N}(\mu_{\mathrm{b}}\boldsymbol{1}, \boldsymbol{\Sigma}_{\mathrm{b}}), \tag{4.40}$$

where

$$\mathbf{1} = (1, \ldots, 1)^{\mathsf{T}}, \tag{4.41}$$

$$\boldsymbol{\Sigma}_{\mathrm{b}} = \mathrm{diag}(\sigma_{\mathrm{n}}^2[0], \ldots, \sigma_{\mathrm{n}}^2[K-1]). \tag{4.42}$$

A vector consisting of observed F_0s

$$\boldsymbol{y} = (y[0], \ldots, y[K-1])^{\mathsf{T}}, \tag{4.43}$$

is given by the sum of $\boldsymbol{x}_{\mathrm{p}}$, $\boldsymbol{x}_{\mathrm{a}}$, and $\boldsymbol{x}_{\mathrm{b}}$:

$$\boldsymbol{y} = \boldsymbol{x}_{\mathrm{p}} + \boldsymbol{x}_{\mathrm{a}} + \boldsymbol{x}_{\mathrm{b}}. \tag{4.44}$$

From Eqs. (4.44), (4.37), (4.38), and (4.40), $\boldsymbol{y}$ is normally distributed such that

$$\boldsymbol{y}|\Theta \sim \mathcal{N}(\boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\mu} + \mu_{\mathrm{b}}\mathbf{1}, \boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\Sigma}_{\mathrm{p}}(\boldsymbol{F}_{\mathrm{p}}^{-1})^{\mathsf{T}} + \boldsymbol{F}_{\mathrm{a}}{}^{-1}\boldsymbol{\Sigma}_{\mathrm{a}}(\boldsymbol{F}_{\mathrm{a}}^{-1})^{\mathsf{T}} + \boldsymbol{\Sigma}_{\mathrm{b}}), \tag{4.45}$$

where $\Theta := \{\theta, \boldsymbol{s}, \psi, \varphi, \mu_{\mathrm{b}}\}$. Overall, the likelihood function of the Fujisaki model parameters Θ given $\boldsymbol{y}$ can be written as

$$\begin{aligned} P(\boldsymbol{y}|\Theta) &= \frac{1}{\sqrt{(2\pi)^K|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right\}, \\ \boldsymbol{\mu} &= \boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}} + \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\mu}_{\mathrm{a}} + \mu_{\mathrm{b}}\mathbf{1}, \\ \boldsymbol{\Sigma} &= \boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\Sigma}_{\mathrm{p}}(\boldsymbol{F}_{\mathrm{p}}^{\mathrm{T}})^{-1} + \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\Sigma}_{\mathrm{a}}(\boldsymbol{F}_{\mathrm{a}}^{\mathrm{T}})^{-1} + \boldsymbol{\Sigma}_{\mathrm{b}}. \end{aligned} \tag{4.46}$$

As for the prior probability of Θ, we assume that the phrase control parameter ψ, accent control parameter φ, and state sequence $\{s[k]\}_{k=0}^{K-1}$ are independent of each other. Recall that we assumed in 4.4.1 that $\{s[k]\}_{k=0}^{K-1}$ is a first-order Markov chain. We further assume that all other parameters are uniformly distributed. Thus,

$$P(\Theta) \propto P(\psi)P(\varphi)P(\boldsymbol{s}), \tag{4.47}$$

$$P(\boldsymbol{s}) = P(s_0)\prod_{k=1}^{K-1} P(s_k|s_{k-1}). \tag{4.48}$$

4.5 Parameter Estimation Algorithm

Here we describe an iterative algorithm, which locally maximizes the posterior density of Θ given $\boldsymbol{y}$, $P(\Theta|\boldsymbol{y}) \propto P(\boldsymbol{y}|\Theta)P(\Theta)$. By regarding the set consisting of the phrase, accent, and base components, $\boldsymbol{x} := (\boldsymbol{x}_{\mathrm{p}}^{\mathsf{T}}, \boldsymbol{x}_{\mathrm{a}}^{\mathsf{T}}, \boldsymbol{x}_{\mathrm{b}}^{\mathsf{T}})^{\mathsf{T}}$, as the complete data, this problem can be viewed as an incomplete data problem, which can be dealt with using the expectation–maximization (EM) algorithm (Dempster et al. 1977; Feder and Weinstein 1988).

The log-likelihood of Θ given, the complete data is given as

$$\log P(\boldsymbol{x}|\Theta) \stackrel{\Theta}{=} \frac{1}{2}\log|\boldsymbol{\Lambda}^{-1}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{x}-\boldsymbol{m}),$$

$$\boldsymbol{x} := \begin{bmatrix} \boldsymbol{x}_{\mathrm{p}} \\ \boldsymbol{x}_{\mathrm{a}} \\ \boldsymbol{x}_{\mathrm{b}} \end{bmatrix}, \ \boldsymbol{m} := \begin{bmatrix} \boldsymbol{F}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}} \\ \boldsymbol{F}_{\mathrm{a}}^{-1}\boldsymbol{\mu}_{\mathrm{a}} \\ \mu_{\mathrm{b}}\mathbf{1} \end{bmatrix}, \tag{4.49}$$

$$\boldsymbol{\Lambda}^{-1} := \begin{bmatrix} \boldsymbol{F}_{\mathrm{p}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{F}_{\mathrm{a}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{F}_{\mathrm{a}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Sigma}_{\mathrm{b}}^{-1} \end{bmatrix},$$

where $\stackrel{\xi}{=}$ denotes equality up to a term independent of ξ. Taking the conditional expectation of Eq. (4.49) with respect to $\boldsymbol{x}$ given $\mathbf{y}$ and $\Theta = \Theta'$, and then adding $\log P(\Theta)$ to both sides, we obtain an auxiliary function

$$Q(\Theta,\Theta') \stackrel{\Theta}{=} \frac{1}{2}\Big[\log|\boldsymbol{\Lambda}^{-1}| - \mathrm{tr}(\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\mathbf{y};\Theta']) + 2\boldsymbol{m}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\mathbb{E}[\boldsymbol{x}|\mathbf{y};\Theta'] - \boldsymbol{m}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{m}\Big] + \log P(\Theta). \tag{4.50}$$

Since the incomplete data and the complete data are related through the linear equation $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$, where $\boldsymbol{H} := [\boldsymbol{I},\boldsymbol{I},\boldsymbol{I}]$, $\mathbb{E}[\boldsymbol{x}|\mathbf{y};\Theta]$, and $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{T}}|\mathbf{y};\Theta]$ are given explicitly as

$$\mathbb{E}[\boldsymbol{x}|\mathbf{y};\Theta] = \boldsymbol{m} + \boldsymbol{\Lambda}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathsf{T}})^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{m}), \tag{4.51}$$

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\mathbf{y};\Theta] = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{H}^{\mathsf{T}}(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}^{\mathsf{T}})^{-1}\boldsymbol{H}\boldsymbol{\Lambda} + \mathbb{E}[\boldsymbol{x}|\mathbf{y};\Theta]\mathbb{E}[\boldsymbol{x}|\mathbf{y};\Theta]^{\mathsf{T}}. \tag{4.52}$$

It can be shown that an iterative procedure consisting of maximizing $Q(\Theta,\Theta')$ with respect to Θ (the maximization step) and substituting Θ into Θ' (the expectation step) locally maximizes the posterior density $P(\Theta|\mathbf{y})$. The expectation step computes $\mathbb{E}[\boldsymbol{x}|\mathbf{y};\Theta']$ and $\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathsf{T}}|\mathbf{y};\Theta']$ to Eqs. (4.51) and (4.52) by substituting the current parameter estimate into Θ'.

Now, let $\mathbb{E}[\boldsymbol{x}|\mathbf{y};\Theta']$ be partitioned into three $K \times 1$ blocks and $\mathbb{E}[\mathbf{x}\boldsymbol{x}^{T}|\mathbf{y};\Theta']$ into nine $K \times K$ blocks such that

$$\mathbb{E}[\boldsymbol{x}|\mathbf{y};\Theta'] = \begin{bmatrix} \bar{\boldsymbol{x}}_{\mathrm{p}} \\ \bar{\boldsymbol{x}}_{\mathrm{a}} \\ \bar{\boldsymbol{x}}_{\mathrm{b}} \end{bmatrix}, \ \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^{\mathsf{T}}|\mathbf{y};\Theta'] = \begin{bmatrix} \boldsymbol{R}_{\mathrm{p}} & * & * \\ * & \boldsymbol{R}_{\mathrm{a}} & * \\ * & * & \boldsymbol{R}_{\mathrm{b}} \end{bmatrix}, \tag{4.53}$$

where $*$ stands for blocks that we can ignore hereafter. The auxiliary function can then be rewritten in a more convenient form:

$$Q(\Theta,\Theta') \stackrel{\Theta}{=} \frac{1}{2}\Big[\log|\boldsymbol{F}_{\mathrm{p}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}}| + \log|\boldsymbol{F}_{\mathrm{a}}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{F}_{\mathrm{a}}| + \log|\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}|$$

$$
\begin{aligned}
&- \mathrm{tr}(\boldsymbol{F}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}}\boldsymbol{R}_{\mathrm{p}}) + 2\boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{F}_{\mathrm{p}}\bar{\boldsymbol{x}}_{\mathrm{p}} \\
&- \mathrm{tr}(\boldsymbol{F}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{F}_{\mathrm{a}}\boldsymbol{R}_{\mathrm{a}}) + 2\boldsymbol{\mu}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{F}_{\mathrm{a}}\bar{\boldsymbol{x}}_{\mathrm{a}} \\
&- \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\boldsymbol{R}_{\mathrm{b}}) + 2\mu_{\mathrm{b}}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\bar{\boldsymbol{x}}_{\mathrm{b}} \\
&- \boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{p}}^{-1}\boldsymbol{\mu}_{\mathrm{p}} - \boldsymbol{\mu}_{\mathrm{a}}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{a}}^{-1}\boldsymbol{\mu}_{\mathrm{a}} - \mu_{\mathrm{b}}^{2}\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{b}}^{-1}\mathbf{1}\big] + \log P(\Theta). \qquad (4.54)
\end{aligned}
$$

The update formula for each parameter in the maximization step can be derived using Eq. (4.54). Owing to space limitations, we omit the update equations for ψ and φ (readers are referred to Kameoka et al. 2010 if interested). In practice, these parameters can be treated as constants, as mentioned earlier.

1) State sequence $s_0, \ldots, s_{K-1}$: Leaving only the terms in $Q(\Theta, \Theta')$ that depend on $s := \{s_k\}_{k=0}^{K-1}$, we have

$$
Q(\Theta, \Theta') \overset{s}{=} -\frac{1}{2}\sum_{k=0}^{K-1}(\boldsymbol{o}[k] - \boldsymbol{c}_{s_k}[k])^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}(\mathbf{o}[k] - \mathbf{c}_{s_k}[k]) + \sum_{k=1}^{K-1}\phi_{s_{k-1},s_k}, \qquad (4.55)
$$

where $\boldsymbol{o}[k] := ([\boldsymbol{F}_{\mathrm{p}}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k, [\boldsymbol{F}_{\mathrm{a}}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k)^{\mathrm{T}}$. Here the notation $[\cdot]_k$ is used to denote the k-th element of a vector. The state sequence $\{s_k\}_{k=0}^{K-1}$ maximizing $Q(\Theta, \Theta')$ can be solved efficiently using the Viterbi algorithm as follows. We first set $\delta_0(i)$ at

$$
\delta_0(i) = -\frac{1}{2}(\boldsymbol{o}[0] - \boldsymbol{c}_i[0])^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}(\mathbf{o}[0] - \mathbf{c}_i[0]) + \phi_i \qquad (4.56)
$$

for all the hidden states i. To let state $\mathrm{r}_{0,0}$ be the starting state, we shall set ϕ_i at

$$
\phi_i = \begin{cases} 0 & (i = \mathrm{r}_{0,0}) \\ -\infty & (i \neq \mathrm{r}_{0,0}) \end{cases}. \qquad (4.57)
$$

We can then compute $\delta_k(i)$ for $k = 1, \ldots, K-1$ recursively via

$$
\delta_k(i) = \max_{i'}[\delta_{k-1}(i') + \phi_{i',i}] - \frac{1}{2}(\boldsymbol{o}[k] - \boldsymbol{c}_i[k])^{\mathrm{T}}\boldsymbol{\Upsilon}^{-1}(\boldsymbol{o}[k] - \boldsymbol{c}_i[k]). \qquad (4.58)
$$

The most likely transition for each state should be registered at each recursion $\Psi_k(i) = \mathrm{argmax}_{i'}[\delta_{k-1}(i') + \phi_{i',i}]$, so that the most likely state sequence can be traced at the end of the recursion with $s_{k-1} = \Psi_k(s_k)$ $(k = K-1, \ldots, 1)$, where $s_K = \mathrm{argmax}_i \delta_K(i)$. Substituting the updated state sequence $\{s_k\}$ into Eq. (4.20), we finally obtain the updated $\boldsymbol{\mu}_{\mathrm{p}}$ and $\boldsymbol{\mu}_{\mathrm{a}}$.

2) Magnitude of phrase command $C^{(\mathrm{p})}[k]$:
$Q(\Theta, \Theta')$ is maximized with respect to $C^{(\mathrm{p})}[k]$ when

$$
C^{(\mathrm{p})}[k] = [\boldsymbol{F}_{\mathrm{p}}\bar{\boldsymbol{x}}_{\mathrm{p}}]_k \ \ (k \in \mathcal{K}_{\mathrm{p}_0}), \ \ \mathcal{K}_{\mathrm{p}_0} = \{k | s_k = \mathrm{p}_0\}. \qquad (4.59)
$$

3) Magnitude of accent command $C_n^{(\mathrm{a})}$:
$Q(\Theta, \Theta')$ is maximized with respect to $C_n^{(\mathrm{a})}$ when

$$
C_n^{(\mathrm{a})} = \frac{1}{|\mathcal{K}_{\mathrm{a}_n}|}\sum_{k \in \mathcal{K}_{\mathrm{a}_n}}[\boldsymbol{F}_{\mathrm{a}}\bar{\boldsymbol{x}}_{\mathrm{a}}]_k, \ \ \mathcal{K}_{\mathrm{a}_n} = \{k | s_k \in \mathrm{a}_n\}. \qquad (4.60)
$$

4) Baseline value μ_{b}:
$Q(\Theta, \Theta')$ is maximized with respect to μ_{b} when

$$\mu_{\mathrm{b}} = \frac{\mathbf{1}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{b}}^{-1} \bar{\boldsymbol{x}}_{\mathrm{b}}}{\mathbf{1}^{\mathrm{T}} \boldsymbol{\Sigma}_{\mathrm{b}}^{-1} \mathbf{1}} = \frac{\sum_k [\bar{\boldsymbol{x}}_{\mathrm{b}}]_k / \sigma_{\mathrm{n}}[k]^2}{\sum_k 1/\sigma_{\mathrm{n}}[k]^2}. \tag{4.61}$$

5) Variances of state emission densities σ_{p}^2, σ_{a}^2:
$Q(\Theta, \Theta')$ is maximized with respect to $\sigma_{\mathrm{p},i}^2$ and $\sigma_{\mathrm{a},i}^2$ when

$$\sigma_{\mathrm{p}}^2 = (\mathrm{tr}(\boldsymbol{F}_{\mathrm{p}}^{\mathrm{T}} \boldsymbol{F}_{\mathrm{p}} \boldsymbol{R}_{\mathrm{p}}) - 2\boldsymbol{\mu}_{\mathrm{p}}^{\mathrm{pT}} \boldsymbol{F}_{\mathrm{p}} \bar{\mathbf{x}}_{\mathrm{p}} + \boldsymbol{\mu}_{\mathrm{T}} \boldsymbol{\mu}_{\mathrm{p}})/K, \tag{4.62}$$

$$\sigma_{\mathrm{a}}^2 = (\mathrm{tr}(\boldsymbol{F}_{\mathrm{a}}^{\mathrm{T}} \boldsymbol{F}_{\mathrm{a}} \boldsymbol{R}_{\mathrm{a}}) - 2\boldsymbol{\mu}_{\mathrm{a}}^{\mathrm{aT}} \boldsymbol{F}_{\mathrm{a}} \bar{\mathbf{x}}_{\mathrm{a}} + \boldsymbol{\mu}_{\mathrm{T}} \boldsymbol{\mu}_{\mathrm{a}})/K. \tag{4.63}$$

To summarize, we obtain the following iterative algorithm that guarantees monotonic convergence to a local maximum of the posterior density $P(\Theta|\boldsymbol{y})$:

1. (E-step) Update $\bar{\boldsymbol{x}}_{\mathrm{p}}$, $\bar{\boldsymbol{x}}_{\mathrm{a}}$, $\bar{\boldsymbol{x}}_{\mathrm{b}}$, $\mathbf{R}_{\mathrm{p}}$, $\mathbf{R}_{\mathrm{a}}$ and $\mathbf{R}_{\mathrm{b}}$ via Eqs. (4.51) and (4.52).
2. (M-step) Increase $Q(\Theta, \Theta')$ w.r.t. Θ through the following updates:
   a) Update $\boldsymbol{s}$ by using the Viterbi algorithm.
   b) Update θ via Eqs. (4.59), (4.60), (4.62) and (4.63).
   c) Update μ_{b} via Eq. (4.61).
   Return to 1 until convergence.

4.6 Evaluation of Parameter Estimation Accuracy

4.6.1 Parameter Estimation Using Synthetic F_0 Contours

To evaluate the pure behavior of the present parameter estimation algorithm, we conducted a command estimation experiment using synthetic F_0 contours. We chose a Fujisaki model parameter extractor developed by Narusawa et al. (2002) as a baseline method for comparison.

The synthetic F_0 contours were artificially created using the original Fujisaki model with the manually annotated command sequences associated with the speech data in the ATR Japanese speech database B-set (Kurematsu et al. 1999). This database consists of 503 phonetically balanced sentences.

The constant parameters were fixed at $N = 10$, $t_0 = 8$ ms, $\alpha = 3.0$ rad/s, $\beta = 20.0$ rad/s, $\upsilon_{\mathrm{p}}^2 = 0.2^2$, $\upsilon_{\mathrm{a}}^2 = 0.1^2$, and $\upsilon_{\mathrm{n}}^2[k] = 10^{15}$ respectively, for unvoiced regions and $\upsilon_{\mathrm{n}}^2[k] = 0.2^2$ for voiced regions. μ_{b} was set at the minimum log F_0 value in the voiced regions. The initial values of Θ were set at the values obtained with Narusawa's method (Narusawa et al. 2002). The EM algorithm was then run for

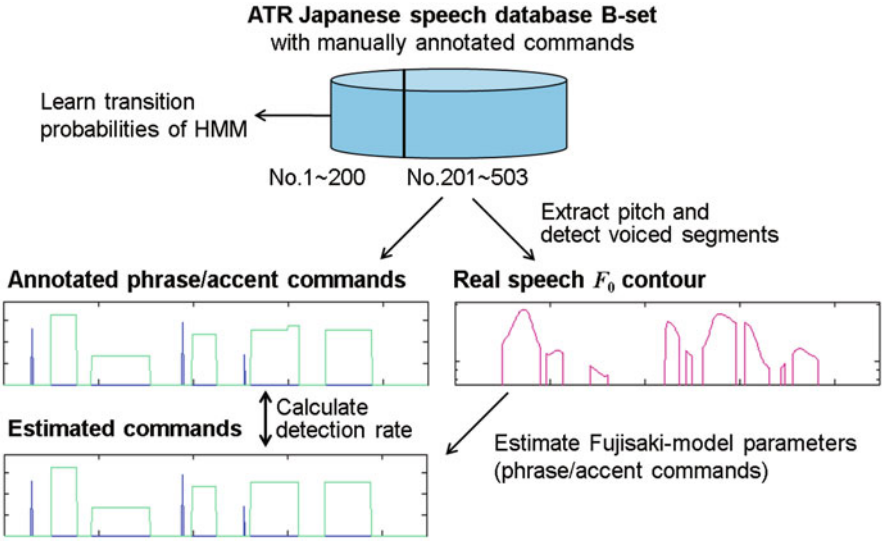


Fig. 4.4 Overview of the experiment described in Sect. 4.6.1

20 iterations. The number of substates in the HMM and the transition probability $\phi_{i',i}$ were determined according to the manually annotated data of the first 200 sentences. The parameter estimation algorithm was then tested on the remaining 303 sentences.

The present method obtained the detection rate of 83.4 % while Narusawa's method only obtained 72.6 %. Readers are referred to (Yoshizato et al. 2012) for more details on the way in which we obtained the detection rate. With this experiment, the present method was shown to be significantly superior to the conventional method in terms of the ability to extract command sequences. Another important conclusion drawn from the results was that the approximation error due to the discretization did not appear to be a crucial matter.

4.6.2 Parameter Estimation Using Real Speech Data

We also conducted an experiment using real speech signals, excerpted from the same database (Kurematsu et al. 1999). Figure 4.4 illustrates an overview of this experiment. F_0 contours were extracted using a method described in (Kameoka 2007). All other experimental conditions were the same as above.

Figures 4.5 and 4.6 show examples of observed F_0 contours and the estimated F_0 contours obtained with the conventional and present methods. We can confirm from these examples that the present method was able to fit the model to observed F_0 contours more accurately than the conventional method.

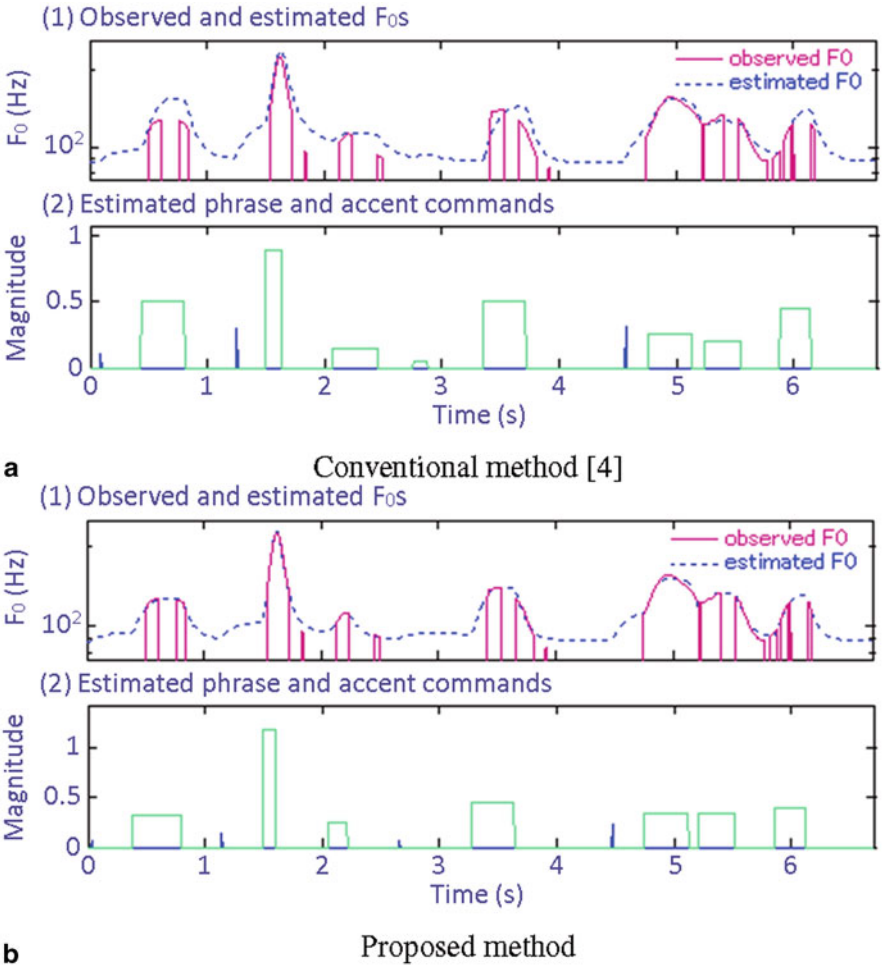


Fig. 4.5 *1* An observed F_0 contour in voiced regions (in *solid line*) and the estimated F_0 contours (in *dotted line*) along with, *2* the estimated phrase and accent commands

4.7 Application to Prosody Generation for Text-to-Speech (TTS) Synthesis

4.7.1 F_0 Contour Synthesis

Statistical parametric speech synthesis has become competitive with established concatenative techniques over the last decade. One successful approach is based on HMMs (Yoshimura et al. 1999; Tokuda et al. 2000). The HMM-based systems model a sequence of spectra, F_0s, and their delta and acceleration components within a unified HMM framework. At the synthesis stage, a sequence of these parameters is

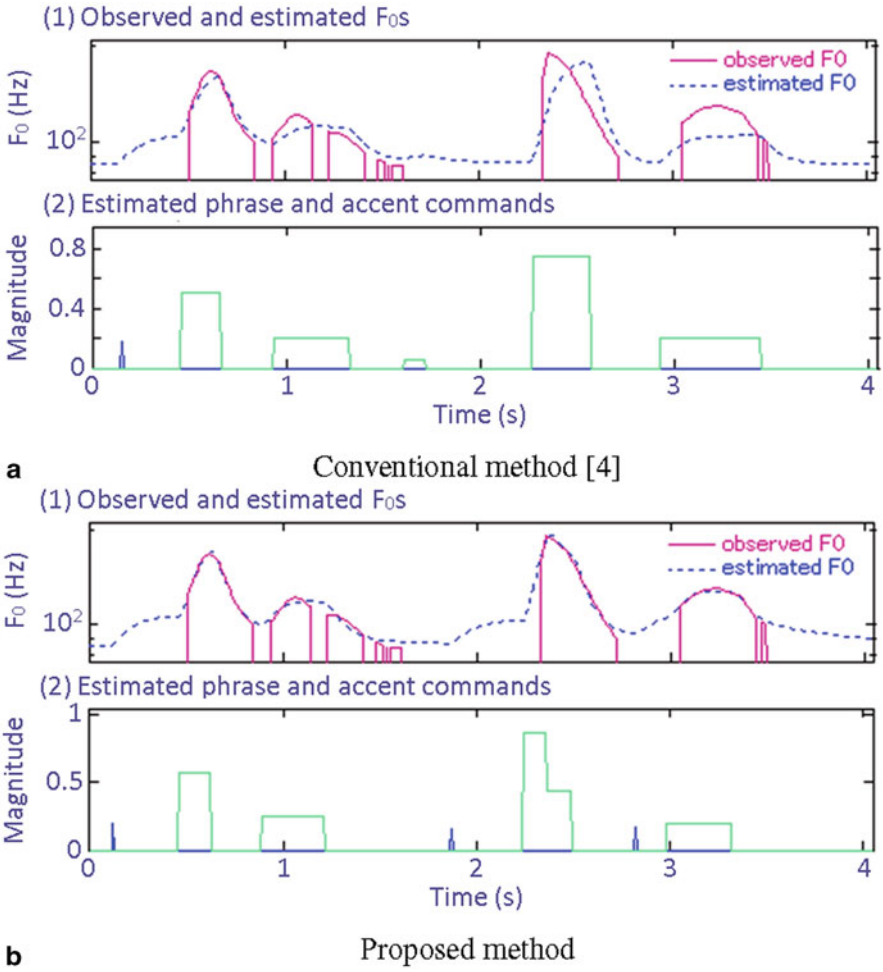


Fig. 4.6 *1* An observed F_0 contour in voiced regions (in *solid line*) and the estimated F_0 contours (in *dotted line*) along with, *2* the estimated phrase and accent commands

generated according to the output probabilities of the trained HMM given an input sentence. The constraints of the dynamic parameters are usually considered during parameter generation in order to guarantee the smoothness of the generated spectral and F_0 trajectories.

Although the quality of synthesized speech generated by the conventional HMM-based systems has steadily improved, it is still easily distinguishable from actual human speech. One problem is that they sometimes produce physically unnatural F_0 contours. This is because the current HMM-based systems fail to take account of the physical mechanism of phonation to express macroscopic F_0 variations. To avoid synthesizing such contours, one reasonable approach would be to learn and

generate the Fujisaki model parameters instead of the F_0 values themselves from text inputs. In this way, we can assure that we can always generate physically natural F_0 contours. Another important advantage of this approach involves the flexibility to change the speaker characteristics and speaking style of the generated speech.

The formulation in the previous sections has not only allowed us to derive an efficient parameter inference algorithm utilizing powerful statistical methods but also to obtain an automatically trainable version of the Fujisaki model. This section further extends this model to a context-dependent one so as to be able to learn and generate the Fujisaki parameters from input sentences. In the following, we assume that a sequence of spectral parameters and the corresponding sequence of context labels are obtained using an existing TTS system (including, but not limited to, HMM-based ones) both at the training and synthesis stages.

It is important to note that phrase and accent commands can be interpreted as quantities related to linguistic information. In the Japanese language, for example, a phrase command and an accent command typically occur at the beginning of each breath group and over the range of accent nucleus in each accentual phrase, respectively. Therefore, we expect that we can obtain an appropriate F_0 contour from a text input by allocating phrase and accent commands to those locations. Here, the problem is how to determine the magnitudes of the phrase and accent commands. Thus, we must treat the magnitude of each command as the model parameter to be trained.

4.7.2 Decision Tree-Based Context Clustering

As mentioned above, the positions of the breath groups and the accent nucleus can be roughly predicted according to the given sequence of the context labels. However, these positions are not always exactly equal to the true positions of the phrase and accent commands that underlie the training utterance. Thus, it would be preferable to estimate the positions as well as the magnitudes of the phrase and accent commands during the training process. To do so, we construct a left-to-right HMM with the states $\mathrm{r}_0, \mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{p}_0, \mathrm{p}_1, \mathrm{p}_2, \ldots, \mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_2, \ldots$ concatenated in appropriate order according to the transcription of the training utterance, as depicted in Fig. 4.7. Each state is then split into substates in the same way as Fig. 7.1 and the transition probabilies associated with the substates are set such that the expected duration of each state becomes equal to the duration predicted from the transcription. Determining the state sequence $\boldsymbol{s}$ under this setting allows us to estimate the precise positions of the phrase and accent commands of the training utterance with the help of the linguistic information.

The generative model under consideration is the same as Eq. (4.45) except that the HMM topology shown in Fig. 4.2 is replaced with the above left-to-right topology. Now, we would like to construct a context dependent model such that the states $\mathrm{p}_0, \mathrm{p}_1, \mathrm{p}_2, \ldots$ and $\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_2, \ldots$ are associated with as many combinations of contextual factors as possible. However, as contextual factors increase, their combinations

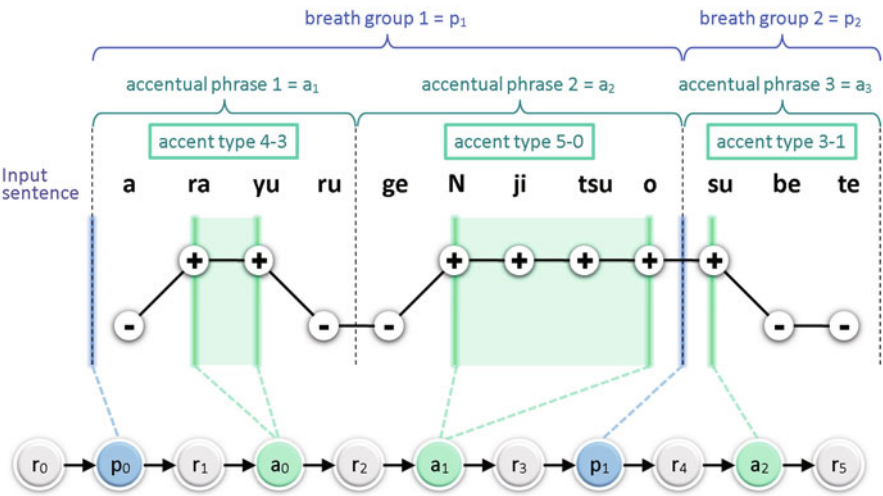


Fig. 4.7 Construction of a *left*-to-*right* HMM using linguistic information

increase exponentially. Furthermore, a significant number of context combinations are usually missing or unseen in the training data. Thus, learning such a model with limited training data often results in overfitting and poor generalization. In conventional HMM-based speech synthesis systems, a decision tree-based context clustering technique has successfully been applied to mitigate this problem. We adopt a similar idea to contruct a context dependent version of the present F_0 contour model.

We construct two decision trees for $\mathrm{p}_0, \mathrm{p}_1, \mathrm{p}_2, \ldots$ and $\mathrm{a}_0, \mathrm{a}_1, \mathrm{a}_2, \ldots$, respectively. Each decision tree is a binary tree in which a context-related yes/no question is attached to each node and is grown in a top–down fashion. Starting with the root node, each node is successively split by selecting a yes/no question that gives the minimum of the description length,

$$-\max_{\Theta^{(L)}} \sum_{d=1}^{D} \log p(\boldsymbol{y}^{(d)}|\Theta^{(L)}) + L \log D/2, \tag{4.64}$$

where $\boldsymbol{y}^{(d)}$ denotes the d-th training example, D the number of the training examples, and $\Theta^{(L)}$ the set consisting of L free parameters. Initially, all the states to be clustered are placed in the root node of the tree and the (locally) maximum log-likelihood of the training data is computed using the algorithm described in Sect. 4.5 subject to the constraint that the emission densities of all the states in the same node are tied. This node is then split into two by choosing an optimal question from a finite set that partitions the states in the parent node so as to give the greatest decrease in the description length. The node to be split next is selected by applying the above procedure to all the nodes and finding the node and the question that give the minimum description length. This procedure is repeated until splitting no longer decreases the description length. Owing to space limitations, the contextual factors and the yes/no

questions we used for the clustering are omitted. For more details, readers are referred to (Kadowaki et al. 2014).

4.7.3 Parameter Generation

Given a text input, we can predict the state sequence $\boldsymbol{s}$, according to which we can generate the sequence of the phrase and accent commands by tracking the mean sequence of the trained HMM. The F_0 contour can then be constructed by using the definition of the Fujisaki model.

Some examples of the synthesized speech generated by the present and conventional methods are demonstrated in our demo site: http://www.brl.ntt.co.jp/people/kameoka/Demos/.

4.7.4 Related Work

It should be noted that some methods have already been developed for predicting the Fujisaki model parameters from text inputs (Hirose et al. 2001; Sakurai et al. 2003). Since this approach treats the Fujisaki model parameters as the features to predict, they must be extracted from the raw F_0 contours of the training utterances prior to learning the prediction model. However, as described earlier, the extraction of the Fujisaki model parameters is not a simple task and the errors made at the parameter extraction stage will directly propagate to the subsequent training stage, resulting in poor prediction. By contrast, the present approach simultaneously performs these two processes with joint optimization. The present F_0 contour model allows for the incorporation of the linguistic information available from the training data set into the parameter estimation (learning) process, which can be done simply by constructing an HMM with a left-to-right topology and running the algorithm described in Sect. 4.5.

4.8 Conclusions

We proposed to introduce a generative model of speech F_0 contours. The present F_0 contour model was formulated by translating the Fujisaki model, a well-founded mathematical model representing the control mechanism of vocal fold vibration, into a probabilistic model described as a discrete-time stochastic process. There were two motivations behind this formulation. One was to derive a general parameter estimation framework for the Fujisaki model, allowing for the introduction of powerful algorithms such as the Viterbi algorithm, forward–backward algorithm, and EM algorithm. The other was to construct an automatically trainable version of the Fujisaki

model so that in future it can be incorporated to statistical speech synthesis and conversion frameworks. We showed an example of the application of the present F_0 contour model to prosody generation for TTS synthesis.

References

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* 39 (1): 1–38.

Feder, M., and E. Weinstein .1988. Parameter estimation of superimposed signals using the EM algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 36 (4): 477–489.

Ferguson, J. D. 1980. Variable duration models for speech. In *Proceedings of Symptoms Application of Hidden Markov Models to Text and Speech*, 143–179.

Fujisaki, H. 1988. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In *Vocal Physiology: Voice Production, Mechanisms and Functions*, ed. O. Fujimura, 347–355. New York: Raven.

Fujisaki, H., and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of japanese. *Journal of the Acoustic Society of Japan (E)* 5 (4): 233–242.

Hirose, K., M. Eto, N. Minematsu, and A. Sakurai. 2001. Corpus-based synthesis of fundamental frequency contours based on a generation process model. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2255–2258.

Kadowaki, K., T. Ishihara, N. Hojo, and H. Kameoka. 2014. Speech prosody generation for text-to-speech synthesis based on generative model of F_0 contours. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association (Interspeech 2014)*.

Kameoka, H. 2007. Statistical approach to multipitch analysis. Ph.D. dissertation, The University of Tokyo.

Kameoka, H., T. Nakatani, and T. Yoshioka. 2009. Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, 45–48.

Kameoka, H., J. Le Roux, and Y. Ohishi. 2010. A statistical model of speech F_0 contours. In *Proceedings of the 2010 ISCA Workshop on Statistical and Perceptual Audition (SAPA 2010)*, 43–48.

Kameoka, H., K. Yoshizato, T. Ishihara, Y. Ohishi, K. Kashino, and S. Sagayama. 2013. Generative modeling of speech F_0 contours. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013)*, 1826–1830.

Kurematsu, A., K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. 1999. ATR japanese speech database as a tool of speech recognition and synthesis. *Speech Communication* 27:187–207.

Levinson, S. 1986. Continuously variable duration hidden Markov models for speech analysis. In *Proceedings of the 1986 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1986)*, 1241–1244.

Mixdorff, H. 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*, vol. 3, 1281–1284.

Narusawa, S., N. Minematsu, K. Hirose, and H. Fujisaki. 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. In *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)*, 509–512.

Russell, M., and R. Moore. 1958. Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition. In *Proceedings of the 1985 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1985)*, 5–8.

Sakurai, A., K. Hirose, and N. Minematsu. 2003. Data-driven generation of F_0 contours using a superpositional model. *Speech Communication* 40 (4):535–549.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2000)*, 1315–1318.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1999)*, 2347–2350.

Yoshizato, K., H. Kameoka, D. Saito, and S. Sagayama. 2012. Statistical approach to fujisaki-model parameter estimation from speech signals and its quantitative evaluation. In *Proceedings of the 6th International Conference on Speech Prosody 2012*, 175–178.

Yoshizato, K., H. Kameoka, D. Saito, and S. Sagayama. 2012. Hidden markov convolutive mixture model for pitch contour analysis of speech. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech 2012)*.

Part II
Para- and Non-Linguistic Issues of Prosody

Chapter 5
Communicative Speech Synthesis as Pan-Linguistic Prosody Control

Yoshinori Sagisaka and Yoko Greenberg

Abstract A series of prosody analyses have been carried out to find the correlations between lexical attributes and communicative prosody variations. By employing lexical attributes showing their impressions, Multi-Dimensional Scaling (MDS) has revealed that three-dimensional perceptual impression space (positive–negative, confident–doubtful, allowable–unacceptable) nicely correlates to F0 heights and their dynamics. Based on these correlations, a communicative prosody generation scheme is newly proposed for input text based on impression attributes of input lexicons using the command–response model in a pan–linguistic language common framework. In this scheme, the communicative component derived by constituent lexicons is added to conventional reading style prosody. This communicative prosody generation formalism not only provides the first methodology for communicative prosody generation, but also creates a new linguistic research paradigm where traditionally abandoned notions like *paralinguistics* will become linguistically essential core items for communicative spoken language science.

5.1 Communicative Prosody

In traditional linguistics, quite physiological topics and colloquial speech intrinsic phenomena have been excluded from the theoretical research targets (Sapir 1921). This restriction of research targets has nicely worked to build many essential linguistic theories without being bothered by treating many irregular linguistic phenomena as performance. This restriction has also been inherited in prosody studies. As a natural result, most research efforts in prosody have been devoted to the studies on speech prosody of well-formed utterances and their linguistic structures or accentual properties, which are closely related to the interests in written language. In this sense, mainly speech phenomena closely related to written language have been widely studied and others such as the variations among speech utterances have neither been well studied from a linguistic viewpoint nor from the speech engineering point of

Y. Sagisaka (✉) · Y. Greenberg
Waseda University, Tokyo, Japan
e-mail: ysagisaka@gmail.com

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_5

view. As even their description has not yet been discussed, we have difficulties describing the existing problems.

On the other hand, recently, as seen in papers of speech related journals and international conferences, many research interests have been attracted to so-called "*paralinguistics*." There are many paper submissions on emotional speech or other prosodic phenomena which have not been treated in traditional studies. As there has been little interest to specify the prosody variations observed in real communications and characterize their communicative functions, these phenomena have been simply treated as *paralinguistics* "out-of-linguistics," though it plays an essential role in speech communication. Are they really to be treated as *paralinguistics* "out-of-linguistics?"

In the field of speech synthesis, there are growing needs of speech output for a man–machine communication system. However, speech needed in communication has not yet been systematically studied. As the contents of some chapters in this book represent, by following the traditional linguistic studies, speech researchers simply treat this characteristic as *paralinguistics* and no further studies have been carried out to link to linguistic attributes. Though some types of speech categorized as emotional speech have been studied only for ease of identifying their output characteristics, they are not sufficient to cover the purpose for a communication system. As we can see in our daily speech, we need speech output possessing characteristics to attain dialogue functions know as *dialogue act* by which the speaker's intention can be smoothly conveyed to listeners. Up to now, there has been little study to characterize this type of speech variation and relate it to linguistic characteristics that we can expect to use as an input for the synthesis of speech for the communication system.

We have started to analyze nonreading speech aiming at synthesis of more realistic human speech for communication systems. To distinguish with conventional reading-style speech which has been a research target for a long time in text-to-speech technology, we would like to use the term "*communicative prosody*" hereafter to imply prosody which is generated by considering its communicative function. It can be considered as speech attributed to some information by a sender to a receiver. To enable communicative speech output, if we can specify some of the communicative characteristics in the prosody domain, we will be able to add new communicative factors to prosody control by considering its function. As it looked quite difficult to characterize the prosody of communicative speech in general, we started the analysis with simple utterances consisting of an adjective with an adverb expressing magnitude (Sagisaka et al. 2005).

5.2 Lexicons and Communicative Prosody

As the first step towards communicative prosody characterization, we started to find the differences of prosody between read speech and communicative speech obtained from simulated conversations (Sagisaka et al. 2005). For speech contents, we designed two-phrase utterances consisting of Japanese adjective and adverb phrases

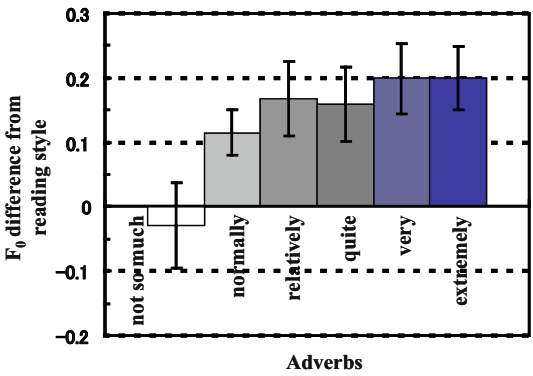


Fig. 5.1 The increase of F0 average difference between reading style and communicative style in proportion to the increase of degree of adverbs when positive adjectives follow

expressing different degrees under designed conversational situations. Five paired adjectives were employed to show either positive or negative meaning (e.g. beautiful/dirty, interesting/boring) and six adverbs expressing degrees were employed. Through both F0 observation of communicative speech with various F0 heights, we could have confirmed the following characteristics.

- There were consistent F0 differences between the communicative speech prosody and the read speech prosody depending on the adjective's attribute and degree of adverb
- Positive/negative adjectives give average F0 increase/decrease respectively to communicative prosody
- The magnitude of F0 increase/decrease is in proportion to the degree of adverbs as shown in Fig. 5.1

Moreover, we found that the above F0 control characteristics coincided with the perceptual naturalness evaluation of communicative speech using MOS (Mean Opinion Score). It turned out that a communicative F0 pattern could be obtained using the mapping from the subjective degree of the adverb to F0 control parameters. The effectiveness of the communicative prosody generation scheme has been confirmed by the perceptual experiment evaluating naturalness of synthesized speech.

This study showed the possibility of communicative prosody control using attributes of output lexicons, i.e., positive/negative of adjectives and degree of adverbs. Though there seem to be many factors affecting communicative prosody, some of them can be explained as a part of the linguistic attributes of lexicons constituting an utterance. This possibility of lexicon-driven communicative prosody generation has been generalized by our series of works using perceptual impressions to characterize communicative prosody (Kokenawa et al. 2005a, b; Greenberg et al. 2006, 2009a, b, 2010; Li et al. 2007a, b; Zhu et al. 2007).

5.3 Communicative Prosody Description Using Perceptual Impressions

We do not have any description framework to discriminate and quantify the differences of prosody observed in real world communication. Though existing prosody descriptions such as ToBI (http://www.ling.ohio-state.edu/∼tobi/) can be served as a description scheme of a prosody shape by itself, we do not have a description system to distinguish them as speech with some specific communicative role. We need a description scheme of communicative functions which cause these distinctions and directly relate to information manifested as prosodic differences. If we can successfully define the description of communicative prosody, we can analyze the mapping from the description to the prosody manifested in communicative speech. We have been suffering this communicative prosody specification problem for more than two decades till we found one way out; through the analysis of "uhm."

Though a single utterance, "uhm" does not have any particular lexical meaning by itself; its intonation can convey various kinds of communicative information. In order to treat the information conveyed by its communicative prosody, we proposed to employ perceptual impressions. Using "uhm" as a target, we can directly associate its prosodic characteristics manifested in F0 and duration with its perceptual impressions without being bothered by intrinsic lexical properties and linguistic structures (Kokenawa et al. 2005a).

Based on our observations of "uhm," 12 single word utterances that were controlled by three types of average F0 height (high, mid, low) and four types of F0 dynamics (rise, flat, fall, rise, and fall) were prepared as speech stimuli. After preliminary listening tests, we decided to use 26 word expressions for the description of perceptual impressions. These impression words can be classified into the three groups, *doubtful–confident* (doubt, ambivalence, understanding, approve), *unacceptable–allowable* (deny, objection, agreement) and *negative–positive* (dark, weakly, not interested, bad mood, heavy, bothering, audacious, anger, annoying, cheerful, delight, gentle, good mood, excited, happy, light, interested, bright) We asked fivesubjects to evaluate 12 "uhm" utterances in terms of 26 impression words with 8-level scaling, 0(not at all)–7(very much).

As easily guessed from these words employed for the description of impressions, they are redundant. To obtain a direct relation between impression attributes and prosody characteristics, we need more compact description by reducing this redundancy. In the field of metric psychology, it is well known that the dimension of this psychological space can be reduced using a mathematical procedure called multi-dimensional scaling (MDS). For the details of its mathematical framework and procedures, please consult with a textbook of MDS (e.g., Borg and Groenen 2010). Using MDS, the dimension of psychological metric space can be reduced to smaller dimensional space by preserving geometrical distances between samples expressed in this psychological space defined as Euclidian distance.

By applying MDS, the above perceptual impression space with 26 dimensions could be reduced to 3 dimensions (Kokenawa et al. 2005a). To interpret the meaning

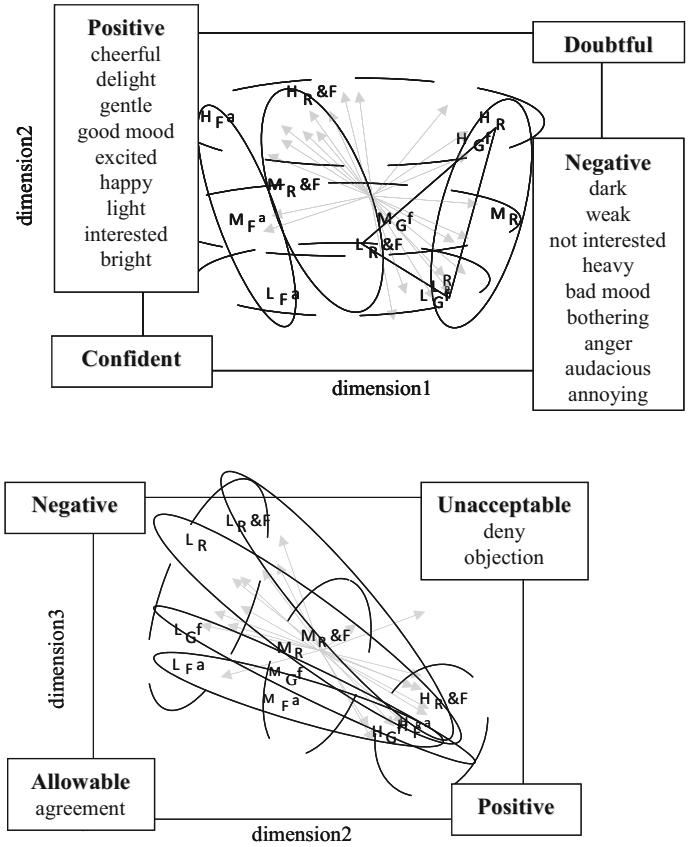


Fig. 5.2 Projection of test word vectors in three dimensional perceptual impression space obtained by Multi-Dimensional Scaling (*INDSCAL*) (Clusters of F0 height (*H,M,L*) and F0 dynamic patterns (*R,Fl,Fa,R&F*) are circled with *dash line* and *full line* respectively)

of three axes obtained by MDS, the average scores corresponding to each impression word were projected onto the three-dimensional spaces. In Fig. 5.2, we plotted the impression words that showed a high correlation between these three axes. As shown in this figure, we can approximate a set of impression words in three dimensions expressing the speaking attitudes of *positive–negative*, *confident–doubtful* and *allowable–unacceptable*. The axes of *confident–doubtful* and *positive–negative* can be projected on the plane spanned by the first and second dimensions. The axes of *allowable–unacceptable* and *confident–doubtful* can be interpreted in the plane spanned by the firstand third spaces. The axes of *allowable–unacceptable* and *positive–negative* are interpreted in the plane of the secondand third dimensions. These results nicely coincide with our intuitive grouping of the 26 basic expressions given in the previous paragraph and support the possibility of treatments of perceptual impressions by a restricted number of freedoms conveyed just by F0 average height and shapes.

5.4 Communicative Prosody Generation Using Impression Attributes of Constituent Lexicons

The studies in the previous section suggested further possibilities of impression attributes as input to specify communicative prosody. As "uhm" has no specific meaning by itself, if we want to generate a specific "uhm" with communicative prosody, we need to assign its impression, whereas, if we want to generate a word utterance directly associated to impressions observed in the "uhm" analysis, it may be possible to get those impressions directly from the lexicon itself. That is, the default impression might be obtained from the lexicon itself. To confirm this idea of the relationship between word impressions and communicative prosody, we have observed the communicative prosody of simple phrase expressions corresponding to a three-dimensional space.

We collected communicative speech data of 16 common Japanese phrases (*doubtful–confident*: doubt, ambivalence, understanding, approval; *unacceptable–allowable*: denial, objection, agreement, sympathy; *negative–positive*: dark, sad, not interested, heavy, bright, happy, interested, and light). We observed their prosodic characteristics in conversational speech (Kokenawa et al. 2005a). The correlation analyses showed that word impressions directly corresponding to three dimensions in a perceptual impression space had the same prosodic characteristics of "uhm" showing the corresponding impressions. The word attributes expressing *confident–doubtful; allowable–unacceptable* could be dependent on the difference of F0 dynamic patterns and duration, while those of *positive–negative* were highly related with the F0 height. These results showed the usefulness of the word impressions for communicative prosody generation.

These correlations between word impressions and communicative prosody characteristics have been confirmed not only as single word utterances but also as phase utterances consisting of multiple words with different impressions (Greenberg et al. 2010). From these observations, we can think of a communicative prosody generation scheme as shown in Fig. 5.3. As shown in this figure, input lexicons are used not only for the calculation of conventional prosody such as phrasing and phrase accents, but also for the calculation of communicative contributions. For the F0 generation, we employed the command-response model (Fujisaki and Hirose 1984) where conventional prosody control and the communicative one could be added ton its control parameter domain (Li et al. 2007a).

Figure 5.4 shows the distributional differences of two parameters: the magnitude of phrase/accent component Ap/Aa (for detailed explanation of them, please see the chap. 3 written by Prof. Mixdorff and chap. 4 by Dr. Kameoka) for different F0 shapes. As shown in these characteristics, superficial F0 shape differences can be well modeled by the differences of these parameters. Perceptual naturalness evaluation experiments on synthesized speech with communicative prosody have shown the usefulness of this communicative prosody control (Greenberg et al. 2006).

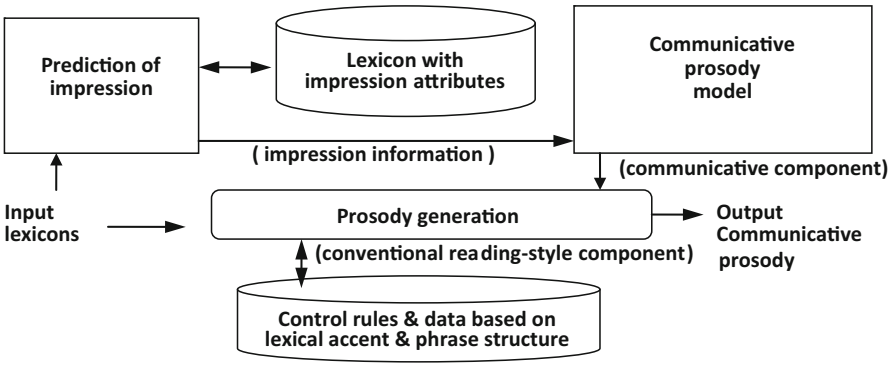
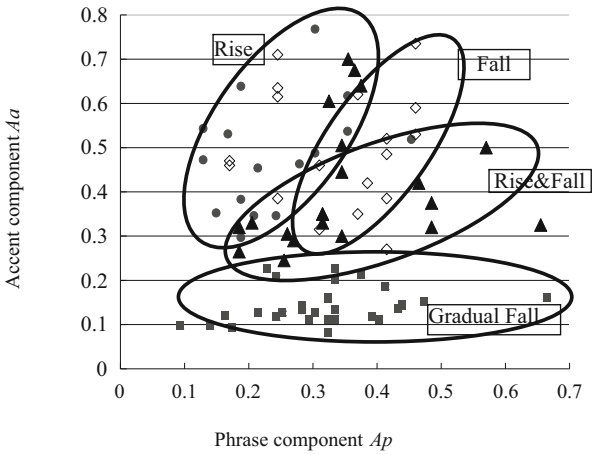


Fig. 5.3 Communicative prosody generation using impression prediction by input lexicons

Fig. 5.4 F0 generation parameters *Ap* and *Aa* for four typical patterns of *"uhm"*



5.5 Prosody Transfer as Pan-Linguistic Characteristics

From the impressions employed for communicative prosody control, we can easily guess the applicability of the proposed lexicon-driven communicative prosody control to other languages. To confirm the applicability of the proposed impression—prosody mapping to other languages, we carried out experiments on the communicative prosody generation for Mandarin and English phrases using the proposed generation scheme (Li et al. 2007b; Greenberg et al. 2009b). By using corresponding utterances of "uhm" uttered by a Japanese speaker, we extracted the communicative prosody component using the command response model. For both Chinese and English words, directly showing the impressions corresponding to six axes of three dimensions were selected.

By analyzing the read speech of these Chinese and English words using the command response model fitting, their word intrinsic prosody parameters were obtained

for the command-response model. By modifying each parameter using corresponding communicative prosody parameters extracted from "uhm" uttered by a Japanese speaker, the communicative speech samples were synthesized using STRAIGHT synthesis (Kawahara et al. 1999). Perceptual evaluation experiments of synthesized speech showed a remarkable increase in the naturalness of the synthesized communicative prosody samples. Though we could have confirmed the applicability of directly associated impressions derived from MDS analyses only to representative words, we could speculate the generality of this communicative prosody control.

5.6 Further Steps Toward Dialogue Prosody Control

As shown in above sections, we could have confirmed the possibility of communicative prosody control based on impression-prosody correlations of constituent lexicons. However, it is also true that the communicative prosody cannot be simply obtained from constituent lexicons. For example, compare the following two utterances in conversation.

- Their rooms were so dirty!
- Are their rooms dirty or clean?

Though the same word "dirty" is employed in these two utterances, their communicative prosody is not identical. The first example shows the speaker's negative opinion to the room using the lexicon "dirty." On the other hand, "dirty" in the second one has one of the possibilities of the status of the room and does not indicate the speaker's opinion. As shown by the differences in these examples, we have to pay attention to the communicative function of the utterance by itself, not only sticking to the impression attributes of constituent words.

In the field of spoken dialogue, statistical models have been proposed to identify the dialogue act and its structure (Young 2009, Young et al. 2010; Quarteroni et al. 2011). Free from taxonomic classification and identification problems of dialogue acts, these statistical dialogue models have the potential to provide a soft description of the utterance act and are expected to be useful for dialogue prosody control. More intensive systematic studies are expected to control dialogue prosody to properly reflect the dialogue act.

5.7 Conclusions

In this chapter, we have introduced prosody variation modeling for communicative prosody characterization. Through the analyses to formulate communicative prosody in terms of its impressions as a descriptor and the corresponding prosody, we have proposed the first approach to handle communicative prosody. Using an F0 control scheme employing the command-response model, we could have confirmed that the

communicative prosody could be generated from constituent lexicons. We believe that this trial demonstrates two important research possibilities in the near future.

The first possibility is a possibility of communicative prosody generation for speech synthesis. Unlike conventional synthesis of pre-fixed categories such as emotional speech, we can expect more detailed and context-sensitive speech output from input lexicons. Like computational modeling of semantics related issues in the field of natural language processing, we will be able to employ NLP related techniques such as the vector space method (Turney and Pantel 2010) to effectively use information embedded in lexicons for synthesis.

The second possibility is a general expansion of research targets in linguistics and phonetics. Some linguists have already awarded the importance of linguistic formulation considering the type of narrators and communication situations (Sadanobu 2012; Teshigawara and Kinsui 2011). By integrating these research efforts together with phonetics, we believe that traditional linguistics and phonetics should cover much wider viewpoints where not only unidirectional written sentences, but also bi-directional spoken language can be formulated. To properly treat communicative information exchanges, the information transferred from a writer/speaker to a reader/listener are to be analyzed by considering their information sending/receiving contexts.

From this viewpoint, spoken language research topics conventionally excluded as *paralinguistics* should be reconsidered as core elements of communicative linguistic and phonetics. We believe that information technology will be able to provide useful methodology which could be useful not only in engineering applications, but also in a scientifically well-defined theoretical framework as computational communicative modeling.

Acknowledgements The authors would like to express sincere thanks to many collaborators. In particular, Hiroaki Kato, Minoru Tsuzaki, Takumi Yamashita, Ming Zhu, Ke Li, for their original contributions and assistance. As shown in the references, this paper consists of their original works and gives a unified view underlying these works. Moreover, the authors would like to give special thanks to Toshiyuki Sadanobu and the member of his group supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research (A), 23320087 which has given us the opportunity to think over our works in terms of linguistics and phonetics.

References

Borg, I., and P. J. F. Groenen. 2010. *Modern multidimensional scaling: Theory and applications*. Springer Series in Statistics.

Fujisaki, H., and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan* 5 (4): 233–242.

Greenberg, Y., M. Tsuzaki, K. Kato, and Y. Sagisaka. 2006. A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech. *Proceedings of the Speech Prosody*, pp. 37–40.

Greenberg, Y., N. Shibuya, M. Tsuzaki, H. Kato, and Y. Sagisaka. 2009a. Analysis on paralinguistic prosody control in perceptual impression space using multiple dimensional scaling. *Speech Communication* 51 (7): 585–593.

Greenberg, Y., M. Tsuzaki, H. Kato, and Y. Sagisaka. 2009b. Communicative prosody generation using language common features provided by input lexicons. *Proceedings of the SNLP2009*, pp. 101–104.

Greenberg, Y., M. Tsuzaki, H. Kato, and Y. Sagisaka. 2010. Analysis of impression-prosody mapping in communicative speech consisting of multiple lexicons with different impressions. *Proceedings of the O-COCOSDA*.

Kawahara, H., I. Masuda-Katsuse, and A. Cheveignè. 1999. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27:187–207.

Kokenawa, Y., M. Tsuzaki, H. Kato, and Y. Sagisaka. 2005a. F0 control characterization by perceptual impressions on speaking attitude using multiple dimensional scaling analysis. *Proceedings of the IEEE ICASSP*, pp. 273–275.

Kokenawa, Y., M. Tsuzaki, K. Kato, and Y. Sagisaka. 2005b. Communicative speech synthesis using constituent word attributes. *Proceedings of the INTERSPEECH*, pp. 517–520.

Li, K., Y. Greenberg, N. Campbell, and Y. Sagisaka. 2007a. On the analysis of F0 control characteristics of non-verbal utterances and its application to communicative prosody generation in NATO security through science series E: Human and societal dynamics vol. 8. The fundamentals of verbal and non-verbal communication and the biometric issue, pp. 179–183, IOS Press.

Li, K., Y. Greenberg, and Y. Sagisaka. 2007b. Inter-language prosodic style modification experiment using word im-pression vector for communicative speech generation. *Proceedings of the INTERSPEECH*, pp. 1294–1297.

Quarteroni, S., A. V. Ivanov, G. Riccardi. 2011. Simultaneous dialog act segmentation and classification from human spoken conversations. *Proceedings of the IEEE ICASSP*, pp. 5596–5599.

Sadanobu, T. 2012. An unofficial guide for Japanese characters. http://dictionary.sanseido-publ.co.jp/wp/author/sadanobu-e/.

Sagisaka, Y., T. Yamashita, and Y. Kokenawa. 2005. Generation and perception of F0 markedness for communicative speech synthesis. *Speech Communication* 46 (3–4): 376–384.

Sapir, E. 1921. *Language: An introduction to the study of speech*. New York: Harcourt, Brace and Company.

Teshigawara, M., and S. Kinsui. 2011. Modern Japanese "Role Language" (Yakuwarigo): Fictionalised orality in Japanese literature and popular culture. https://www.equinoxpub.com/journals/index.php/SS/article/view/7911.

Turney, P. D., and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Young, S. 2009. Cognitive user interfaces: An engineering approach. *Plenary talk in IEEE ICASSP*.

Young, S., M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech and Language*, 24 (2): 150–174.

Zhu, M., K. Li. Greenberg, and Y. Sagisaka. 2007. Automatic extraction of paralinguistic information from communicative speech. *Proceedings of the 7th symposium on natural language processing*, pp. 207–212.

http://www.ling.ohio-state.edu/∼tobi/.

Chapter 6
Mandarin Stress Analysis and Prediction for Speech Synthesis

Ya Li and Jianhua Tao

Abstract Expressive speech synthesis has recently received much attention. Stress (or pitch accent) is the perceptual prominence within words or utterances, and is one important feature in forming the highs and lows of the pitch contour, which makes the speech sounds more expressive. In this chapter, we introduce a large-scale stress annotated continuous Mandarin corpus. Then the stress distribution and its stability are thoroughly analyzed from aspects of rhythm level and tone pattern. Based on these results, we propose a novel hierarchical Mandarin stress modeling method. The top level emphasizes stressed syllables, while the bottom level focuses on unstressed syllables for the first time due to its importance in both naturalness and expressiveness of synthetic speech. We also carried out several experiments to assign the Mandarin stress from textual features by using the classification and regression tree (CART) and maximum entropy (ME) model respectively. The work could be beneficial to speech synthesis systems for generating high natural and expressive speech.

6.1 Introduction

Prosody is a super-segmental feature of speech and consists of rhythm, stress, and intonation, among which stress is a hot topic in recent years due to the growing demand for expressive text-to-speech (TTS). Stress (in English, the "pitch accent" or "accent" is often used with the similar sense to the "stress" in Mandarin) is the perceptual prominence within words or utterances. Mandarin stress can be categorized as word stress and sentence stress, and there are three levels of word stress, stressed, regular, and unstressed. When a syllable is stressed, its pitch goes higher and the duration becomes longer. It serves as one important feature in forming the ups and downs in a pitch contour, which makes the speech sound more expressive. However, Mandarin stress processing is a complicated problem. The difficulties lie in

J. Tao (✉) · Y. Li
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China
e-mail: jianhua.tao@gmail.com

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_6

three aspects. First, Mandarin stress perception and labeling are tough work. In English, ToBI (Silverman et al. 1992) is a well-accepted annotation system for prosody patterns of utterances. It defines prominence from the word's pitch movements or configurations and is easy to practice. Nevertheless, this is not the case for tonal language, such as Mandarin, in which each syllable has a tone and a relative steady pitch contour. Due to the frequent perceptual conflict among tone, intonation, and stress, it is quite tough to find an exact definition of Mandarin stress, which makes stress labeling more difficult compared to other corpus labeling work. Li et al. (2000) build a read corpus ASCCD with four-level stress annotation, and they report that the consistency in stress labeling is about 66 %. However, the corpus is a discourse corpus and is not phonetically balanced, so it is not suitable for TTS. Wang et al. (2003b) have annotated stress for 300 utterances selected from the Microsoft large-scale TTS speech corpus; nevertheless, the 300 utterances are insufficient for some machine learning methods for stress processing. The second obstacle for Mandarin stress processing is that the performance of automatic stress prediction from raw text is not up to our expectation. Although we can utilize acoustic features of speech or combine textual and acoustic features together to build an excellent stress detector (Shao et al. 2007; Ni et al. 2009; Ananthakrishnan and Narayanan 2008), a prediction module just using textual features is still required for a TTS system. Shao et al. (2007) utilized an artificial neural network (ANN) model to predict Chinese sentential stress using acoustic features, textual features, and both of them. The F-score for predicting a stressed syllable using just the text-based ANN model is only 36.1 %. Ni et al. (2009) compared the performance of the support vector machine (SVM), classification and regression tree (CART), and AdaBoost with CART model for Mandarin stress prediction with acoustic features, textual features, and both of them. They argue that AdaBoost with CART can achieve more favorable results than a single decision tree, and SVM is inferior to the CART model. Similarly, Wightman (Wightman and Ostendorf 1994) and Hirschberg (1995) adopted decision trees to predict pitch accent in English and obtain encouraging results, which are above 80 % in precision. The third bottleneck for Mandarin stress processing is how to balance perceptual prominence and naturalness in synthesizing speech even if we can assign stress from textual features appropriately. Early research on Mandarin stress generation are based on a concatenation system, and the primary drawback is that the expressiveness of synthetic speech still relies on the audio corpus they used. The growing demand for expressive speech synthesis forces us to seek an alternative technique. An HMM-based speech synthesis system can overcome this drawback by easily modifying the prosodic parameters of synthetic speech. Much research has already successfully been done on expressive speech synthesis with HMM-based TTS. Yamafishi et. al. (2004) reported their recent progress in generating Japanese expressive speech synthesis. The idea of their work is that prosodic features and spectral features should be controlled properly to model expressive synthetic speech. They use speaking style interpolation and adaptation for HMM-based speech synthesis. In the style adaptation, the maximum likelihood linear regression is often adopted. Yu et. al. (2010) utilize two decision tree models and extract word-level emphasis patterns from natural English speech, and then embed the emphasis model in an HMM-based speech synthesis framework. They argue that due to the weakness of

emphasis cues, directly using emphasis context features and the traditional adaptation method does not work well. Badino et.al. (2009) automatically detect contrastive word pairs with textual features only and use enhanced context dependent labels to synthesize the emphasis. They point out that their methods can convey contrast information effectively. The drawback lies in that the realization of emphasis turned out to be occasionally strong and therefore less contextually appropriate.

In this chapter, we try to address the first two problems in Mandarin stress processing. First, we built a large scale stress annotated corpus. Then the stress distribution and its stability are thoroughly analyzed from aspects of rhythm level and tone pattern. Based on these results, we propose a novel hierarchical Mandarin stress modeling method. The top level emphasizes stressed syllables, while the bottom level focuses on unstressed syllables for the first time due to its importance in both naturalness and expressiveness of synthetic speech. We also carried out several experiments to assign the Mandarin stress from textual features by using the CART and maximum entropy (ME) model respectively. The results show that we can get a relative reliable stress assignment automatically. The work could be beneficial to speech synthesis systems for generating high natural and expressive speech.

6.2 Mandarin Stress Perception and Analysis in Continuous Speech

6.2.1 Corpus Construction

The audio corpus used in this work contains 6000 phonetic balanced sentences (about 73,000 syllables), which are first designed for speech synthesis and read by a professional female speaker. All the utterances are segmentally labeled according to the audio data by research assistants.

For the stress labeling, we have tracked many prosody transcribing systems. In English, ToBI (Silverman et al. 1992) is a well-accepted annotation system for prosody patterns. Many ToBI-like labeling system are proposed for other languages. C-ToBI (Li 2002) and Pan-ToBI (Tseng and Chou 1999) are designed for Mandarin. However, ToBI-like labeling principle is hard to master, and therefore the labeling consistency cannot be guaranteed. Therefore, we simplified the stress labeling in this work, and label the streets only according to the prominence we perceived. Three levels of stress are adopted here, namely, stressed, regular, and unstressed syllables according to their prominence degrees within a prosodic word (for word stress labeling) or an utterance (for sentence stress labeling).

During the stress labeling, three assistants were first trained with a subset corpus several times. A small percentage of disagreement is acceptable due to the frequent perception confliction among tone, intonation, and stress. The aim of training is to keep the consistency of each annotator with their own during the whole annotating process and among annotators as much as possible. In the word stress labeling, we segmented the utterance into prosodic words and stored them according to their tone patterns separately to reduce the impact of the surrounding syllables on the

Table 6.1 Stress distribution in the final corpus

Stress category	Stress degree distribution (%)		
	Unstressed	Regular	Stressed
Sentence stress (in prosodic words)	0.22	0.54	0.24
Word stress (in syllables)	0.04	0.48	0.48

perception of the current syllable. In the sentence stress labeling, we assign stress to each word according to its prominence degree within an intonation phrase. We adopt an intonation phrase instead of an utterance in sentence stress labeling for two reasons. First, pitch is reset at an intonation phrase boundary, and to compare the prominence degree within an intonation phrase can reduce the impact of intonation change. Second, it is hard for listeners to distinguish the prominence difference in a long sentence due to the short-term memory limit, and thus the labeling consistency cannot be guaranteed. The sentence stress labeling consistencies before and after training are 42 % and 65 % of three annotators. The word stress labeling consistency is above 72 %.

The corpus construction lasts for 3 months. The average stress degree of three annotations was used as the ground truth in the text-based stress prediction model training and testing process. In the ground truth data, the percentages for unstressed, regular, and stressed syllables shown in Table 6.1.

6.2.2 Word Stress Analysis

It is claimed that when a syllable is stressed, probably, the duration goes longer and the pitch range goes wider than when it is unstressed. With this large scale stress annotated corpus, we may find more evidences for this claim, and hopefully, we can obtain a specific prosodic parameter variation pattern when a syllable is stressed or not.

As a tonal language, Mandarin stress is highly related with tone, rhythm level, and intonation in continuous speech. The coarticulation among syllables is also an important factor. Therefore, we adopted the tone pattern category from Xu (1994) which categorized tone patterns as "compatible" and "conflicting" patterns. The compatible context is an environment that the adjacent syllables have similar pitch value. While in a conflicting context, the adjacent syllables have very different pitch values. Tone three was eliminated due to its strong coarticulation in continuous speech. Figure 6.1 shows the tone patterns of disyllables and trisyllables (Quadrisyllables are rare in the corpus). The prosodic boundaries are categorized as syllable boundary (denoted as b0), prosodic word boundary (denoted as b1), prosodic phrase boundary (denoted as b2), and intonation phrase boundary (denoted as b3). Through this method, we defined 120 syllable context categories, i.e., 5 (tone categories) $\times$ 4 (prosodic

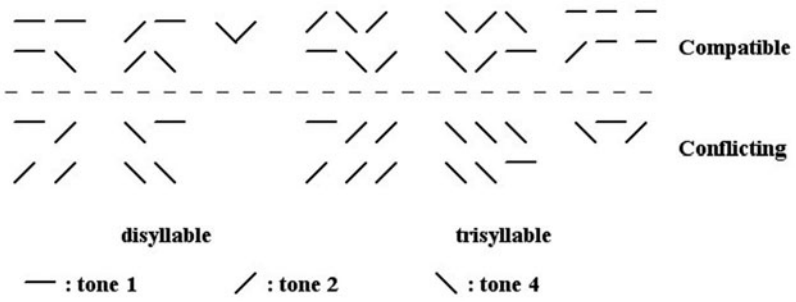


Fig. 6.1 Tone patterns of *disyllables* and *trisyllables*

boundary categories) × 3 (stress levels) × 2 (tone pattern categories). The normalized pitch, including the mean pitch, pitch range as well as the normalized duration of all the syllables were statistically summed and then averaged according to their context categories.

6.2.2.1 Prosodic Parameter Analysis Methodology

Two normalization methods for syllable duration are adopted and listed below.

$$d_1 = d/d_{PSDWord} \tag{6.1}$$

$$d_2 = d/d_{Final} \tag{6.2}$$

d_2, d_2 represent two normalized syllable duration values by two methods. d is the syllable's real duration. $d_{PSDWord}$ is the average syllable duration of each prosodic word, and d_{Final} is the average duration of the syllables with the same finals. Equation 6.1 represents the relative duration within a prosodic word, while Eq. 6.2 represents the duration change of the syllable itself in different stress levels.

For pitch normalization, we have adopted three methods. In Eq. 6.3–6.5, p_1, p_2, and p_3 are the three normalized mean pitch values of each syllable. pis the real mean pitch, and $p_{PSDWord}$ is average mean pitch per syllable of each prosodic word. p_1 represents the relative mean pitch within a prosodic word, and the unit for p_1 and $p_{PSDWord}$ is Hz. p_2 and p_3 are semitones which is a better measurement for intonation analysis (Li 2005). p_{ref} is set as 20 Hz, and p_{hw}, p_{lw} are the maximum and minimum pitch in a prosodic word.

$$p_1 = p/p_{PSDWord} \tag{6.3}$$

$$p_2 = 12\log(p/p_{ref})/log(2) \tag{6.4}$$

$$p_3 = 12\log(p/p_{lw})/log(p_{hw}/p_{lw}) \tag{6.5}$$

We also calculated the maximum pitch, minimum pitch and the pitch range in semitones, denoted as p_h, p_l, and p_{range}, respectively.

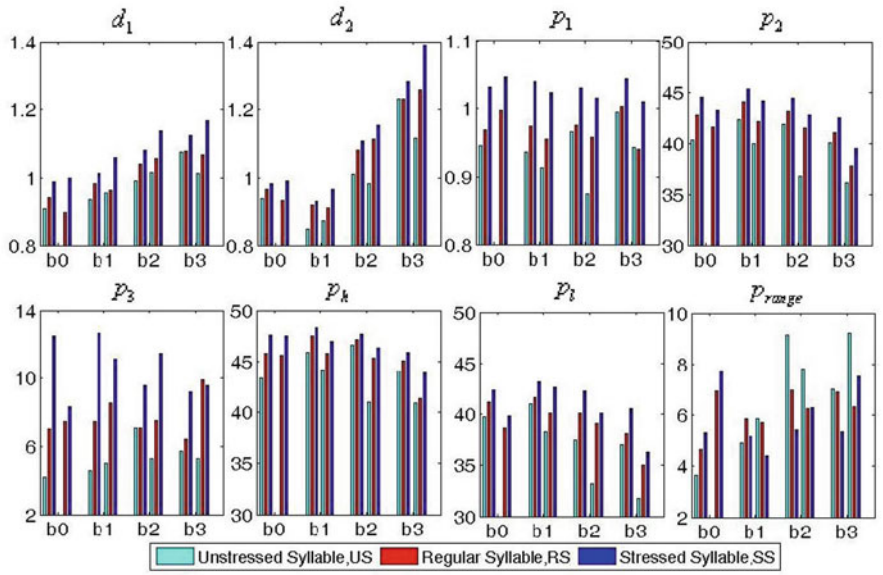


Fig. 6.2 Normalized prosodic parameters in different syllable contexts. (*b0*, *b1*, *b2*, and *b3* denote the syllable, prosodic word, prosodic phrase and intonation phrase boundaries respectively)

6.2.2.2 Prosodic Parameter Analysis Results

Figure 6.2 illustrates the above normalized prosodic parameters in different phrase boundaries and tone patterns, in which, the unstressed syllables' prosodic parameters in compatible context are not included because the sample amount is too low. The results show that the perceptual difference of Mandarin word stress on the different rhythm levels and tone patterns has obvious regularities:(1) Both the syllables' duration goes longer and the pitch goes higher when the syllable is stressed as its rhythm level increases; in addition, these prosodic differences increase when the prosodic boundary goes higher; (2) the difference of d_2 in different syllable contexts is more noticeable than d_1, which indicates that d_2 can convey the duration variation when a syllable's stress level change more clearly; (3) Starting from prosodic word boundary, p_2, p_h, and p_l decrease when the prosodic boundary goes higher. This verifies the pitch-decline trends in statements.

We also calculate the correlation coefficients of the normalized prosodic parameter and stress level in different syllable contexts, in which p_1 is not included, because p_1 is ratio scaled, and the difference cannot reflect the difference perceived by the human ear (Li 2005). The results show that pitch has greater impact on stress perception, and stress perception is more influenced by duration in a conflicting tone pattern context than that in a compatible tone pattern. (For detailed results, please refer to (Li 2012)).

6.3 Hierarchical Mandarin Stress Modeling

Over the past decades, researchers have paid a lot of attention to stressed syllables, but up to now it is still a controversial question and hasn't been well solved. To name a few, first, the three classification prediction models are not as good as expected; Second, if two-level of stress is adopted, namely sentence stress and word stress, and in each level, three-prominence degree is adopted, the whole prediction model becomes complicated. Third, the conventional "stressed" and "nonstressed" category regardless of word level or sentence level cannot fully represent the undulation of the pitch contours. Therefore, we have to take a closer look at the nature of Mandarin stress.

6.3.1 Sentence Level Stress

Sentence stress is the prominence within a sentence, generally, it is assigned to prosodic word firstly, and then obtained by a syllable in the prosodic word.

From our experience, when hearing a whole utterance, people intuitively recognize a few prominent syllables, and then they will compensate for other less prominent or noise-masked syllables and phonemes using their knowledge of the spoken language, which is similar to the phonemic restoration effect (Timothy 2002). The top of these stressed syllables' pitch contours are often significantly higher in the whole utterance. Therefore, the sentence-level stressed syllables should be intensively investigated for speech technology, especially for speech synthesis, which aims to produce human-like natural speech.

6.3.2 Word Level Stress

In the scope of prosodic word, by contrast, the prominence difference within syllables does exist, but is not so distinct.

The growing demand for language learning and speech technology forces researchers to switch their focus to nonstressed syllables, since in daily life, people tend to pronounce words or syllables with little effort, thus bringing about more and more weak syllables in real speech. These weak syllables have lower pitch and shorter duration, and are also important to form the ups and downs in pitch contours. Nowadays there are approximately 20 % weak syllables in real Mandarin speech, and even more in Beijing dialect.

The terms "unstressed syllable" and "neutral-tone syllable" are often wrongly treated as interchangeable, because both of them are weak syllables. To clarify, the unstressed syllable in this work is slightly different from the neutral tone syllable. Yuenren Chao thinks that there are static weakened syllables and provisional weakened syllables. Most auxiliary and suffix words, e.g., "的 (de5) ", "了 (le5) ", "们 (men5) ", which have small semantic value or a purely grammatical function

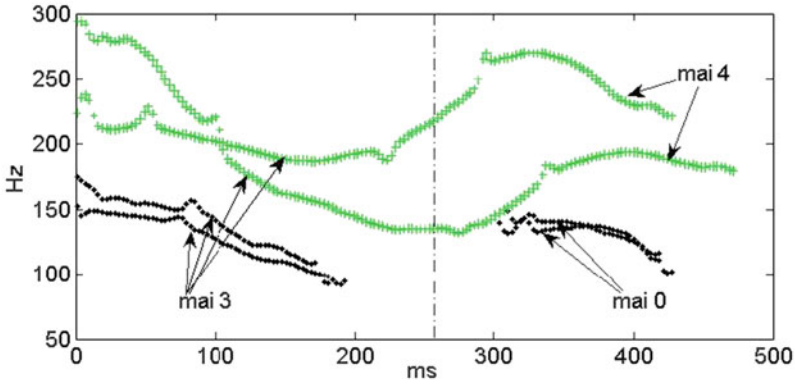


Fig. 6.3 The pitch contours of "买卖" in natural speech

are steadily weakened in speech and are defined as the neutral tone syllable (Chao 1968). The neutral tone syllables in Mandarin do have some relatively stable patterns in acoustic realization (Li et al. 2010).

However, it is not the case for the unstressed syllable, whose acoustic realization varies as the context changes. In addition, some may change the word sense when unstressed. The following example demonstrates the influence of an unstressed syllable in word sense and their prosodic feature.

我做的是小买卖。(What I do is a small trade.)

wo3 zuo4 de5 shi4 xiao3 mai3 mai0.

买卖双方都同意这个安排。(Both the buyer and the seller agree to this arrangement.)

mai3 mai4 shuang1 fang1 dou1 tong2 yi4 zhe4 ge4 an1 pai2

Figure 6.3 shows four pitch contours of the content word "买卖" in natural speech, among which two "卖" labeled as "mai0" are unstressed, and the other two labeled as "mai4" are not. The four "买" are not unstressed. We can see that the acoustic features can change a lot when a syllable is unstressed[1]. The pitch is dropping, the pitch range becomes narrower and duration becomes shorter. They have more than one type of pitch contours and should be synthesized accordingly in a TTS system. Otherwise, the utterance sounds unnatural and the meaning of the whole sentence may change.

6.3.3 Two-Level Stress Modeling

In summary, people only address a small number of syllables to make them significantly prominent in the upper level (e.g., sentence level), and these upper-level

[1] Because of the coarticulation, the acoustic realization of surrounding syllables are also changed.

stressed syllables greatly improve the speech expressiveness. Meanwhile, at the lower level (e.g., word level) the prominence syllable is restricted by the upper level and cannot be distinctively perceived, thus people weaken other syllables to make that syllable stand out. Through such a mechanism, the speech sounds more expressive and without too much physical effort. Additionally, people sometimes use an unstressed syllable to express different meanings. Based on this investigation, we proposed a novel two-level Mandarin stress modeling method, in which word level unstressed syllable investigation is emphasized for the first time, and stressed syllables in the sentence level are studied as per traditional methods. This modeling method emphasizes different prosodic units modeling according to their functions, and it can offer a more precise model because each layer is a binary classification model and can generally provide a better performance than ternary classification, and thus will not bring too much accumulative error.

6.4 Stress Prediction from Text

Automatic stress prediction from textual features has been widely studied due to its critical role in TTS. Many statistical machine learning techniques are introduced to this task.

6.4.1 Multiple Textual Features

We choose the ME model to predict stress in this work which has been successfully applied to many prosody information predictions. The ME model seeks the probability distribution with the ME subject to certain constraints. Such constraints force the model to match its feature expectations with those observed in the training data. Therefore, selecting the most effective and discriminative features can greatly improve the performance of the ME model. Hirschberg (Hirschberg 1995) indicates that though the detailed syntactic, semantic, and discourse level information can enhance the prediction of pitch accent, it is indeed possible to get a fair success with some automatically extracted features. Hence, only the shallow grammatical information which could be easily and reliably acquired from raw text is considered in this work. The atomic text feature templates used in this work are listed in Table 6.2.

We also combined the above atomic feature templates to get more sophisticated feature templates and the sliding window method was adopted in feature extraction. For instance, the feature template "B_1&B0&B1" denotes the previous, current, and the next syllable prosodic boundaries and the number after the feature templates indicates the window offset. For the upper level stress prediction, we set the window width to five to capture the long distance relation. In lower level stress prediction, the window width is set to three. We also use a wrapper method for selecting the

Table 6.2 Atomic textual features used in stress prediction

Phonetic features	PINYIN transcript (PY) and tone identity (T)
	Syllable's prosodic boundary (B)
	Part-of-Speech and the length of a word (P, L)
	Syllable description and the word, prosodic word descriptions (C, W, PW)
	Prosodic word length (PL)
Position features	Normalized index of the current syllable in the prosodic word (RPW)
	Index of the syllable in current word and current prosodic word (IW, IPW)
Distance features	Distance from current syllable to the previous and the next word (DPW, DNW)
	Distance from current syllable to previous and next prosodic phrase (DPP, DNP)
	Distance from current syllable to the beginning and the end of the utterance (DB, DE)
	Distance from current word to the beginning and the end of the utterance (DBW, DEW)
	Distance from current prosodic word to the beginning and the end of the utterance (DBPW, DEPW)
Prior stress features	The stress ratio of the current syllable (SRC)
	The stress ratio of the current prosodic word (SRW)

effective feature templates to achieve better performance. The stop criterion is that the improvement of the average correct rate is less than 0.1 %.

Regarding the training corpus, the word level stress prediction model only utilizes the word level stress annotated data and the combined data, which having multiplied the two level stress annotation and then discretized (Li et al. 2011), is finally used in sentence stress prediction. The ratio of the training set and testing set is 9:1. In the training process, the training corpus is divided into ten parts and a tenfold cross validation is conducted. In this work, two ME prediction models are binary, namely, stressed and non-stressed classification in sentence level stress prediction, and unstressed and non-unstressed classification in word level stress prediction.

6.4.2 Experiments and Discussions

The first experiments are conducted to select the most effective feature templates for word and sentence stress predictions. The two baseline systems utilize all the atomic feature templates mentioned above. The average correct rates of two models before and after template selection are given in Table 6.3. Although it is possible to combine the templates with prior knowledge to get more sophisticated feature

Table 6.3 Average correct rate of two stress prediction models before and after feature selection (FS)

Experiment	Before FS (%)	After FS (%)
Word level	92.7	94.1
Sentence level	70.1	75.9

Table 6.4 Selected feature templates for word-level unstressed syllable and sentence-level stressed syllable prediction

	Word level	Sentence level
1	B0&DEW	B0&SRW0
2	B1&PW0	B1&T0&T_1
3	DB&SRW0	C0&RPW0
4	DNP0&DNP1	DBPW&SRC0
5	DNP0&PL0	DNP_2&PW0
6	DNP_1&PL0	DPP_1
7	DNP_1&P_1	PL2&PW0
8	PL1&SRC0	T0&T_1
9	T0	P1&PY1
10	T0&T1	P_1&T1

Table 6.5 Hierarchical stress prediction performance

Experiment	Precision (%)	Recall (%)	F-score (%)
Word level	94.2	67.6	78.7
Sentence level	73.1	73.4	73.3

templates, automatically selecting can achieve better performance with fewer templates. The original atomic templates (e.g., B_1, B0, B1) are more than 80, and after two independent template selections only ten templates are selected for each stress prediction model. Table 6.4 shows the detailed selected templates.

The second experiments are performed to evaluate the performances of concerned stress types in each prediction model, and the results are shown in Table 6.5. For sentence level stress prediction, the F-score of stressed syllables is 73.3 %, and for word level stress prediction, the F-score of unstressed syllable is 78.7 %. In (Li et al. 2010), we also use CART model to predict unstressed syllable, and the precision, recall, and F-score are 86.3, 56.3, and 68.1 % respectively. It implies that ME model performs better than CART in text-based stress prediction. As for the comparison between our results with the previous Mandarin stress prediction studies, no direct comparison could be presented because all other related works only perform a single layer sentence stress prediction, not to mention the difference in the corpus. In (Ni et al. 2009), Ni et al. construct several sentence stress prediction models and the best overall correct rate, which using a Boosting method with CART is above 80 %. However, the corpus they used, ASCCD, is a discourse corpus and there is no detailed information of stressed syllables prediction results in Ni et al. (2009).

The two level stress labels are helpful to describe the long term and short term prosodic characteristics of each syllable as far as possible. In addition, the experimental performances demonstrate the feasibility of the proposed hierarchical stress modeling. With the relatively ideal stress assignation, the next step for stress generation in a TTS system could be carried out. Fortunately, the latest pitch modeling research which utilizes a similar hierarchical approach for generating natural speech (Qian et al. 2008; Zen and Braunschweiler 2009; Lei et al. 2010) has already shown its effectiveness. The future work will take the advantage of the proposed hierarchical stress modeling (serve as the front-end of a TTS system) as well as the hierarchical prosodic parameter generation (Qian et al. 2008; Zen and Braunschweiler 2009; Lei et al. 2010) (serve as the back end of a TTS system) to get more natural and expressive speech.

6.5 Conclusions

The ultimate goal of this work is to synthesize human-like speech with stress, thus making the synthetic speech more expressive. In this chapter, we introduced a large scale stress annotated corpus, and carried out a thorough prosodic parameter analysis in various aspects. Based on these investigations, we proposed a hierarchical stress modeling method to get a fine-grained stress description. In the model, sentence level stressed syllables were studied as traditional methods are, while in the word level, we emphasized unstressed syllable research for the first time due to its importance in both naturalness and expressiveness. According to this architecture, a two-level stress prediction model under the ME framework was constructed. Experiments showed that the proposed method could obtain a fine-gained stress structure description reliably. This hierarchical stress model could be further integrated into the TTS system to generate more expressive speech.

Acknowledgements This work is supported by the National Natural Science Foundation of China (NSFC) (No. 60873160, 61011140075, 90820303, 61273288, 61233009, 61203258, and 61305003), and partly supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM (CSIDM) Programme Office.

References

Ananthakrishnan, S. and Narayanan S. S. 2008. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transaction on Audio, Speech, and Language Processing* 16 (1): 216–228.

Badino, L., J. S. Andersson, J. Yamagishi, and R. A. J. Clark. 2009. Clark Identification of contrast and its emphatic realization in HMM-based speech synthesis. INTERSPEECH, Brighton, 520–523.

Chao, Y. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.

Hirschberg, J. 1995. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence* 63:305–340.

Lei, M., Y. Wu, F. K. Soong, Z. Ling, and L. Dai. 2010. A hierarchical f0 modeling method for HMM-based speech synthesis. INTERSPEECH, Chiba, 2170–2173.

Li, A. 2002. Chinese prosody and prosodic labeling of spontaneous speech. Speech Prosody 2002, International Conference, Aix-en-Provence, France.

Li, A. 2005. The psycho-acoustic units for intonation study, report of phonetic research, 13–17.

Li, Y. 2012. Research on hierarchical analysis and prediction of stress in Chinese. Doctoral Diss., Chinese Academy of Sciences.

Li, A., X. Chen, G. Sun, W. Hua, Z. Yin, Y. Zu, F. Zheng, and Z. Song. 2000. The phonetic labeling on read and spontaneous discourse corpora. ICSLP: 724–727.

Li, Y., J. Tao, M. Zhang, S. Pan, and X. Xu. 2010. Text-based unstressed syllable prediction in Mandarin. INTERSPEECH, Chiba, 1752–1755.

Li, Y., J. Tao, and X. Xu. 2011. Hierarchical stress modeling in Mandarin text-to-speech. INTERSPEECH, 2013–2016.

Ni, C., W. Liu and Xu B. 2009. Mandarin pitch accent prediction using hierarchical model based ensemble machine learning. IEEE Youth Conference on Information, Computing and Telecommunication, YC-ICT '09. Beijing, 327–330.

Qian, Y., H. Liang and Soong F. K. 2008. Generating natural f0 trajectory with additive trees. INTERSPEECH, 2126–2129.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, Price, P. J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labeling English prosody. Proceedings of the ICSLP, 867–892.

Shao, Y., J. Han, Y. Zhao, and T. Liu. 2007. Study on automatic prediction of sentential stress for Chinese Putonghua text-to-speech system with natural style. *Chinese Journal of Acoustic* 26 (1): 49–92.

Timothy, B. J. 2002. *The psychology of language*. New Jersey: Prentice Hall.

Tseng, C., and F. Chou. 1999. A prosodic labeling system for Mandarin speech database. Proceedings of the XIV International Congress of Phonetic Science, 2397–2382.

Wang, Y., M. Chu, and L. He. 2003a. Location of sentence stresses within disyllabic words in Mandarin. Proceedings of the 15th ICPhS, Barcelona, 1827–1830.

Wang, Y., M. Chu, and L. He. 2003b. Labeling stress in continuous mandarin speech perceptually. Proceeding of the 15th International Congress of Phonetic Science, Barcelona, 2095–2098.

Wightman, C. and Ostendorf M. 1994. Automatic labeling of prosodic patterns. *IEEE Transaction on Speech and Audio Processing* 2(4): 469–481.

Xu, Y. 1994. Production and perception of coarticulated tones. *Journal of the Acoustical Society of America* 95:2240–2253.

Yamafishi, J., T. Masuko, and T. Kobayashi. 2004. HMM-based expressive speech synthesis-towards TTS with arbitrary speaking styles and emotions. Special Workshop in Maui (SWIM).

Yu, K., F. Mairesse and Young S. 2010. Word-level emphasis modeling in HMM-based speech synthesis. ICASSP, Dallas, 4238–4241.

Zen, H., and N. Braunschweiler. 2009. Context-dependent additive log f0 model for HMM-based speech synthesis. INTERSPEECH, Brighton, 2091–2094.

Chapter 7
Expressivity in Interactive Speech Synthesis; Some Paralinguistic and Nonlinguistic Issues of Speech Prosody for Conversational Dialogue Systems

Nick Campbell and Ya Li

Abstract This chapter explores the role of prosody in expressive speech synthesis and goes beyond present technology to consider the interrelated multimodal aspects of interactive spoken dialogue systems for human–machine or human–human interaction. The chapter stresses that social aspects of spoken dialogue are now ripe to be considered in the design of interactive systems and shows how three modalities can be combined—utterance content, speech expressivity, and facial or bodily gestures—to express social factors and manage the interaction. Linguistic prosody has been well described in the literature but the social aspects of managing a spoken dialogue remain as a next-step for speech synthesis research. This chapter shows how voice quality functions socially as well as linguistically, and describes an application of speech synthesis in a robot dialogue system that makes complementary use of visual information and peaking-style variation.

7.1 Introduction

Prosody control in speech synthesis has traditionally been concerned with the signaling of linguistic information, phrase structure, and semantics, but is now increasingly having to cope with the demands of expressive speech and social information processing (Campbell 2007; Vinciarelli et al. 2008). Speech synthesis itself is also facing new challenges as we enter the multimodal phase of human–computer interaction, and new streams of information become available for use in the human–computer

N. Campbell (✉)
Speech Communication Lab, Centre for Language and Communication Studies, Trinity College Dublin, The University of Dublin, Dublin, Ireland
e-mail: nick@tcd.ie

Y. Li
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

Trinity College Dublin, The University of Dublin, Dublin, Ireland
e-mail: yli@nlpr.is.ac.cn

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_7

interface (Tao et al. 2006). This chapter will look at some recent changes in speech processing, from an interface development point of view, and discuss some of the social issues of prosody processing that have now become more relevant.

We are already in the age of Siri, and our personal computing devices have become voice-aware so that they can listen and speak back to us. Information is no longer 'just at the tip of our fingers' but is now there 'simply for the asking'—the devices will tell us all; while Siri may have been the first, Apple is by no means alone—Google, Microsoft, and Intel are all in the race, with perceptual computing and cognitive processing forming core elements of human–computer interface technology, with Amazon and eBay soon to follow. Computers are starting to wake up and think; or more correctly, computers are now starting to process information in much the same way as humans do, to include social and interpersonal aspects of information sharing. They are becoming proactive and using massively parallel processes to guess our goals as part of the human–machine interface—they no longer just wait to take commands but are now beginning to 'read between the lines' and interpret the actions of the user to infer the underlying intentions to provide a more efficient form of interface for general information processing.

7.2 Perceptual Computing

The Human Interface Guidelines laid out in the recent Intel perceptual computing SDK (Intel Developer Zone n.d.) explain how the new generation of devices 'sense and perceive the user's actions in a natural, immersive, and intuitive way'. Input modalities include mid-air hand gestures, touch, voice, mouse, and keyboard, but voice is no longer limited to spoken text: 'People do not speak the way they write (Intel Developer Zone n.d., p. 19)' and in the near future nonverbal and nonspeech vocalisations will also be processed as part of the message.

The multimodal human interface accepts gestures, facial expressions, hand movements, and spoken instructions and is already a core part of the operating system, functioning alongside the more traditional keyboard and mouse and well on the way to replacing them in future personal computing, telecommunications, and conferencing. Already swipe and pinch are ubiquitous and well understood as interaction modalities, and 'thumbing' along with speech are replacing the traditional keyboard as standard input modalities for many interaction types. Gesture recognition is close behind, as computers are beginning to appear intelligent.

People can now interact with their personal devices in a more natural way, thanks to recent developments in these computer interface technologies, and the technologies in turn will affect the ways that people interact with each other when communicating remotely or through devices. Texting has replaced spoken phone calls in many cases, and people respond to incoming messages differently now that they can see before they answer who the call is from.

Speech-based interface technology has become multimodal, but the software drivers have yet to catch up—and the understanding required to process the way people speak in 'natural' interactions is not yet complete. Further research is required

in the ways that people communicate informally with each other so that machines can learn the rules and catch up. Social elements of speech processing must be taken into consideration to function alongside the more traditional linguistic ones.

7.3 Cognitive Speech Processing

The linguistic content of an utterance such as 'take two ounces of cherry wine' is rich and each word carries almost equal weight in lending meaning to the whole, but a social comment such as 'did ja see the game last night then?' is almost entirely formulaic and the utterance itself serves more as a greeting than as a question; as an informal invitation to enter into a social dance by responding with something from the heart like 'your man's kick!' . . . or 'worst ever' where the exchange of new information is minimal but the maximisation of social contact is primary. Only a fool would answer 'yes', and a bigger fool 'no'! In texting, a smiley-face would be an appropriate response. Social engagement is required in these situations, not just dry information about facts.

For speech synthesis, such social interactions will need a new form of prosody control; where the effect of voice quality takes precedence over the rise and fall of the intonation. Since the purpose of these utterances is not to impart linguistic information but rather to show a form of engagement or bonding, alongside any transfer of propositional content, the tone-of-voice is what carries the crucial information. A test question for the effectiveness of such an utterance might be 'are you with me?', rather than 'did you hear me?' or 'do you follow that?'

In such socially motivated interactions the exact nature of an appropriate response might be quite variable and a wink as good as a nod, but the timing of that 'wink' will be crucial. Though using 'wink' here metaphorically, it has a literal meaning in this context as well, as head or facial movement or a shrug of the shoulders can function equivalently alongside or instead of speech in social communication. The combined modalities work together to achieve the desired expression of engagement. Ideally they should also match.

Societies have evolved sophisticated codes of language usage that show inclusiveness within a group. Human ears are perhaps especially tuned to the fine prosodic and tonal nuances of these social verbal exchanges but the speech synthesiser is not yet capable of producing them. An utterance is an utterance; regardless of 'when' or 'where' or 'to whom' it is produced, and an identical sound sequence will normally be generated as output for any given input text sequence, regardless of utterance context(s).

7.4 Social Prosody

If the transmission of linguistic content can be thought of as a predominantly left-brain activity, and the inflections of affect as a predominantly right-brain activity, then current speech synthesis is almost completely left-brained; it only tells half the

story. When we talk to others, our voices also reveal how we feel about them and about what they are saying. The subtle differences in intonation and voice quality are social signals that human listeners are well attuned to, but that are missing from computer speech.

The linguistic uses of speech prosody are now well-known and models have already been implemented successfully in most text-to-speech synthesis systems (Sproat 1998; Santen et al. 1996). So as speech synthesis evolves from being a 'reading machine' to becoming an interactive 'talking device' the *social* uses of speech prosody are now more important to model for an effective conversational human–computer interface. Ideally, both hemispheres should contribute.

Voice-based interfaces using computer speech synthesis are already being used to mediate human–human conversations through, for example, speech translation, and to stand-in for a human in a machine-mediated dialogue such as a call centre interaction, or the early stages thereof (Edlund and Heldner 2005).

Computer speech is already required to make nonlinguistic vocalisations; to laugh, to insert backchannel utterances, and to make socially appropriate noises in a 'conversation' (Trouvain 2014), where in many cases the prosodic variation of these paralinguistic and universal/generic speech sounds carries sophisticated and complex information.

If you overhear someone you know well talking on the phone, for example, you can usually make a fair guess about the nature of that conversation and the speaker's relationship with the interlocutor without actually having to 'listen in'.

Much research has been carried out on the effects of emotion on the voice for purposes of speech synthesis and speaker analysis (Scherer 1989; AAAC n.d.), but in our experience, emotion was not the main factor that accounted for the voice-quality and speaking-style variation—it was far more closely related to 'nature of the interlocutor' or 'relationship with the other person' (Campbell and Mokhtari 2003).

Our analyses of speech from a very large corpus of interactive and spontaneous real-world conversations showed that a combination of fundamental frequency (F0) and normalised amplitude quotient (NAQ) (acoustic measures of pitch and vocal pressure) reliably covaried with 'nature of the interlocutor' so that we were able to make a good guess at who the person was talking to from aspects of her intonation and tone-of-voice.

We focussed on one commonly-used example; the word 'honma' (a Japanese Kansai-dialect expression that can be used whenever the English word 'really' would be appropriate; including expressive affective and interjective uses as well as acting as a simple amplifier to adjectives). Our corpus includes more than 3500 examples of this word in everyday use and while detailed acoustic and pragmatic analyses are still being undertaken, we speculate that there might be in the order of 35 different functional varieties of the word.

For the present work we analysed voice quality in several utterances of 'hai' (a common Japanese word meaning 'yes', with all it's different usages). We compared 100 utterances of each, randomly selected to represent eight different known interlocutors on a range of familiarity from the 600-hr FAN subset of the same ESP

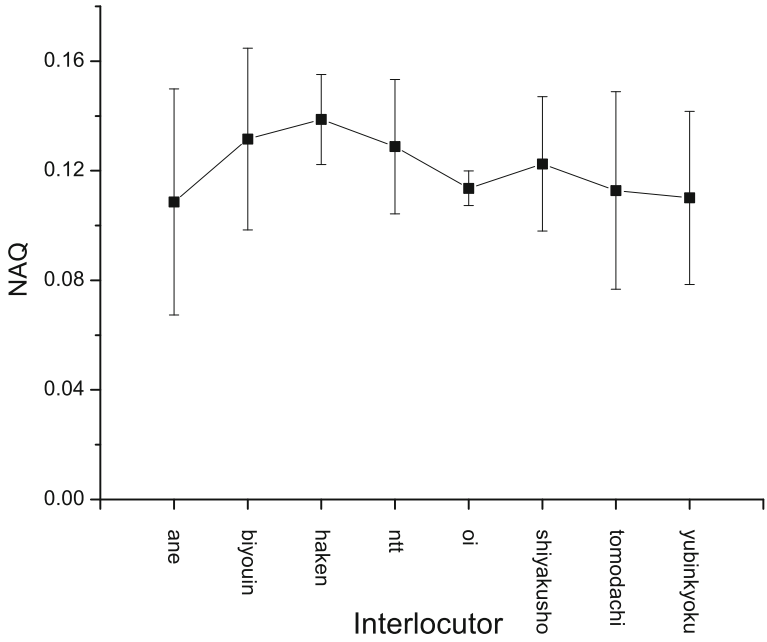


Fig. 7.1 Showing normalised amplitude quotient (*NAQ*) scores for speaker FAN from the JST/ESP corpus when addressing various different *interlocutors*. There is considerable variation in voice quality but the mean correlates well with 'relationship to the interlocutor' or 'formality/care' in the speech. We notice high formality when talking to *haken* (company) and *biyouin* (beauty shop) but low when talking with *tomodachi* (friends), *oi* (cousin), and *ane* (older sister. We might extrapolate from these results and wonder whether she also had a friend in the *yubinkyoku* (post office)

corpus. We (the second author) computed NAQ values (Alku et al. 2002) to estimate vocal tension for each class of interlocutor. Figure 7.1 shows that although there is considerable variation, which is to be expected since so many different 'meanings' are included, there is a clear difference in the means for different interlocutors, with friends being typically low-valued (indicating a harsher voice and speaking style), and more formal interactions (softer brighter voice) higher in the range. This type of variation in voice quality is consistent with our earlier findings, reflecting a kind of social relationship with the interlocutor that is rarely modelled in current speech synthesis.

The expressiveness of each pronunciation varies a little in each case according to the underlying emotion(s) and mood of the speaker but to a far greater extent in the expression of surprise, agreement, understanding, belief, shock, and a whole gamut of 'distance-expressing' or 'closeness-expressing' tones of voice. Expression of social and interpersonal 'distance' could most clearly be seen in speech between family members, and when talking to friends or business partners (strangers). However, even within the same social pairing (talking with a close friend for example) the range of expressiveness varies greatly as the speaker indicates different degrees

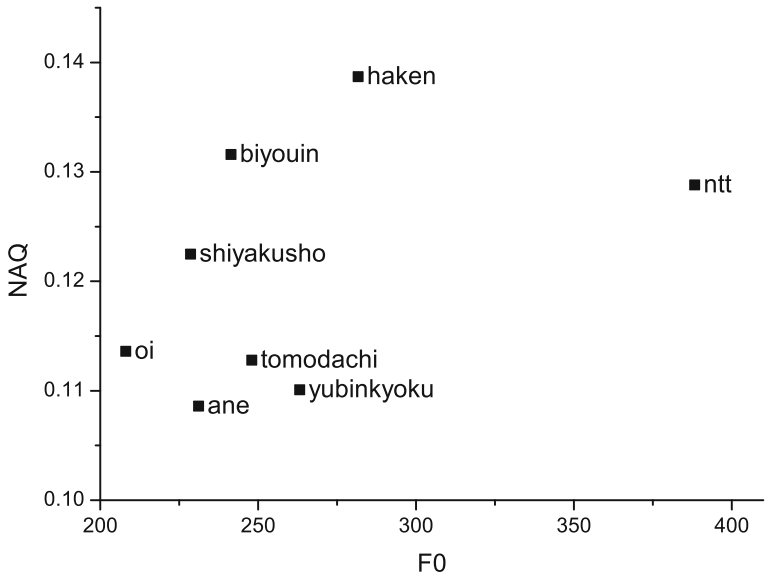


Fig. 7.2 Showing fundamental frequency (*F0*) plotted against normalised amplitude quotient (*NAQ*) scores for speaker FAN addressing various different interlocutors. Clear groupings emerge with familiar interlocutors clustered together in the *lower left* and the three business partners (*biyouin* (beauty-shop), *haken* (part-time employer), and *ntt* (the telephone company)) being distinguished by *F0* but not *NAQ*. The city hall (*shiyakusho*) falls in intermediate position perhaps due to its noncommercial relationship

or types of cognitive engagement with her partner. Figure 7.2 clarifies these relationships by plotting NAQ in conjunction with F0, which also shows similarly consistent nonlinguistic variation.

7.5 Interactive Speech Synthesis

No computer speech synthesiser that we know of is yet 'interactive' in the sense that it responds differently to individual listeners in different situations (but cf (Creative Speech Technology n.d.; Moore 2013; Moore and Nicolao 2011), and we are working to understand how such technology might be produced. Human speakers adjust their timing and speech content interactively according to how the interlocutor is perceived to be processing that content in real time. We watch our partners closely when talking (or produce more backchannel feedback when speaking over a phone) and the interaction becomes an active two-way process of mutual adjustments so that maximal (or at least efficient) speech processing can take place.

We conducted an extended experiment in casual robot–human conversation at the Science Gallery in Trinity College Dublin (TCD) a few years ago. The theme of the high-tech/art exhibition was 'Human+' (Science Gallery n.d.), and it featured

bionic developments as well as artificial-intelligence exhibits. Ours was a small robot, Herme (Science Gallery n.d.), that struck up a conversation with passers-by. It worked by using image processing and searching the scene in front of it for a face to appear, when it began its conversation routine. The sequence of utterances it was able to make was completely fixed and the robot had no faculty for speech recognition(the environment was so noisy that no speech recogniser would have worked anyway!).

The trick in maintaining the flow of a conversation was for the robot to take and keep the initiative at each stage of the conversation and appear to be listening to what the visitor was saying to it ... and the trick in keeping the visitor engaged in that conversation was to get the timing of each subsequent utterance right, within a small window of allowable time. Too early and we risked talking over a visitor's response; too late and the robot displayed an apparent lack of 'interest' in that reply.

We tested variation in timing control by using human operators in a wizard-of-Oz scenario, remotely triggering each utterance in sequence while viewing the interaction from afar and, alternatively, by using the robot's sensors to detect signs of voice activity and facial movement in order to trigger each utterance automatically. The only difference between the two scenarios was in the timing of the utterance release ... but the human wizards (mostly students in our lab) often succeeded in keeping a visitor present for the entire sequence of utterances (the whole conversation ending in the signing of a consent form allowing us subsequent use of the data) while the automatic system failed regularly! (but provided good training material for later comparison with the successful dialogues). Humans can read signals, albeit corrupted over a Skype connection, which our sensors failed to register.

No use of expressive speech per se was necessary for these interactions, but the utterances were deliberately childish and the voice (the default synthesiser provided with an Apple operating system) was modified by signal processing to raise the pitch and compress the formants so that it appeared small and appropriate to the device (we used Apple's 'Princess' voice as the base with default prosody). When somewhat cheeky questions were asked by the 'cute' voice, most participants reacted favourably and responded positively (see the Appendix for a full listing of the dialogue sequence).

The nature of the utterances themselves was particularly conversational; brief and chatty. They were sentences designed for people passing the time together—casual and often 'ill-formed' but contextually appropriate. Thus, the triad of necessary components was complete: something interesting to say, a cute (or appropriate/appealing) tone-of-voice, and 'appropriate' timing for each utterance. There was some socially determined and pattern-based evolution in the selection of the utterances but apparently very little latitude in the timing by which they could be produced. The tone-of-voice (the core 'personality' selected by the speaker at any moment) sets the scene for an interpretation of the overall context and reliably triggers the viewer's response.

7.6 Where Next?

In this post-Snowden age of ubiquitous information processing and metadata gathering, it may be necessary to include a few words about data protection and ethics with regard to interactive voice-based systems. The conversational speech synthesis that we are working towards needs to be aware of the person it is talking with (though in the case of plural listeners little personal information would be required). Sensors are needed that take in voice activity and process acoustic features, as well as for image processing to capture facial and bodily dynamics.

Different languages and cultures make different use of gesture, intonation, voice quality ranges, etc., and the norms for each society or subcultures within a society are subtly different. By incorporating memory into the system, both short- and long-term, it becomes possible to use normalisation techniques to overcome the individual differences; by recalibrating the range of sensitivity for each conversation or individual user. Such use of memory raises ethical issues.

However, the face and voice per se would not need to be stored or processed, only the dynamics thereof, or the relative change in certain features over time, so no personal data would be required for such a system to work. Herme did not have to listen to what her partners said to her so long as she was able to respond 'appropriately' and with the right timing—no speech recognition was used for processing content. However, the possibility of personalisation might require storage of parameters that allow better normalisation of input feature dynamics. It is important to design a system where the user has the right to enable or disable these features, and is made aware of how little or how much information about them is processed or stored.

7.7 Summary and Conclusion

This chapter discussed some aspects of paralinguistic and nonlinguistic prosody for potential application in multimodal and interactive speech synthesis. It described work-in-progress at the Speech Communication Lab in TCD (the University of Dublin) towards the development of expressive dialogue speech synthesis and has put forward the view that awareness of the interlocutor is an essential component of efficient spoken interaction and that current speech technology might be fruitfully engineered in that direction. The fusion of visual and vocal cues can be helpful in predicting the appropriate timing of a speech utterance and the selection of an appropriate voice quality (and expressiveness) is crucial for an appropriate rendering of any utterance in context.

Our present and future work involves the TinWoman, a large static interactive device for placing at home or office where people might come together for social chit-chat, somewhat like a coffee machine in a large organisation. It employs a wider range of sensors that enable us to monitor the progress of an extended Herme-like conversation so that we can improve the timing and voice quality and test a range of informal utterance types that characterise the so-called chit-chat. Our goal is to

produce a talking machine that is sensitive to its interlocutor, has metacognitive abilities (Metalogue, EU FP7 research n.d.), and is able to impart a sense of humor into a conversation (JOKER - FP7 Chist-Era funded research n.d.). Speech synthesis has come a long way in the past years, but it is still far from being able to replicate the types of voice and speaking style that are necessary for satisfying social interactions.

Acknowledgement The authors would like to acknowledge the contribution of SFI (through the FastNet (09/IN.1/I2631) and CNGL (12/CE/I2267) projects, and the Stokes Professorship (07/SK/I1218)), as well as joint work with NAIST in Japan (Kaken-hi 24500256 & 23242023) and with Nanjing Normal University in China (parts of this work are supported by the Major Program for the National Social Science Fund of China (13&ZD189)). The principal author further wishes to thank the Chinese Academy of Sciences for the kind loan of the second author to our lab. We also thank Emer Gilmartin for her perceptive contributions to the development of Herme's script.

Appendix

Full text of the Herme utterances as used in the Science Gallery experiment:
Each line represents a new utterance to be synthesised separately, with a variable-length pause, and some groups of lines (starting with a dash) form a mini subdialogue section—we experimented by varying the timing of the pauses, i.e. utterance onsets, according to the nature of each response from the visitors to the exhibition—Herme's chat partners. Where possible, the content of any response was ignored as the wizard's task was simply to trigger the next utterance and the automatic version had no content processing module.

```
- hello? hi
 hello
 hi
- my name is hur-mi.
 h e r m e
  hur-mi
what's your name?
- how old are you?
 really
  I'm nearly seven weeks old
- do you have an i d number
  i need an i d number to talk to you
  i d numbers are on your right
  thank you
- are you from dublin?
 really
 I'm from the Speech Communication Lab here in Tee See Dee
- tell me about you
 really
```

```
 owe
- tell me something else
 owe
 really
- why are you here today?
 really
 why
- do you like the exhibition
 really
 why?
i like your hair
- do you know any good jokes?
 tell me a funny joke
 ha ha haha ha
- tell me a knock knock joke
 who's there
 who?
 who
ha ha haha ha
- I know a joke
what's yellow and goes through walls
 a ghost banana
ha ha hehe he.
 ho hoho ho ho
- thanks for your help
 goodbye, see you later
  goodbye
```

Note how these utterance chunks typically group into triads (sets of three) and how they maintain the initiative of the conversation throughout. Note also some spelling hacks (e.g., using 'owe' to ensure that the synthesiser correctly pronounced 'oh!'). The repeated use of 'oh' and 'really' (sometimes with 'why') with various punctuation served to keep the conversation interactive and was key to Herme's supposed conversational abilities.

References

AAAC. 2014. Research on affective computing, emotions and human-machine interaction. http://emotion-research.net.

Alku, P., T. Bäckström, and E. Vilkman. 2002. Normalised amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America* 112 (2): 701–710.

Campbell, N. 2007. Expressive speech processing & prosody engineering. In *New trends in speech based interactive systems*, ed. Fang Chen and Kristiina Jokinen. New York: Springer.

Campbell, N., and P. Mokhtari. 2003. Voice quality: The 4th prosodic dimension. In Proceedings of the 15th international congress of phonetic sciences (ICPhS'03), Barcelona, Spain, 2417–2420.

Creative Speech Technology. 2014. http://crestnetwork.org.uk/page/beyond-speech.
Edlund, J., and M. Heldner. 2005. Exploring prosody in interaction control. *Phonetica* 62 (2–4): 215–226.
Intel Developer Zone. 2014. Intel®Perceptual Computing SDK 2013. https://software.intel.com/en-us/vcsource/tools/perceptual-computing-sdk/home.
JOKER—FP7 Chist-Era funded research. 2014. http://www.chistera.eu/projects/joker.
Metalogue. 2014. EU FP7 research. http://www.metalogue.eu.
Moore, R. K. 2013. Spoken language processing: Where do we go from here? In *Your virtual butler, LNAI,* ed. R. Trappl, vol. 7407, 111–125. Heidelberg: Springer.
Moore, R. K., and M. Nicolao. 2011. Reactive speech synthesis: Actively managing phonetic contrast along an H&H continuum. 17th international congress of phonetics sciences (ICPhS), Hong Kong.
Scherer, K. R. 1989. Vocal correlates of emotion. In *Handbook of psychophysiology: Emotion and social behavior,* ed. A. Manstead and H. Wagner, 165–197. London: Wiley.
Science Gallery. 2011. Human+: The future of our species. https://dublin.sciencegallery.com/humanplus/.
Science Gallery. 2011. Human+: The future of our species. Talking with robots. https://dublin.sciencegallery.com/humanplus/talking-robots/.
Sproat, R. 1998. *Multilingual text-to-speech synthesis: The Bell Labs approach*. Boston: Kluwer.
Tao, J., L. Huang, Y. Kang, and J. Yu. 2006. The friendliness perception of dialogue speech. Proceedings of Speech Prosody, Germany.
Trouvain, J. 2014. Laughing, breathing clicking—The prosody of nonverbal vocalisations. Proceedings of Speech Prosody (SP7), Dublin, 598–602.
Van Santen, J. P. H., R. W. Sproat, and J. P. Olive, et al. eds. 1996. *Progress in speech synthesis*. New York: Springer-Verlag.
Vinciarelli, A., M. Pantic, and H. Bourlard. 2008. Social signal processing: Survey of an emerging domain. *Image and Vision Computing* 27:1743–1759.

Chapter 8
Temporally Variable Multi attribute Morphing of Arbitrarily Many Voices for Exploratory Research of Speech Prosody

Hideki Kawahara

Abstract Morphing provides a flexible research strategy for non- and para linguistic aspects of speech. Recent extension of the morphing procedure has made it possible to interpolate and extrapolate physical attributes of arbitrarily many utterance examples. By using utterances representing typical instantiation of the non- and para linguistic information in question and introducing systematic perturbation of trajectories in a high-dimensional space spanned by a set of indexed weights for the physical parameters of utterances, the physical correlates of such information can be represented in terms of a differential geometrical concept. Formulation of this extended morphing framework in generalized representations and a few representative cases of applications are discussed with comments on the limitations of the current implementation and possible solutions.

8.1 Introduction

It is not trivial to identify contributing physical parameters of perceptual attributes in spoken utterances. Identification of perceptual attributes, which play important roles in conveying non- and para linguistic information embedded in speech sounds as a result of multi stage encoding processes (Fujisaki 1996), is itself a hard problem. Solving it requires interdisciplinary and exploratory research strategies.

Morphing (Kawahara and Matsui 2003) introduced a unique means for enabling *quantitative* example-based exploitation as well as verification for studying speech prosody for existing research strategies, such as statistical analysis or recognition of speech materials and perceptual tests with synthetic speech or converted speech materials (Schröder 2001; Douglas-Cowie et al. 2003; Turk and Schroder 2010; Schuller et al. 2011). The morphing introduced in this chapter is based on STRAIGHT, a framework for analysis, modification, and synthesis of speech sounds, which allows precise control of physical parameters while preserving manipulated speech quality naturally. This flexibility with reasonable quality made it possible to measure the effects of perceptual attributes quantitatively. For example, a stimulus continuum

H. Kawahara (✉)
Wakayama University, 930 Sakaedani, Wakayama, 640-8510 Wakayama, Japan
e-mail: kawahara@sys.wakayama-u.ac.jp

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_8

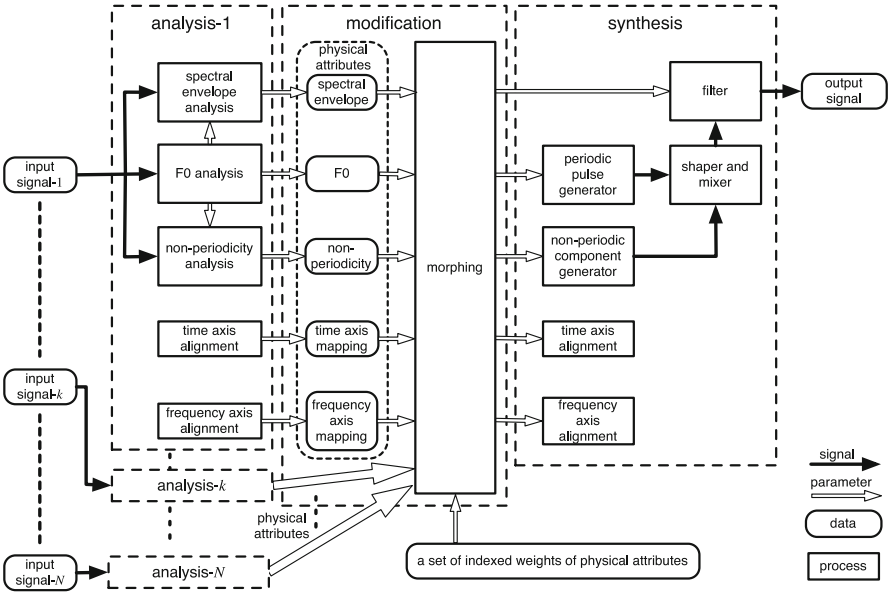


Fig. 8.1 Schematic diagram of N-way morphing based on TANDEM-STRAIGHT

generated by morphing was used to quantify after effects in the gender perception of voices (Schweinberger et al. 2008). Independent control of physical attributes is also applicable for studying the contribution of each physical attribute on speaker perception (Schweinberger et al. 2014). Averaging many voices with this morphing recursively revealed that averaging makes voice more attractive (Bruckert et al. 2010). The last example and following collaborations led to substantial extension of the morphing procedure. This extended morphing makes it possible to interpolate and extrapolate parametric representations of arbitrarily many utterance examples (Kawahara et al. 2013) on the basis of a completely reformulated analysis and synthesis framework, TANDEM-STRAIGHT. Please refer to references (Kawahara et al. 2008; Kawahara and Morise 2011) for technical details of TANDEM-STRAIGHT analysis and synthesis procedures. This section is an introduction to the underlying idea and prospective application examples of this extended morphing in speech prosody research.

8.2 Morphing Framework

Figure 8.1 shows a schematic diagram of the extended morphing based on TANDEM-STRAIGHT. Each input signal, usually an utterance in prosodic research, is analyzed and yields a set of parameters. The extended morphing procedure merges all sets of parameters made from corresponding inputs to generate a set of parameters on the

basis of a set of weights for individual utterances and constituent parameters as functions of time. The generated set of parameters is used to synthesize the output. The following sections are an introduction on how to build this extended morphing framework, starting from the relationships between constraints posed to parameters and the corresponding relevant way of morphing.

8.2.1 Constraints and Extended Morphing

Morphing, which is capable of extrapolation, has to assure that the extrapolated parameter does not violate constraints posed to the specific parameter. Depending on the type of posed constraint, speech parameters can be categorized into the three groups given below.[1]

8.2.1.1 No Constraint

When F0 is represented in terms of musical cent, theoretically, it can be any real number. Extrapolation based on arithmetic operation of real numbers yields real numbers. This assures that the extrapolated parameters do not violate the original constraint.

In this case, morphed parameter g_{m} made from N examples is defined by the following equation.

$$g_{\mathrm{m}} = \sum_{k=1}^{N} w_k g_k, \tag{8.1}$$

where subscript k represents the index of examples, w_k represents the corresponding weight, and g_k represents the specific parameter of interest. Usually, normalization condition $\sum_{k=1}^{N} w_k = 1$ is used. However, this condition is not mandatory.

8.2.1.2 Positivity Constraint

A spectral envelope obtained by TANDEM-STRAIGHT analysis is a power spectral envelope and is always positive. Logarithmic conversion of positive values yields real numbers. Extrapolation based on arithmetic operation on logarithmic conversion of a spectral envelope followed by exponential conversion assures that the extrapolated values are always positive.

[1] This grouping is specific to speech analysis results obtained by TANDEM-STRAIGHT. There can be more groups for other parameters.

Morphed parameter g_{m} made from N examples obeying the positivity constraint is defined by the following equation, where the same note holds for the weight w_k.

$$g_{\mathrm{m}} = \exp\left(\sum_{k=1}^{N} w_k \log\left(g_k\right)\right). \tag{8.2}$$

8.2.1.3 Monotonicity Constraint

Each set of analyzed parameters is aligned on each time axis. Mapping from the underlying index to the corresponding time is a monotonic increasing function. The extrapolated result has to obey this constraint. Taking into account the fact that the time derivative of the mapping is always positive, integration of the exponential conversion of the extrapolated logarithmic derivative of mappings assures that the extrapolated mapping from index to time is a monotonic increasing function.

Morphed function $g_{\mathrm{m}}(\tau)$ of abstract index τ made from N examples obeying increasing monotonicity is defined by the following equation.

$$g_{\mathrm{m}}(\tau) = C_1 \int_{\tau_0}^{\tau} \exp\left(\sum_{k=1}^{N} w_k(\lambda) \log\left(\frac{dg_k(\lambda)}{d\lambda}\right)\right) d\lambda + C_0, \tag{8.3}$$

where C_0 and $C_1 (> 0)$ are constants for fulfilling the boundary condition. The weight $w_k(\tau)$ in this case is also a function of abstract index τ. Usually, the normalization condition on $\sum_{k=1}^{N} w_k(\tau) = 1$ is posed. However, this condition is not mandatory. For example, the right hand side can be a function of the abstract time τ.

8.2.2 Parameter Alignment

The analysis stage of TANDEM-STRAIGHT decomposes an input signal into three types of parameters; spectral envelope $P(f, t)$, fundamental frequency $f_0(t)$, and aperiodicity parameters $f_{\mathrm{c}}(t)$ and $\alpha(t)$, which represent the inflection point and slope of the sigmoid model (Kawahara et al. 2010). Each parameter in different examples has to be aligned to match each other for them to be morphed properly. This implies that mapping from the underlying abstract temporal index to the time axis of each example and that for the frequency axis also have to be morphed by using Eq. 8.3 before morphing these three types of parameters by using Eq 8.1 or Eq. 8.2.

Let $t_k(\tau)$ represent the temporal axis mapping of the k-th example. Also, let $f_k(\nu, \tau)$ represent the frequency axis mapping of the k-th example, where τ and ν represent abstract time and frequency indices. Parameters are aligned to these abstract indices by using the following inverse functions.

$$\tau = t_k^{-1}(t_k) = t_k^{-1}(t_k(\tau)) \tag{8.4}$$

$$\nu = f_k^{-1}(f_k, t_k) = f_k^{-1}(f_k(\nu, \tau), t_k(\tau)) \tag{8.5}$$

By using these inverse functions, the parameter set of the k-th example on its own time and the frequency axes is converted to the set of parameters on the abstract time and frequency indices. Equations for representing the elements of the set are given below.

$$P_k^{(\mathrm{C})}(\nu, \tau) = P_k^{(\mathrm{C})}(f_k^{-1}(f_k, t_k), t_k^{-1}(t_k)) = P_k(f_k, t_k) \tag{8.6}$$

$$f_{0,k}^{(\mathrm{C})}(\tau) = f_{0,k}^{(\mathrm{C})}(t_k^{-1}(t_k)) = f_{0,k}(t_k) \tag{8.7}$$

$$\{f_{\mathrm{c},k}^{(\mathrm{C})}(\tau), \alpha_k^{(\mathrm{C})}(\tau)\} = \{f_{\mathrm{c},k}^{(\mathrm{C})}(t_k^{-1}(t_k)), \alpha_k^{(\mathrm{C})}(t_k^{-1}(t_k))\} = \{f_{\mathrm{c},k}(t_k), \alpha_k(t_k)\}, \tag{8.8}$$

where superscript C represents that parameters are aligned on the abstract indices. These aligned parameters are morphed by using Eq. 8.1 or Eq. 8.2 depending on their constraints.

8.2.3 Morphing Weight Representations

For morphing N examples, weight has to be indexed by example identifier $k \in \{1, \ldots, N\}$, attribute identifier[2] $X \in \Lambda = \{\mathrm{F0}, \mathrm{A}, \mathrm{P}, \mathrm{t}, \mathrm{f}\}$, and abstract time index τ.[3] Depending on the constraints on weights, morphing is categorized into four groups.

Let $w_k^{(X)}(\tau)$ represent one instance of a specific weight.

8.2.3.1 Temporally Static Tied-Attribute N-Way Morphing

The simplest case of N-way morphing[4] is when weights depend only on example identifier. The following constraint is posed.

$$w_k^{(X_1)}(\tau_1) = w_k^{(X_2)}(\tau_2) = w_k. \tag{8.9}$$

This type of morphing is uniquely determined by the weight vector $\vec{w}=[w_1, \ldots, w_N]^T$, where superscript T represents transposition.

[2] F0, A, P, t, and f represent fundamental frequency, aperiodicity parameter, power spectral envelope, time axis, and frequency axis, respectively.

[3] An abstract frequency index can also be used for indexing weights in principle. However, in this article, only abstract time is used for indexing weights.

[4] Instead of writing "morphing of arbitrarily many voices," "N-way morphing" is used.

8.2.3.2 Temporally Variable Tied-Attribute N-Way Morphing

The next extension is to make weights temporally variable. The following constraint is posed.

$$w_k^{(X_1)}(\tau) = w_k^{(X_2)}(\tau) = w_k(\tau). \tag{8.10}$$

This type of morphing is uniquely determined by the vector function $\vec{w}(\tau) = [w_1(\tau), \ldots, w_N(\tau)]^T$.

8.2.3.3 Temporally Static Multi-Attribute N-Way Morphing

The other extension is to weigh attributes independently while keeping them temporally static. The following constraint is posed.

$$w_k^{(X)}(\tau_1) = w_k^{(X)}(\tau_2) = w_k^{(X)}. \tag{8.11}$$

This type of morphing is uniquely determined by the weight matrix $W = [\vec{w}^{(\mathrm{F0})}, \vec{w}^{(\mathrm{A})}, \vec{w}^{(\mathrm{P})}, \vec{w}^{(\mathrm{t})}, \vec{w}^{(\mathrm{f})}]$.

8.2.3.4 Temporally Variable Multi-Attribute N-Way Morphing

When no constraint is posed, it is the most flexible morphing, temporally variable multi attribute N-way morphing. This type of morphing is uniquely determined by the matrix function $W(\tau) = [\vec{w}^{(\mathrm{F0})}(\tau), \vec{w}^{(\mathrm{A})}(\tau), \vec{w}^{(\mathrm{P})}(\tau), \vec{w}^{(\mathrm{t})}(\tau), \vec{w}^{(\mathrm{f})}(\tau)]$.

8.2.4 Implementation

This extended morphing procedure is implemented by using a scientific computation environment (Matlab (2013) version 8.2.0.701 (R2013b) 2013). This morphing scheme requires determining mappings of time and frequency axes for each example. In the current implementation, piecewise linear functions are determined by using anchoring points.

Figure 8.2 shows the graphical user interface (GUI) for assigning anchor points to each analysis result. The left GUI is for assigning anchoring points to the first example. Anchoring points of the other examples are assigned by using the center GUI, where the assignment of the first example is displayed on the right side of the GUI. The upper image of each GUI is a spectrographic representation of the spectral envelope analyzed by TANDEM-STRAIGHT. The lower plot of each GUI shows the waveform, spectral flow (black lines), and voiced/unvoiced indicator (cyan box). These displays are viewing ports of the underlying whole image. The size and location of the visible area from the ports are manipulated by using Matlab GUI tools

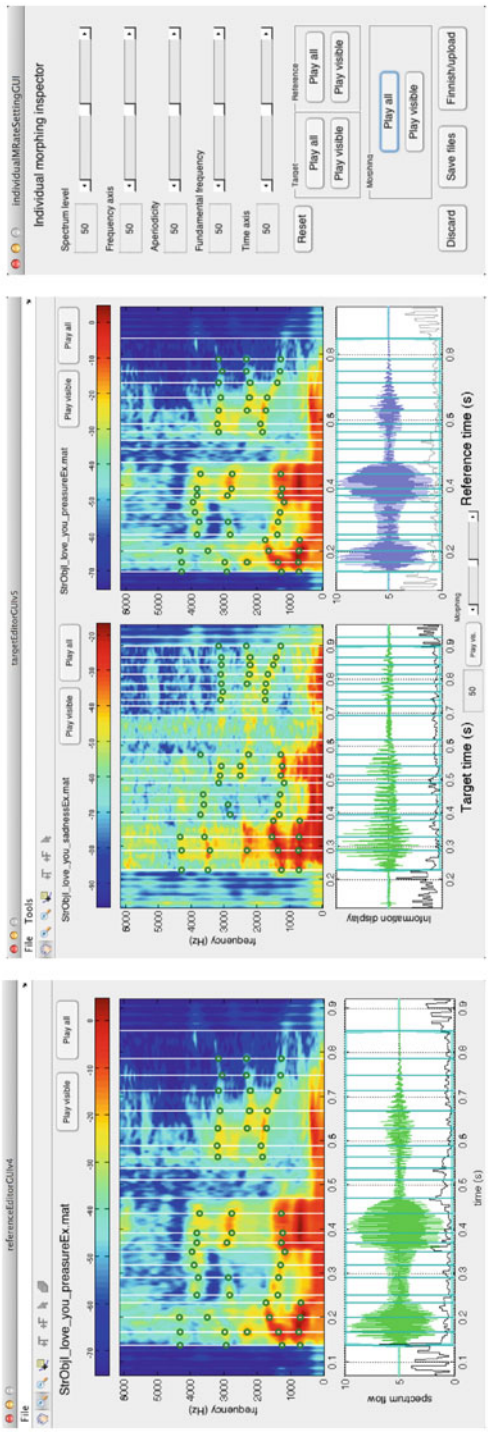


Fig. 8.2 Graphical user interface displays for assigning anchoring points and morphing control panel for interactive inspection

(pan, zoom-in, and zoom-out) shown on their menu bars. Users can add, adjust, and delete the vertical white lines in the upper image to assign temporal anchor points. Users also can add, adjust, and delete the open circles on the vertical lines to assign frequency anchor points.

The right image shows the morphing control panel, which sets weights to control "temporally static multi attribute two-way morphing." It allows interactive inspection by listening to morphed speech as a whole or visible portion. The morphed speech and used weight matrix in this inspection are exported as a sound file and a data file consisting of the weight matrix.[5]

After interactive manual assignment and adjustment of anchoring points, the resultant anchor information is stored. The morphing procedure and the synthesis procedure are integrated as a Matlab function to perform "temporally variable multi attribute N-way morphing with synthesis." Four types of morphing procedures are implemented by using this function by setting a set of weight matrices properly.

8.3 Application to Speech Prosody Research

In this section, application of this extended morphing framework to prosody research is discussed. Conventional morphing is understood to be a special case of four types of two-way morphing models in this extended morphing framework. Discussion starts from this two-way morphing.

8.3.1 Two-Way Morphing

Two-way morphing of all four types is already useful. For example, the simplest temporally static tied-attribute morphing can be applied to generate a stimulus continuum spanning two examples having competing non- or para linguistic information. Such a continuum can be used as a reference scale to quantify the amount of perceived information in given unknown utterances. It can also be applied to testing if perception of the attribute in question is categorical or graded. Temporally variable tied-attribute morphing is useful to evaluate contributions of different parts of the utterance for the attribute in question. Temporally static multi attribute morphing is also useful for evaluating the contributions of different physical attributes on a specific perceptual attribute. The most flexible temporally variable multi attribute morphing can be used to quantify perceptual effects of physical parameter perturbation. In other words, it is applicable to derive the element of the Jacobian $\frac{\partial \Psi_l}{\partial x^{(X)}}$ of the mapping from physical attributes $\{x^{(X)}\}_{X \in \Lambda}$ to perceptual attributes $\{\Psi_l\}_{l \in \Omega}$, where Ω represents a set of perceptual attributes (or evaluated non- and para linguistic information).

[5] Please note that this functionality is for initial inspection only.

8.3.2 Three-Way Morphing and General N-Way Morphing

In a more general case, N-way morphing adds another possibilities. Figure 8.3 shows an example of three-way morphing with a GUI designed mainly for demonstrations. This time, a temporally static tied-attribute morphing GUI is shown in the right panel, and the supporting GUI for binding utterances and their meta data is shown in the left panel. The left bottom list shows the contents of the data file generated with this supporting GUI. The data file consists of analysis data and an anchor information bundle.

Once relevant binding is established, the GUI in the right panel is used to interactively assign a weight of each example utterance and synthesize the morphed sound. Three colored spheres display the weights for each utterance by their projected area and shape. When the corresponding weight is negative, a bowl shape is used instead of a sphere shape. The semitransparent gray sphere is a control knob for the user to manipulate. The signed areas of the three triangles formed by the gray sphere and two of the three colored spheres divided by the area of the triangle formed by three colored spheres is used as the weights for example utterances. This weight definition always fulfills the condition $\sum_{k=1}^{3} w_k = 1$. The upper right panel of the GUI displays F0 trajectories of each utterance (using the corresponding color) and that of the morphed one (using the black line). The lower right panel displays the morphed spectral envelope.

This screen shot shows an example of morphing with extrapolation. The triangle formed by gray, green, and blue is outside of the triangle formed by the yellow bars and has a negative area. This negative area (negative weight) is represented by the red bowl. Weights are also represented by the color bars and numbers under the GUI manipulator. In this example, red, green, and blue markers are assigned to emotional utterances expressed with pleasure, sadness, and anger, respectively. The text of the utterances is “I love you” spoken by a male actor. The morphed speech represents “somewhat angry and slightly sad and, at the same time, negatively happy” emotional expression. The morphed speech was perceived to be consistent with this confusing description.

This example can be interpreted differently. Setting the gray sphere on the center of the triangle generates a morphed sound with emotionally neutral expression. Deviation from the center adds some specific emotional expression. The magnitude of deviation indicates the strength of the expression. Therefore, moving outside of this triangle means exaggerating or caricaturing the expression. Consequently, moving toward the center of the triangle means neutralization.

This exaggeration and neutralization approach is also applicable to prosody-related attributes and scalable to more general N-way morphing. Combining this approach with the other three types of morphing categories may open interesting research as well as application possibilities.

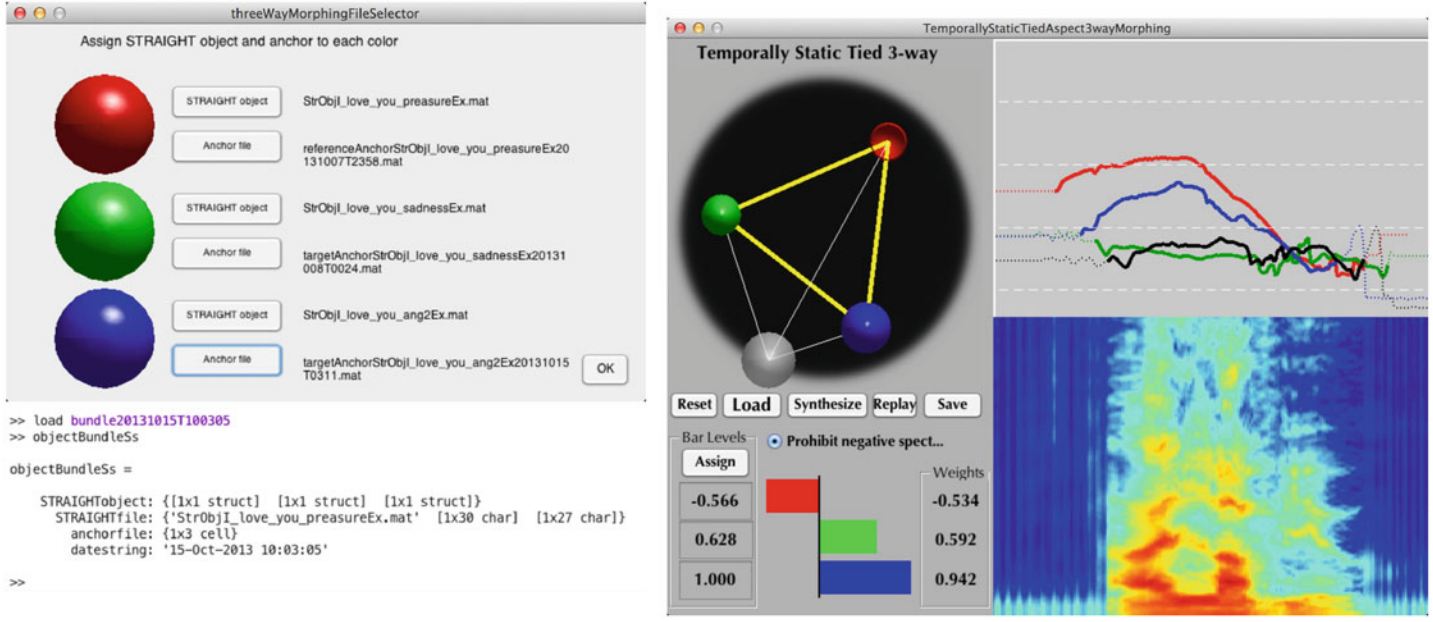


Fig. 8.3 Demonstration of temporally static tied-attribute three-way morphing of emotional speech

8.3.3 Limitations and Further Study

This extended morphing procedure assures that extrapolated parameters do not violate conditions posed to them. However, spectral envelope level extrapolation introduces annoying perceptual effects such as unnatural loudness variations and musical noise when extrapolation is large. The demonstration shown in Fig. 8.3 alleviates this annoyance by using the spectral level at the point where the yellow and white lines cross of the manipulation GUI when the gray sphere is in the extrapolation area. This suggests two issues to be investigated. First, it may be better to separate power and the spectral shape manipulation since they have close relationships to different salient perceptual attributes, loudness and timbre, respectively. Second, it may also be better to manipulate parameters in physically meaningful domains such as a vocal tract area function derived from the STRAIGHT spectrum (Arakawa et al. 2010). This study will also lead to a more relevant and easy to use definition method of time and frequency mapping than the current manual anchor assignment.

8.4 Conclusion

A flexible morphing procedure based on TANDEM-STRAIGHT, a speech analysis, modification, and a synthesis framework is introduced. The procedure can morph arbitrarily many examples with temporally varying weights for each constituent physical speech parameter independently. Each of the four types of morphing has prospective applications in speech prosody research by providing flexible tools for *quantitative* example-based exploitation.

References

Arakawa, A, Y Uchimura, H Banno, F Itakura, and H Kawahara. 2010. High quality voice manipulation method based on the vocal tract area function obtained from sub-band LSP of STRAIGHT spectrum. In Proceedings of ICASSP2010, IEEE, pp 4834–4837.

Bruckert, L., P. Bestelmeyer, M. Latinus, J. Rouger, I. Charest, G. A. Rousselet, H. Kawahara, and P. Belin. 2010. Vocal attractiveness increases by averaging. *Current Biology* 20 (2): 116– 120.

Douglas-Cowie, E., N. Campbell, R. Cowie, and P. Roach. 2003. Emotional speech: Towards a new generation of databases. *Speech communication* 40 (1): 33–60.

Fujisaki, H. 1996. Prosody, models, and spontaneous speech. In *Computing prosody*, ed. Y. Sagisaka, N. Campbell, and N. Higuchi, 27–42l. New York: Springer.

Kawahara, H., and H. Matsui. 2003. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In Proceedings of ICASSP2003, Hong Kong, vol I, pp 256–259.

Kawahara, H., and M. Morise. 2011. Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework. *Sadhana* 36 (5): 713–727.

Kawahara, H., M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. 2008. A temporally stable power spectral representation for periodic signals and applications to interference-free

spectrum, F0 and aperiodicity estimation. In Proceedings of ICASSP 2008, IEEE, pp. 3933–3936.

Kawahara, H., M. Morise, T. Takahashi, H. Banno, R. Nisimura, and T. Irino. 2010. Simplification and extension of non-periodic excitation source representations for high-quality speech manipulation systems. In Proceedings of Interspeech2010, ISCA, pp. 38–41.

Kawahara, H., M. Morise, H. Banno, and V. Skuk. 2013. Temporally variable multi-aspect N-way morphing based on interference-free speech representations. In 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), OS.28-SLA.9.

Matlab. 2013. version 8.2.0.701 (R2013b). The MathWorks Inc., Natick, Massachusetts, USA.

Schröder,M. 2001. Emotional speech synthesis: a review. In INTERSPEECH, pp. 561–564

Schuller, B., A. Batliner, S. Steidl, and D. Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53 (9): 1062–1087.

Schweinberger, S. R., S. Casper, N. Hauthal, J. M. Kaufmann, H. Kawahara, N. Kloth, and D. M. Robertson. 2008. Auditory adaptation in voice perception. *Current Biology* 18:684–688.

Schweinberger, S. R., H. Kawahara, A.P. Simpson, V. G. Skuk, and R. Zäske. 2014. Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science* 5 (1): 15–25.

Turk, O., and M. Schroder. 2010. Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. *IEEE Trans Audio Speech and Language Processing* 18 (5): 965–973.

Part III
Control of Prosody in Speech Synthesis

Chapter 9
Statistical Models for Dealing with Discontinuity of Fundamental Frequency

Kai Yu

Abstract The accurate modelling of *fundamental frequency*, or *F0*, in HMM-based speech synthesis is a critical factor for achieving high quality speech. However, it is also difficult because F0 values are normally considered to depend on a binary voicing decision such that they are continuous in voiced regions and undefined in unvoiced regions. Namely, estimated F0 value is a discontinuous function of time, whose domain is partly continuous and partly discrete. This chapter investigates two statistical frameworks to deal with the discontinuity issue of F0. *Discontinuous F0 modelling* strictly defines probability of a random variable with discontinuous domain and model it directly. A widely used approach within this framework is *multi-space probability distribution* (MSD). An alternative framework is *continuous F0 modelling*, where continuous F0 observations are assumed to always exist and voicing classification is modelled separately. Both theoretical and experimental comparisons of the two frameworks are given.

9.1 Statistical Parametric Speech Synthesis with Discontinuous Fundamental Frequency (F0)

Compared to traditional unit concatenation speech synthesis approaches, statistical parametric speech synthesis has recently attracted much interest due to its compact and flexible representation of voice characteristics. Hidden Markov model (HMM)-based synthesis (Yoshimura et al. 1999) is the most widely used approach of statistical parametric speech synthesis and is the focus of this chapter. Based on the source-filter model assumption, phonetic and prosodic information are assumed to be conveyed primarily by the spectral envelope, fundamental frequency (also referred to as *F0*) and the duration of individual phones. The spectral and F0 features can be extracted from a speech waveform (Kawahara et al. 1999a), and durations can be manually labelled or obtained through forced-alignment using pre-trained HMMs. A unified

K. Yu (✉)
Computer Science and Engineering, Shanghai Jiao Tong University, 800, Dongchuan Road, Minhang District, Shanghai, P. R. China
e-mail: kai.yu@sjtu.edu.cn

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_9

HMM framework may then be used to simultaneously model these features, where the spectrum and F0 are typically modelled in separate streams due to their different characteristics and time scales. During the synthesis stage, given a phone context sequence generated from text analysis, the corresponding sequence of HMMs are concatenated and spectral parameters and F0 are generated (Tokuda et al. 2000). These speech parameters are then converted to a waveform using synthesis filters (Imai 1983).

The modelling of F0 is difficult due to the differing nature of F0 observations within voiced and unvoiced speech regions. F0 is an inherent property of periodic signals and in human speech it represents the perceived *pitch*. During *voiced* speech such as vowels and liquids, the modulated periodic airflow emitted from the glottis serves as the excitation for the vocal tract. Since there is strong periodicity, F0 values can be effectively estimated over a relatively short-time period (e.g. a speech frame of 25 ms) using (Kawahara et al. 1999b). These F0 observations are continuous and normally range from 60 to 300 Hz for human speech (Huang et al. 2001). However, in *unvoiced* speech such as consonants, energy is produced when the airflow is forced through a vocal-tract constriction with sufficient velocity to generate significant turbulence. The long term spectrum of turbulent airflow tends to be a weak function of frequency (Talkin 1995), which means that the identification of a single reliable F0 value in unvoiced regions is not possible. However, in most F0 modelling approaches, F0 is assumed to be observable for all time instances[1]. Consequently, any practical F0 modelling approach must be capable of dealing with two issues:

- **Voicing classification**: classify each speech frame as voiced or unvoiced;
- **F0 observation representation**: model F0 observations in both voiced and unvoiced speech regions.

Voicing classification is often performed during F0 extraction (Kawahara et al. 1999b), and hence, the voicing label of each frame is usually assumed to be observable. Since the nature of each F0 observation depends on the type of voicing condition, voicing labels are normally considered together with F0 observations rather than being separately modelled. A widely accepted assumption for F0 values in unvoiced speech regions is that they are *undefined* and must be denoted by a discrete unvoiced symbol. Consequently, F0 is a time-varying variable whose domain is partly continuous and partly discrete. This is referred to as a *discontinuous* variable. Note that the "discontinuous" F0 does not just mean the lack of smoothness when viewed as a function of time. Real-value function can also be discontinuous in that sense. Here, the domain with mixed types of values is the essential property for being "discontinuous". Due to the mixed data types of the variable domain, discontinuous F0 observations are not readily modelled by standard HMMs. *Discontinuity* of F0 is, therefore, an essential problem to address in HMM based speech synthesis.

[1] Unobservable unvoiced F0 has also been investigated in (Ross and Ostendorf 1999).This is out of the scope of both discontinuous and continuous F0 frameworks, hence not discussed here.

One solution is to directly model discontinuous F0 observations. The *multi-space probability distribution HMM* (MSDHMM) was proposed for this purpose (Tokuda et al. 2002). In (Yoshimura 2002), this discontinuous F0 distribution is interpreted as a mixture of two distributions for continuous and discrete values, respectively. There is no explicit analysis of the relationship between voicing labels and discontinuous F0 observations. This interpretation using "a mixture of two distributions" can lead to a misunderstanding that the MSDHMM is a Gaussian mixture model (GMM). In this chapter, a formal general mathematical framework is provided for *discontinuous F0* HMM (DF-HMM) (Yu et al. 2010) and the treatment of voicing labels is discussed explicitly. MSDHMM is shown to be a special case of DF-HMM. Within the general DF-HMM framework, extensions of traditional MSDHMM are also discussed.

With a multi-space state-output distribution for discontinuous F0, HMM training can be efficiently performed and good performance can be achieved (Yoshimura 2002). However, there is still significant scope for improving F0 modelling accuracy. An alternative solution to discontinuous F0 modelling is to assume that continuous F0 observations also exist in unvoiced regions and use standard GMMs to model them. This is referred to as *continuous F0* HMM (CF-HMM) framework. A number of approaches with different independency assumption between voicing label and F0 observation have been proposed (Yu et al. 2009; Yu and Young 2011a, b).

The rest of this chapter will use consistent mathematical notations to describe the two F0 modelling frameworks in detail. The two frameworks are then compared in both theory and experiments.

9.2 Discontinuous F0 Modelling

As indicated in Sect. 9.1, a common assumption is that F0 is observable for all time instances and it has a real value in voiced regions while undefined in unvoiced regions. Since F0 values are always considered as *observable*, a specific form of representation needs to be chosen for the *observations in unvoiced regions*. A natural representation is to use a discrete symbol. F0 is, therefore, a *discontinuous* variable, whose domain is partly discrete and partly continuous, which will be denoted as f_+:

$$f_+ \in \{\texttt{NULL}\} \cup (-\infty, \infty), \tag{9.1}$$

where NULL is the discrete symbol representing the observed F0 value in unvoiced regions. It is worth noting that NULL is not a voicing label, it is an *F0 observation value* which must be introduced to satisfy the assumption that F0 is observable. Though it can be normally determined by the voicing label output from a F0 extractor, it is different from a voicing label because it is a *singleton* only used for denoting an unvoiced F0 observation.

Having introduced f_+, it is necessary to define a proper probability distribution for it. Though the domain of f_+ is a mixture of a discrete symbol and real values, a distribution can still be defined using measure theory, as shown in the appendix. The

distribution in this case is defined via the probability of events, A_{f_+}:

$$P(A_{f_+}) = \lambda^d \, \delta(f_+, \texttt{NULL}) + \lambda^c \int_{f_+ = f \in A_{f_+}} \mathcal{N}(f) \, df \tag{9.2}$$

where $f \in (-\infty, +\infty)$ denotes a real number, $\mathcal{N}(\cdot)$ is a Gaussian density of f, $\delta(\cdot, \cdot)$ is a discrete delta function defined as

$$\delta(a, b) = \begin{cases} 1 & a = b \\ 0 & a \neq b \end{cases}$$

$\lambda^d + \lambda^c = 1$ are prior probabilities of f_+ being discrete or continuous respectively and A_{f_+} is the event defined as:

$$A_{f_+} = \begin{cases} \texttt{NULL} & f_+ = \texttt{NULL} \\ (f, f + \Delta) & f_+ = f \in (-\infty, +\infty) \end{cases}$$

where Δ is a small interval. Equation (9.2) is a valid probability mass function. It is also possible to use a density-like form of Eq. (9.2) for the state output distribution in an HMM as follows

$$p(f_+) = \lambda^d \, \delta(f_+, \texttt{NULL}) + \lambda^c \, \mathcal{N}(f)(1 - \delta(f_+, \texttt{NULL})) \tag{9.3}$$

The use of the density form, Eq. (9.3), is equivalent to using the probability form, Eq. (9.2), during HMM training. Refer to the appendix for a more detailed explanation.

9.2.1 General Form Of Discontinuous F0 HMM

As discussed above, the discrete symbol NULL is different from a voicing label which in this chapter will be denoted explicitly as

$$l \in \{\texttt{U}, \texttt{V}\} \tag{9.4}$$

The issue here is that a typical F0 extractor usually outputs a *single* observation stream representing both voicing (V/U) decision and the estimate of real F0 values in voiced regions. Although the voicing decision of F0 extractors is reflected by the switching between NULL and real F0 values, it is not guaranteed to give the real voicing boundaries due to voicing classification errors. Hence, the voicing label is assumed to be *hidden* and the output distribution of f_+ for state s should, therefore, be expressed as

$$p(f_+|s) = P(\texttt{U}|s) p_{\texttt{u}}(f_+|s) + P(\texttt{V}|s) p_v(f_+|s) \tag{9.5}$$

$$= (c_{\mathrm{u}}^{s}\lambda_{\mathrm{u}}^{d} + c_{\mathrm{v}}^{s}\lambda_{\mathrm{v}}^{d})\delta(f_{+}, \mathtt{NULL}) + \big(c_{\mathrm{u}}^{s}\lambda_{\mathrm{u}}^{c}\mathcal{N}(f|s, \mathtt{U}) + c_{\mathrm{v}}^{s}\lambda_{\mathrm{v}}^{c}\mathcal{N}(f|s, \mathtt{V})\big)(1 - \delta(f_{+}, \mathtt{NULL})), \tag{9.6}$$

where $P(\mathtt{U}|s) = c_{\mathrm{u}}^{s}$ and $P(\mathtt{V}|s) = c_{\mathrm{v}}^{s}$ are state dependent voicing probabilities subject to $c_{\mathrm{u}}^{s} + c_{\mathrm{v}}^{s} = 1$, $p_{\mathrm{u}}(f_{+}|s)$ and $p_{\mathrm{v}}(f_{+}|s)$ are conditional distributions of f_{+}, which take the form of Eq. (9.3) and lead to the form of Eq. (9.6).

By definition, $c_{\mathrm{u}}^{s}\lambda_{\mathrm{u}}^{c}\mathcal{N}(f|s, \mathtt{U})$ is the likelihood contribution of the real F0 values detected within unvoiced regions. This term arises because the observed NULL symbol does not correspond exactly to the underlying voicing label l. It can be regarded as modelling erroneous voiced F0 values arising from a voicing classification error in an F0 extractor. Similarly, $c_{\mathrm{v}}^{s}\lambda_{\mathrm{v}}^{d}$ accounts for the error in misclassifying voiced speech as unvoiced. Therefore, Eq. (9.6) offers a complete framework for modelling both voicing classification and discontinuous F0 values. An HMM using Eq. (9.6) as its state output distribution is referred to as a *discontinuous F0 HMM* (DF-HMM) (Yu et al. 2010). Once DF-HMMs are trained, they can be used for classifying the voicing condition of each state and generating voiced F0 parameters during synthesis. The state voicing classification can be naturally made by comparing $c_{\mathrm{v}}^{s}\lambda_{\mathrm{v}}^{c}$ to a predetermined threshold. Then, the voiced F0 parameters can be generated from $\mathcal{N}(f|s, \mathtt{V})$. One problem with this general form of DF-HMM is that voicing labels are hidden, hence the distinction between $\mathcal{N}(f|s, \mathtt{U})$ and $\mathcal{N}(f|s, \mathtt{V})$ relies solely on the difference in statistical properties between the erroneous F0 values and the correct F0 values, which could be hard to capture.

9.2.2 Multi-Space Probability Distribution HMM

MSDHMM is a special case of DF-HMM in which voicing labels are assumed to be observable and the F0 extractor is assumed to be perfect. Therefore, the observation stream for the MSDHMM also includes the voicing label l and all terms modelling F0 extraction error will be zero

$$\lambda_{\mathrm{u}}^{c} = \lambda_{\mathrm{v}}^{d} = P(\mathtt{NULL}|\mathtt{V}) = 0 \tag{9.7}$$

$$\lambda_{\mathrm{v}}^{c} = \lambda_{\mathrm{u}}^{d} = P(\mathtt{NULL}|\mathtt{U}) = 1 \tag{9.8}$$

Eq. (9.6) then becomes[2]

$$p(\mathbf{o}|s) = p(l, f_{+}|s) = P(l)p(f_{+}|l, s) = \begin{cases} c_{\mathrm{u}}^{s} & l = \mathtt{U} \\ c_{\mathrm{v}}^{s}\mathcal{N}(f|s, \mathtt{V}) & l = \mathtt{V} \end{cases} \tag{9.9}$$

where $c_{\mathrm{u}}^{s} + c_{\mathrm{v}}^{s} = 1$ are the prior voicing probabilities. In (Yoshimura 2002), Eq. (9.9) is interpreted as using different forms of distributions for discrete and continuous

[2] Strictly speaking, $\delta(\cdot, \cdot)$ should appear in Eq. (9.9) to denote that, under the MSDHMM assumption, it is not possible to observe $(\mathtt{U}, f)$ or $(\mathtt{V}, \mathtt{NULL})$. This is omitted for clarity.

space respectively, which results in the name *multi-space* distribution. Though a GMM-like form is used in (Yoshimura 2002), it is worth noting that the state output distribution of the MSDHMM is not a mixture of expert model. From Eq. (9.9), it is clear that it is a joint distribution of voicing label and discontinuous F0 values, where due to the assumption of perfect F0 extraction, there will not be any cross-space terms. This approximation is convenient for both HMM training and voicing classification during synthesis. Hence, it has been widely used. The parameter estimation formula and details of using MSDHMM during synthesis stage can be found in (Tokuda et al. 2002).

9.3 Continuous F0 Modelling

Although the MSDHMM has achieved good performance, the use of discontinuous F0 has a number of limitations. Due to the discontinuity at the boundary between voiced and unvoiced regions, dynamic features cannot be easily calculated and hence separate streams are normally used to model static and dynamic features (Masuko et al. 2000). This results in redundant voicing probability parameters which may not only limit the number of clustered states, but also weaken the correlation modelling between static and dynamic features. The latter would then limit the model's ability to accurately capture F0 trajectories. In addition, since all continuous F0 values are modelled by a single continuous density, parameter estimation is sensitive to voicing classification and F0 estimation errors. Furthermore, due to the nature of the discontinuous F0 assumption, one observation can only be either voiced or unvoiced, but not both at the same time. Consequently, during the forward–backward calculation in HMM training, the state posterior occupancy will always be wholly assigned to one of the two components depending on the voicing condition of the observation. This hard assignment limits the possibility of the unvoiced component to learn from voiced data and vice versa. Also, it forces the voiced component to be updated using all voiced observations making the system sensitive to F0 extraction errors.

To address these limitations, an alternative solution, *continuous F0 modelling*, is proposed (Yu et al. 2009; Yu and Young 2011a, b). In this framework, continuous F0 observations are assumed to exist in both voiced and unvoiced speech regions and hence both F0 and the voicing labels can be modelled by regular HMMs, referred to as *continuous F0 HMM* (CF-HMM).

Figure 9.1 shows the relationship between discontinuous and continuous F0 modelling where Fig. 9.1a represents the discontinuous case. As it can be seen, continuous F0 assumes real F0 value for all regions, i.e.

$$f \in (-\infty, \infty). \tag{9.10}$$

Then the unvoiced F0 values have to be generated. They can be the 1-Best candidates from an F0 extractor, random samples or interpolated values between neighbouring voiced regions (Yu and Young 2011a). Another important issue is the modelling of voicing label, referred to as $l \in \{\mathtt{U}, \mathtt{V}\}$, where $\mathtt{U}$ means unvoiced and $\mathtt{V}$ voiced. It

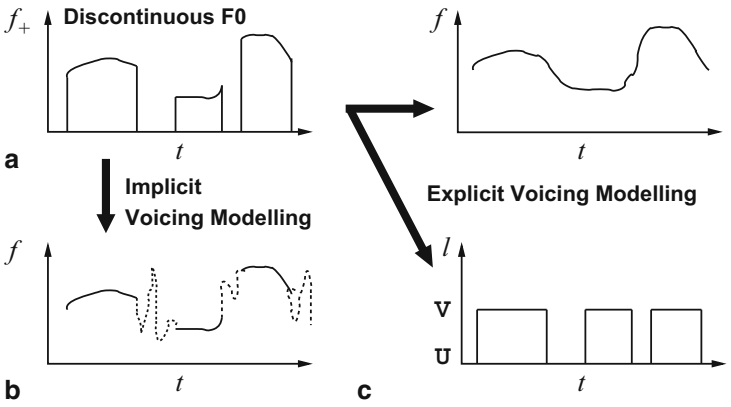


Fig. 9.1 Relationship between discontinuous F0 modelling (**a**) and continuous F0 modelling with implicitly determined voicing condition (**b**) and explicitly determined voicing condition (**c**)

can be modelled as a hidden variable (implicitly) or an observable variable (explicitly). Different treatments of voicing labels lead to different CF-HMM approaches. Figure 9.2 shows the dynamic Bayesian networks [3] of MSDHMM and various continuous F0 approaches. The upcoming sections discuss various aspects of continuous F0 modelling in detail.

9.3.1 Determining F0 in Unvoiced Regions

If F0 is considered to exist in unvoiced regions, then there must, in practice, be some method of determining it. One approach is to make use of the pitch tracker used in F0 observation extraction, such as STRAIGHT (Kawahara et al. 1999a). In many pitch trackers, multiple F0 candidates are generated for each speech frame regardless of whether it is voiced or unvoiced. A post-processing step is then used to assign voicing labels. For voiced regions, the 1-best F0 candidates are reliable. They normally have strong temporal correlation with their neighbours and form a smooth trajectory. In contrast, for unvoiced regions, the 1-best F0 candidates do not have strong temporal correlation and tend to be random. The 1-best F0 candidates of unvoiced regions can, therefore, be used as F0 observations. This will be referred to as *1-best selection*.

[3] A DBN is a graph that shows the statistical dependencies of random variables. In a DBN, a circle represents a continuous variable, a square represents a discrete variable, unshaded variables are hidden and shaded variables are observed. The lack of an arrow from A to B indicates that B is conditionally independent of A. Note that for convenience the notation of continuous random variables is also used here for the discontinuous f_+.

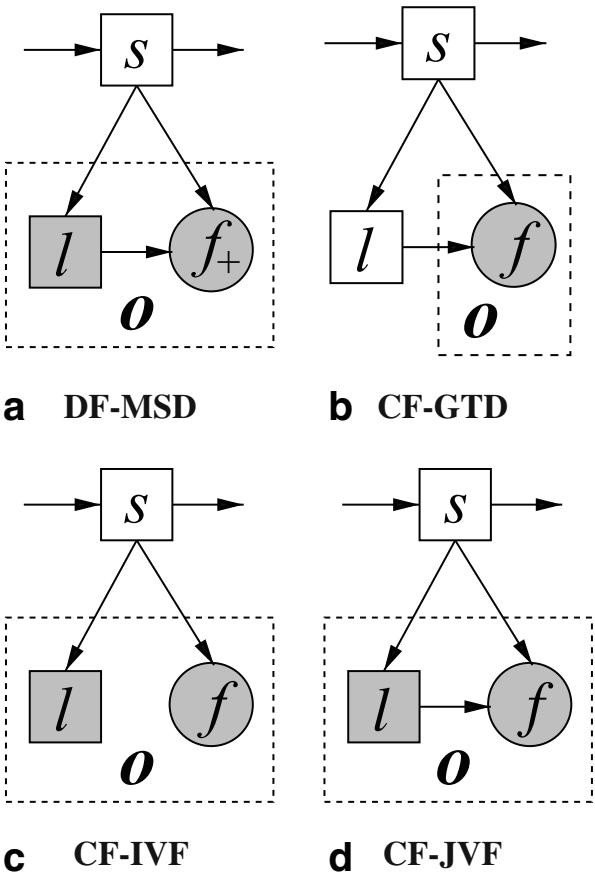


Fig. 9.2 DBN comparison between F0 modelling approaches

Note that unvoiced F0 observations near the boundaries of voiced regions may have temporal correlation which is useful when calculating dynamic features.

Other methods of determining F0 in unvoiced regions may also be used, such as sampling from a pre-defined distribution with large variance (Freij and Fallside 1988; Yu et al. 2009), using SPLINE interpolation (Lyche and Schumaker 1973) or choosing the F0 candidate which is closest to the interpolated F0 trajectory (Yu et al. 2009). Instead of one-off generation, dynamic random generation can also be used, where unvoiced F0 values are regenerated after each parameter estimation iteration (Yu and Young 2011b). It has been observed in various experiments, although synthesis quality may be occasionally affected, there is no consistent conclusion that one particular unvoiced F0 generation approach is best (Yu and Young 2011b; Yu et al. 2009). Hence, only 1-best selection is used in experiments of this chapter.

9.3.2 *Different Forms of Continuous F0 Modelling*

As indicated before, voicing label can be modelled either implicitly or explicitly. There can also be different assumptions on the dependency between F0 observations and voicing labels. These lead to different forms of continuous F0 modelling.

9.3.2.1 Continuous F0 Modelling with Globally Tied Distribution

Here, implicit voicing assumption is used, i.e. voicing label is not observable. By generating real F0 values for unvoiced regions and assuming hidden voicing labels, the continuous F0 modelling with globally tied distribution (Yu et al. 2009), *CF-GTD* in Fig. 9.2b, is obtained. If a frame is voiced then the extracted F0 value is used as the observation, otherwise some other method of computing F0 is used to derive the observation as discussed in the previous section [4].

As voicing labels are assumed to be hidden, a GMM (normally two-component) is used to model the continuous F0 observation f, with one component corresponding to voiced F0 and the other corresponding to unvoiced F0. Due to the uncorrelated nature of unvoiced F0 observations, the distribution of unvoiced F0 is assumed to be independent of the HMM states. The output distribution of an observation $\mathbf{o}$ at state s can then be written as

$$\begin{aligned} p(\mathbf{o}|s) = p(f|s) &= \sum_{l\in\{\mathrm{U},\mathrm{V}\}} P(l|s)p(f|l,s) \\ &= P(\mathrm{U}|s)\mathcal{N}(f;\mu_{\mathrm{U}},\sigma_{\mathrm{U}}) + P(\mathrm{V}|s)\mathcal{N}(f;\mu_s,\sigma_s), \end{aligned} \tag{9.11}$$

where the observation is just the continuous F0 value $\mathbf{o} = f$, $P(\mathrm{U}|s)$ and $P(\mathrm{V}|s)$ are the state-dependent unvoiced or voiced component weights respectively, $P(\mathrm{U}|s) + P(\mathrm{V}|s) = 1$. μ_{u} and σ_{u} are parameters of the globally tied distribution (GTD) for unvoiced speech, and μ_s and σ_s are state-dependent Gaussian parameters for voiced speech. Since the F0 observation is continuous, dynamic features can be easily calculated without considering boundary effects. Consequently, static, delta and delta–delta F0 features are modelled in a single stream using Eq. (9.11).

During HMM training, the initial parameters of the globally tied *unvoiced* Gaussian component can be either pre-defined or estimated on all unvoiced F0 observations. The subsequent training process is similar to standard HMM training. With global tying and random unvoiced F0 observations, the estimated parameters of the unvoiced Gaussian component will have very broad variance and be distinctive from the voiced Gaussian components which model specific modes of the F0 trajectory with much tighter variances. The state-dependent weights of the two components will reflect the voicing condition of each state. During the synthesis stage, similar to MSDHMM, the weight of the voiced component is compared to a predefined threshold

[4] As implicit voicing condition modelling requires distinct statistical properties between voiced and unvoiced distributions, the interpolation approach in Sect. 9.3.1 is not appropriate here.

to determine the voicing condition. Then the parameters of the voiced Gaussians are used to generate an F0 trajectory for voiced regions as in MSDHMM. For unvoiced states, no F0 values are generated and instead white noise is used for excitation of the synthesis filter.

With the continuous F0 assumption, the limitations of MSDHMM in Sect. 9.2.2 are effectively addressed. Since there is only one single F0 stream, there are no redundant voicing probability parameters. When using the MDL criterion in state clustering (Shinoda and Watanabe 1997) , the removal of redundancy will lead to more clustered states which may model richer F0 variations. More importantly, compared to MSDHMM, the use of a single stream introduces a stronger constraint on the temporal correlation of the continuous F0 observations and this will lead to the generation of more accurate F0 trajectories. It is also worth noting that the use of GTD not only contributes to voicing classification, it has an additional advantage. During HMM training, due to the use of multiple (two) Gaussian components, F0 observations within voiced regions are no longer exclusively assigned to voiced Gaussians. F0 extraction errors may be subsumed by the "background" GTD. This will lead to more robust estimation of the voiced Gaussian parameters than MSDHMM.

9.3.2.2 Continuous F0 Modelling with Independent Voicing Label and F0 Value

To improve the voicing classification performance, voicing labels can be assumed to be observable (Yu and Young 2011). Here, an independent data stream is introduced to explicitly model voicing labels, referred to as continuous F0 modelling with independent voicing label and F0 value (CF-IVF) in Fig. 9.2c. The state output distribution at state s is then defined as

$$p(\mathbf{o}|s) = p(l, f|s) = p(f|s)^{\gamma_f} P(l|s)^{\gamma_l}, \tag{9.12}$$

where the observation $\mathbf{o} = [f \;\; l]$, $p(f|s)$ and $P(l|s)$ are the distributions for the continuous F0 and voicing label streams respectively, γ_f and γ_l are stream weights. γ_f is set to be 1 and γ_l is set to be a very small positive value ε [5].

Since it is real valued, f is augmented by dynamic features, as in the implicit voicing case. No dynamic features are required for the voicing label l. In CF-IVF, the two streams share the same state clustering structure. Using (9.12), standard maximum likelihood HMM training can be used to estimate parameters of $p(f|s)$ and $P(l|s)$. During synthesis, state voicing status is only determined by the voicing label stream. Each state s is classified as voiced if $P(\mathrm{V}|s)$ is greater than a predefined threshold and unvoiced otherwise. The F0 trajectory is then generated using the same approach as in section MSDHMM.

[5] This means, in HMM training, the voicing labels do not contribute to the forward–backward state alignment stage but their model parameters are updated once the state alignment has been determined.

Since the voicing condition is modelled by an independent data stream, there is no requirement for the statistical properties of the voiced and unvoiced regions to be distinct. Hence, for example, SPLINE interpolation could be used in unvoiced regions in the hope that its tighter variance might lead to better trajectory modelling in V/U boundary regions (Lyche and Schumaker 1973).

In Eq. (9.12), the continuous F0 density $p(f|s)$ can have any form, for example, a single Gaussian. However, even though voicing classification is now explicit, it is still better to use the GTD model defined by (9.11) since the globally tied distribution will absorb F0 estimation errors and therefore provide more robust modelling.

9.3.2.3 Continuous F0 Modelling with Joint Voicing Label and F0 Value

Though using observable voicing labels can improve voicing classification performance, it is still weak compared to MSDHMM due to the weak correlation between the two streams. Here is a refined approach, where only one stream is used to simultaneously model both observable voicing labels and continuous F0 values, referred to as continuous F0 modelling with joint voicing label and F0 value (CF-JVF) in Fig. 9.2d. The state output distribution is

$$p(\mathbf{o}|s) \;=\; p(l, f|s) \;=\; P(l|s)\, p(f|s, l) \tag{9.13}$$

Compared to CF-IVF, CF-JVF introduces correlation between voicing labels l and continuous F0 values f and allows voicing labels to affect the forward–backward state alignment process. This will naturally strengthen the voicing label modelling. It is interesting to see that the DBN of CF-JVF is the same as MSDHMM. However, observation definition is different. In MSDHMM, each observation dimension is a discontinuous variable as defined in Eq. (9.1). In contrast, CF-JVF uses different data types for different dimensions. Each dimension is either discrete or continuous, but not mixed. Only continuous F0 dimensions require calculation of dynamic features.

It can be shown that the update formula for the parameters of $p(f|s, \mathtt{V})$ is the same as the standard ML update formula except for changing the form of state occupancy calculation (Yu and Young 2011). Although the observation of CF-JVF consists of voicing label and continuous F0 value, during decision tree based state clustering, only the continuous F0 Gaussian is considered for convenience. With this approximation, the clustering process remains unchanged. During synthesis stage, each state of the HMMs is classified as voiced or unvoiced state by comparing $P(l|s)$ to a predefined threshold.

9.4 Experimental Comparison Between MSDHMM and CF-HMM

The continuous F0 modelling techniques described above have been compared to MSDHMM on two CMU ARCTIC speech synthesis data sets (Kominek and Black 2003). A US female English speaker, `slt`, and a Canadian male speaker, `jmk`, were

used. Each data set contains recordings of the same 1132 phonetically balanced sentences totalling about 0.95 h of speech per speaker. To obtain objective performance measures, 1000 sentences from each data set were randomly selected for the training set, and the remainder were used to form a test set.

All systems were built using a modified version of the HTS HMM speech synthesis toolkit version 2.0.1 (HMM-based Speech Synthesis System (HTS)). Mixed excitation using STRAIGHT was employed in which the conventional single pulse train excitation for voiced frames is replaced by a weighted sum of white noise and a pulse train with phase manipulation for different frequency bands. The weights are determined based on *aperiodic component* features of each frequency-band (Kawahara et al. 2001). This mixed excitation model has been shown to give significant improvements in the quality of the synthesized speech (Yoshimura 2002).

The speech features used were 24 Mel-Cepstral spectral coefficients, the logarithm of F0, and aperiodic components in five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 KHz). All features were extracted using the STRAIGHT programme (Kawahara et al. 1999a). Spectral, F0 and aperiodic component features were modelled in separate streams during context-dependent HMM training.

For MSDHMM, as indicated in Sect. 9.2.2, separate streams have to be used to model each of the static, delta and delta–delta F0 features (Masuko et al. 2000). In contrast, all CF-HMM systems used a single stream for static and dynamic features of the continuous F0 observations. The CF-HMM with explicit voicing condition modelling also had an extra data stream for voicing labels. During HMM training for all systems, the stream weight for the aperiodic components was set to to be a very small positive value ϵ, similar to that of the voicing label in Sect. 9.3.2.2. MDL-based state clustering (Shinoda and Watanabe 1997) was performed for each stream to group the parameters of the context-dependent HMMs at state level. The MDL factor for MSDHMM is tuned so that it has similar number of parameters as the continuous F0 modelling techniques. The same MDL factor is used for comparing CF-IVF and CF-JVF. The duration of each HMM state is modelled by a single Gaussian distribution (Yoshimura et al. 1998). A separate state clustering process was then performed for the duration model parameters. During the synthesis stage, global variance (GV) was used in the speech parameter generation algorithm to reduce the well-known over-smoothing problem of HMM based speech synthesis (Toda and Tokuda 2007).

Figure 9.3 shows an example of the F0 trajectories generated by the two models compared to natural speech. Similar trends as shown by the objective measures can be observed: the CF-HMM F0 trajectory is a closer match to the natural speech whilst the MSDHMM has more accurate voicing classification. When listening to the speech, it can be perceived that both the natural speech and CF-HMM synthesised speech have a distinct rise at the end, whilst the MSDHMM speech was flat. In contrast, the effect of the voicing classification errors was not perceptible. Quantative comparisons are given as below.

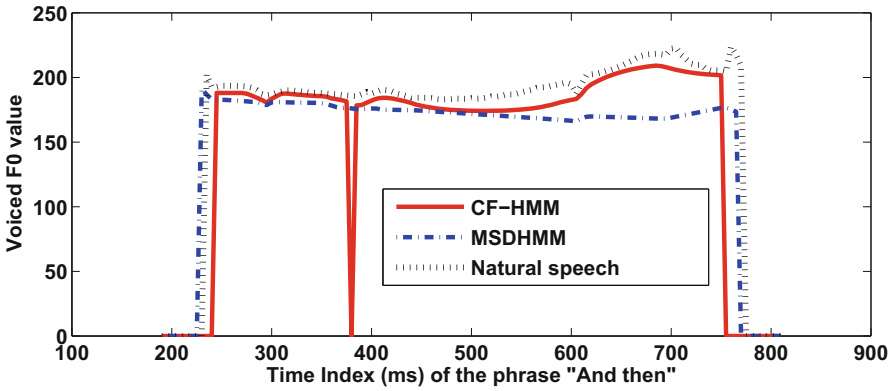


Fig. 9.3 Example F0 trajectories generated by the MSDHMM and CF-HMM models compared to natural speech

9.4.1 Objective Comparison

To quantitatively compare discontinuous and continuous F0 modelling, the root mean square error (RMSE) of F0 observations and the voicing classification error (VCE) were calculated for both the MSDHMM and CF-HMM systems. To reduce the effect of the duration model when comparing the generated F0 trajectories, state level durations were first obtained by forced-aligning the known natural speech from the test set. Then, given the natural speech durations, voicing classification was performed for each state, followed by F0 value generation within the voiced regions. By this mechanism, natural speech and synthesised speech were aligned and could be compared frame by frame. The root mean square error of F0 is defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t\in\mathcal{V}} (f(t) - f_r(t))^2}{\#\mathcal{V}}}, \tag{9.14}$$

where $f_r(t)$ is the extracted F0 observation of natural speech at time t, $f(t)$ is the synthesized F0 value at time t, $\mathcal{V} = \{t : l(t) = l_r(t) = \mathrm{V}\}$ denotes the time indices when both natural speech and synthesized speech are voiced, $\#\mathcal{V}$ is the total number of voiced frames in the set. The voicing classification error is defined as the rate of mismatched voicing labels

$$\mathrm{VCE} = 100\frac{\sum_{t=1,T} \delta(l(t), l_r(t))}{T} \tag{9.15}$$

where $\delta(l, l_r)$ is 1 if $l = l_r$ and 0 otherwise, and T is the total number of frames.

From Table 9.1, CF-HMM approaches effectively reduce the average F0 synthesis errors (RMSE) in both training and test sets compared to MSDHMM. This demonstrates the effectiveness of using continuous F0 observations. On the other hand, VCE performance becomes worse when continuous F0 assumption is used.

Table 9.1 Objective comparisons between MSDHMM and CF-HMM approaches

Data set	F0 modelling	Female		Male	
		RMSE	VCE (%)	RMSE	VCE (%)
train	MSD	16.39	4.71	12.32	5.16
	CF-GTD	11.98	17.74	8.52	18.84
	CF-IVF	11.33	7.01	9.18	8.09
	CF-JVF	10.56	6.49	8.09	6.81
test	MSD	16.65	5.85	13.37	7.17
	CF-GTD	14.67	18.36	11.12	19.49
	CF-IVF	12.58	7.29	11.90	8.43
	CF-JVF	12.87	7.12	11.13	8.13

CF-GTD has the worst performance due to weak modelling of voicing labels. By explicitly modelling observable voicing labels, CF-IVF obtains significant improvement. CF-JVF can achieve further improvement on VCE due to the strengthened correlation between voicing label and continuous F0 values. CF-JVF also improves the RMSE, i.e. F0 trajectory modelling, of most test sets except for the female test set. However, the VCEs of CF-HMM are still worse than MSDHMM. This is expected since MSDHMM assumes observable voicing labels and dependency between F0 observations, which leads to stronger voicing condition modelling.

9.4.2 Subjective Listening Tests

To properly measure performance of the synthesis systems, two forms of subjective listening tests were conducted. The CF-IVF system was first used as the representative approach of CF-HMM and then was compared with other CF-HMM approaches.

First, a mean opinion score (MOS) test was conducted. Thirty sentences were selected from the held-out test sets and each listener was presented with ten sentences randomly selected from them, of which five were male voices and the other five were female. The listener was then asked to give a rating from 1 to 5 to each utterance. The definition of the rating was: 1-bad, 2-poor, 3-fair, 4-good, 5-excellent. In total, 12 non-native and 11 native speakers participated in this test. In order to focus the evaluation on F0 synthesis, the state durations were obtained by forced-aligning the natural speech with known phone context transcriptions. Also, the spectral and aperiodic component features used were extracted from natural speech. Thus, the CF-HMM and MSDHMM models were only used to perform voicing classification of each state and generate F0 trajectories for the voiced regions. In addition, vocoded

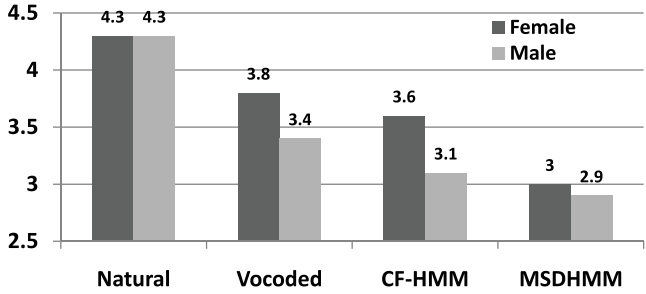


Fig. 9.4 Mean opinion score comparison of CF-HMM (CF-IVF) vs MSDHMM for F0 modelling (spectral, aperiodic component and durational features are identical across all systems). Also included for comparison are the MOS scores for natural and vocoded speech

speech [6] and natural speech were also included in the test to determine the effects of vocoder artifacts on the assessment.

Figure 9.4 shows the resulting MOS scores. It can be observed that the CF-HMM system outperformed the MSDHMM system for both male and female speakers. Vocoded speech, which may be regarded as the best possible speech that could be synthesised from any statistical model, was better than speech synthesized using either the CF-HMM or MSDHMM systems. However, the degradation from natural speech to vocoded speech was much larger than the degradation from vocoded speech to CF-HMM synthesised speech. It can also be observed that speech quality degradation of the female speaker is less than that of the male speaker. Pair-wise two-tail Student's t-tests were performed to evaluate the statistical difference between different systems. With a 95% confidence level, CF-HMM was significantly better than MSDHMM for the female speaker ($p = 0.004$), while the gain for the male speaker was not statistically significant ($p = 0.18$). This suggests that male speech is less sensitive to continuous F0 modelling. The vocoded speech was not significantly different from CF-HMM for both speakers (female: $p = 0.20$, male: $p = 0.08$). Thus, as far as statistical F0 modelling is concerned, on this data, the CF-HMM system is comparable in naturalness to vocoded speech [7].

The above MOS test used ideal duration, spectral and aperiodic component features. To compare the actual performance of complete synthesis systems, pair-wise preference tests were conducted. For the test material 30 sentences from a tourist information enquiry application were used. These sentences have quite different text patterns compared to the CMU ARCTIC text corpus and they therefore provide a

[6] Vocoded speech is the speech synthesized from the original spectral, F0 and aperiodic component features of natural speech. The only loss during this process comes from feature extraction and synthesis filter.

[7] Although there was no significant difference between CF-HMM and MSDHMM for the male speaker, the t-test showed that vocoded speech was significantly better than MSDHMM for both speakers (female: $p = 0.00005$, male: $p = 0.005$).

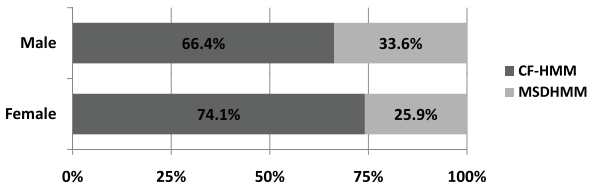


Fig. 9.5 Comparison between CF-HMM (CF-IVF) and MSDHMM

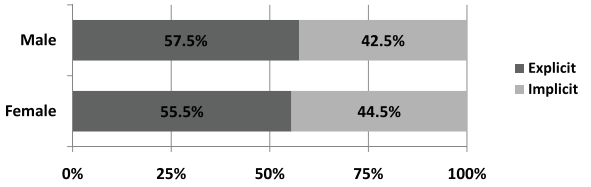


Fig. 9.6 Comparison between implicit and explicit voicing condition modelling

useful test of the generalization ability of the systems. Two wave files were synthesised for each sentence and each speaker, one from the CF-HMM system and the other from the MSDHMM system. Five sentences were then randomly selected to make up a test set for each listener, leading to 10 wave file pairs (5 male, 5 female). To reduce the noise introduced by forced choices, the 10 wave file pairs were duplicated and the order of the two systems were swapped. The final 20 wave file pairs were then shuffled and provided to the listeners in random order. Each listener was asked to select the more natural utterance from each wave file pair. Altogether 12 native and 10 non-native speakers participated in the test. The result is shown in Fig. 9.5.

It can be observed that the CF-HMM system outperformed the MSDHMM system for both male and female speakers. Statistical significance tests were also performed assuming a binomial distribution for each choice. The preference for CF-HMM was shown to be significant at 95 % confidence level (p-values for both speakers are approximately 0). Similar to the MOS test, the CF-HMM was also more dominant for the female speaker than the male speaker.

The above CF-HMM system is a CF-IVF system with 1-best selection approach for unvoiced F0 generation. It is also interesting to compare different CF-HMM approaches. First, CF-GTD is compared to CF-IVF, which is also the comparison between *implicit* and *explicit* voicing condition modelling. A panel of 21 subjects (10 non-native and 11 native speakers) was used.

As can be seen in Fig. 9.6, explicit modelling is better than implicit modelling for both speakers. This is consistent with the objective comparison in Table 9.1. Statistical significance tests showed that the difference was significant for the male speaker ($p = 0.01$) and almost significant for the female speaker ($p = 0.05$).

When explicit voicing condition modelling is used, the dependency between voicing label and F0 observation can be reserved, which leads to CF-JVF. The two approaches are compared in Fig. 9.7. It can be observed that the CF-JVF system outperformed the CF-IVF system for both male and female speakers. Statistical significance tests were performed for the result assuming a binomial distribution for

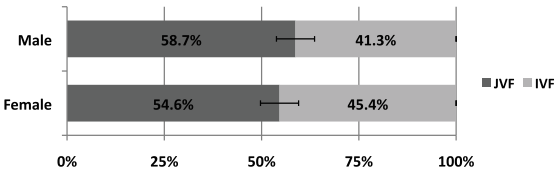


Fig. 9.7 Comparison between CF-IVF and CF-JVF. Confidence interval of 95 % is shown

each choice. The preference for CF-JVF was shown to be significant at 95 % confidence level (p-values: 0.03 for female and 0.0002 for male). This is also consistent with the objective measures.

In summary, the CF-HMM framework addresses many of the limitations of the most widely used DF-HMM approach, MSDHMM. It has been shown to yield improved performance. It is also more consistent with HMM-based speech recognition systems and it can, therefore, more easily share existing techniques, algorithms and code.

9.5 Further Analysis

Although the CF-HMM has been shown to yield significant improvement in speech quality compared to the MSDHMM in the previous section, it is not clear which aspects of the CF-HMM contribute most to the improvements. It is, therefore, useful to investigate the individual techniques used in the CF-HMM in more detail. The specific points of difference between the MSDHMM and the CF-HMM are:

1. *A single F0 stream* is used for both static and dynamic F0 features to provide a consistent voicing label probability and strong temporal correlation modelling.
2. A GTD is used to yield robust unvoiced F0 estimation.
3. *The continuous F0 assumption* avoids the problem of modelling a discontinuity at V/UV boundaries. This allows a single F0 stream to be used and it also avoids the hard assignment of state posterior during HMM training.

It is interesting to note that only the continuous F0 assumption is an inherent property of CF-HMM. A single F0 stream can also be obtained for MSDHMM by constructing dynamic F0 features at unvoiced/voiced boundaries. For example, in (Zen et al. 2001), the boundary dynamic F0 features are calculated from the nearest voiced F0 observations across the unvoiced segment. It is then possible to use a single stream for both static and dynamic F0 features as they have the same voicing boundary. GTD is also not intrinsic to the CF-HMM. From the general DF-HMM, Eq. (9.6), GTD can be easily introduced. Assuming the F0 extraction error is independent of states and combining the prior weights together, Eq. (9.6) becomes

$$p(f_+|s) = c_1^s \delta(f_+, \mathtt{NULL}) + \left(c_2^s \mathcal{N}(f|\mathtt{U}) + c_3^s \mathcal{N}(f|s, \mathtt{V})\right)(1 - \delta(f_+, \mathtt{NULL})) \tag{9.16}$$

and $c_1^s = c_\mathrm{u}^s \lambda_\mathrm{u}^d + c_\mathrm{v}^s \lambda_\mathrm{v}^d$, $c_2^s = c_\mathrm{u}^s \lambda_\mathrm{u}^c$, $c_3^s = c_\mathrm{v}^s \lambda_\mathrm{v}^c$, $c_1^s + c_2^s + c_3^s = 1$.

Table 9.2 Objective comparison between MSDHMM extensions and CF-HMM (CF-IVF)

Data set	F0 modelling	Female		Male	
		RMSE	VCE (%)	RMSE	VCE (%)
train	MSD	16.14	4.48	12.00	4.90
	+ 1str	15.94	5.76	11.53	6.68
	+ GTD	21.19	5.44	19.09	6.51
	CF-IVF	11.33	7.01	9.18	8.09
Test	MSD	16.76	5.85	13.34	6.90
	+ 1str	15.77	6.85	12.79	8.26
	+ GTD	23.44	7.06	20.25	8.10
	CF-IVF	12.58	7.29	11.90	8.43

Given that a single F0 stream and GTD can both be implemented within the DF-HMM framework, the MSDHMM can be extended to include these and thereby, allow a direct comparison with the CF-HMM. To use a single F0 stream, SPLINE interpolation is first performed for all unvoiced segments and dynamic real-valued F0 features are then constructed at the unvoiced/voiced boundaries. Consequently, a single F0 stream can be used to model the discontinuous F0 vectors, which are partly discrete NULL symbols and partly three-dimensional real-valued vectors (here only first and second derivatives are used). Furthermore, the GTD technique can be applied to the single stream MSDHMM. A globally tied Gaussian component is used as $\mathcal{N}(f|\mathtt{U})$ in Eq. (9.16) and c_1^s, c_2^s, c_3^s are updated independently given the sum-to-one constraint. The GTD component is initialized using all voiced F0 values and is never updated during HMM training[8]. During synthesis, c_1^s is compared to a pre-determined threshold (0.5 in this chapter) to determine the voicing classification for each synthesis frame.

Experiments comparing the extended MSDHMM systems and the CF-HMM system were performed to demonstrate the above. Data and experimental set up are the same as in Sect. 9.4. Again, CF-IVF is used as CF-HMM in the experiments.

From the objective comparison Table 9.2, it can be seen that compared to the standard MSDHMM, the single stream MSDHMM (MSD+1str) can slightly reduce the average F0 synthesis errors (RMSE) in both training and test sets presumably due to better temporal correlation modelling. However, it is still less accurate than the CF-HMM. The use of the GTD technique in the MSDHMM led to the worst RMSE performance. This shows that the GTD component cannot accurately capture F0 extraction errors. Instead, it will spoil the estimation of the other voiced

[8] Additional experiments showed that updating the GTD component will lead to worse performance. This is because the parameters of the GTD will be heavily affected by the dominant voiced F0 data during training. Consequently, the updated GTD component will have a small variance although globally tied. This GTD will then fail to model outliers of voiced F0 and will adversely affect the training and state clustering process.

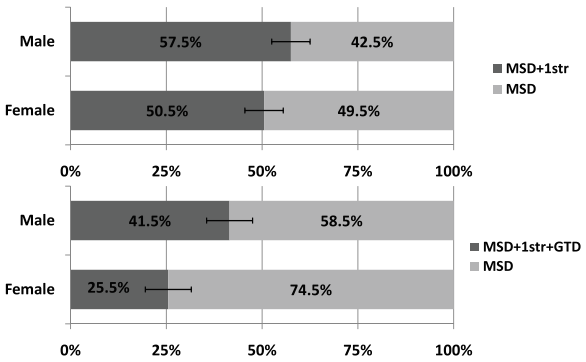


Fig. 9.8 MSDHMM vs. extended MSDHMM. Confidence interval of 95% is shown

Gaussian component because it can absorb mass from real-valued F0 observations in voiced regions. In contrast to the MSDHMM, the CF-HMM has randomly generated unvoiced F0 values which provide a strong statistical constraint (especially in the dynamic features) that prevents the GTD component from subsuming the correctly estimated voiced F0 observations. Hence, although the GTD can absorb F0 outliers and yield robust F0 estimation in the CF-HMM, it cannot do the same for the MSDHMM (Yu and Young 2011). It is worth noting that from the definition of RMSE, Eq. (9.14), only the F0 values well inside voiced regions are considered. This implies that GTD with the continuous F0 assumption does not only apply to boundary observations, it also effectively applies to normal voiced speech regions. In terms of voicing classification error, all DF-HMM approaches obtained better results than the CF-HMM. This is expected since the CF-HMM assumes independence between voicing label and F0 observations, hence the voicing label modelling is weaker. In particular, MSDHMM yielded the best VCE performance because it not only assumes observable voicing labels, but also assumes dependency between F0 observations and voicing labels.

Figure 9.8 shows the comparison between the two extended MSDHMM systems and the traditional MSDHMM. Eight native and 12 non-native listeners conducted the tests. As can be seen, the results are largely consistent with the objective measures. Using a single F0 stream improved the temporal correlation modelling and resulted in better synthesised speech. The effect on the male speaker is much stronger than the female speaker. Statistical significance tests show that the improvement on the quality of the male speech is significant at a 95 % confidence level. For the female voice, there is almost no difference when using a single F0 stream. In contrast, adding GTD to the single F0 stream MSD system significantly degraded the quality of synthesised speech for both voices. This shows that GTD alone is not directly useful within the MSDHMM framework.

Figure 9.9 shows the comparison between CF-HMM and MSDHMM with a single F0 stream, which outperformed the traditional MSDHMM. Eight native and 10 non-native listeners participated in the test. As can be seen, the CF-HMM outperformed the MSDHMM with a single F0 stream. The improvement for the female voice is

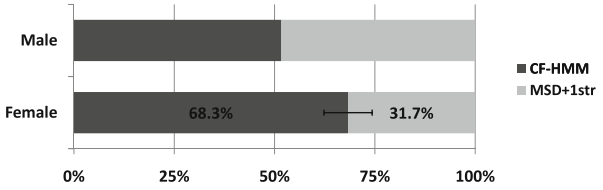


Fig. 9.9 CF-HMM vs. MSDHMM with single F0 stream. Confidence Interval of 95 % is shown

significant while insignificant for the male voice. This is expected since the single F0 stream MSDHMM achieved a significant improvement for the male voice compared to the standard MSDHMM. The only difference between the two systems in Fig. 9.9 is that the CF-HMM uses GTD with continuous F0 values, whilst the MSDHMM uses discontinuous F0 values. This shows that the continuous F0 assumption is an important factor in enabling the CF-HMM to achieve performance improvements.

Acknowledgement Most of the above works were done at the Machine Intelligence Lab, Cambridge University Engineering Department, during the author's tenure. The author would like to thank all his colleagues, especially Steve Young, Blaise Thomson and Tomoki Toda, at Cambridge University for their valuable contributions. The work was also supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning and the China NSFC project No. 61222208.

Appendix

The definition of $p(f_+)$ follows the standard approach for distributions of mixed type (discrete and continuous). (Papoulis 1984) provides discussions on the use of mixed distributions. A short discussion is included below for completeness. All terms used in this appendix are discussed in (Rudin 1987).

To define the probability distribution via measure theory, one must first define the collection of measurable events, called the σ-algebra. In the case discussed here the σ-algebra is the smallest σ-algebra containing the open intervals and also containing the singleton NULL (This exists by Theorem 1.10 of (Rudin 1987)). The probability measure, P, is defined in terms of the events, A. For values $a, b \in \mathbb{R}, a < b$, the probability function is defined as:

$$P(A) = \begin{cases} \lambda^d & A = \{\texttt{NULL}\} \\ \lambda^c \int_{f \in (a,b)} \mathcal{N}(f)\, df & A = (a,b) \end{cases},$$

where $\lambda^d + \lambda^c = 1$. Note that the probability function has only been defined in terms of open intervals and the {NULL} singleton. This is sufficient because the σ-algebra used is the smallest σ-algebra containing these sets.

Despite the use of a mixed distribution, a probability density function may still be defined by using Lebesque integration. The corresponding probability function is

defined as a function of $f_+ \in \{\texttt{NULL}\} \cup (-\infty, \infty)$ by:

$$p(f_+) = \lambda^d \, \delta(f_+, \texttt{NULL}) + \lambda^c \, \mathcal{N}(f)(1 - \delta(f_+, \texttt{NULL})). \tag{9.17}$$

This form of density function can be used in likelihood calculation during HMM training as if it were a standard density function.

To formalize the use of this function, one requires a measure to integrate over. Let the measure μ be defined as follows (with $a, b \in \mathbb{R}, a < b$):

$$\mu(\{\texttt{NULL}\}) = 1, \tag{9.18}$$

$$\mu((a, b)) = (b - a). \tag{9.19}$$

Using Lebesgue integration (Rudin 1987) of the probability density p, Eq. (9.17), with respect to this measure gives that:

$$P(A) = \int_A p \, d\mu. \tag{9.20}$$

Substituting in for the event A, the above formula in terms of traditional integration becomes (with $a, b \in \mathbb{R}, a < b$):

$$P(\{\texttt{NULL}\}) = p(\texttt{NULL}) = \lambda^d, \tag{9.21}$$

$$P((a, b)) = \int_{f \in (a,b)} p(f) df, \tag{9.22}$$

$$= \lambda^c \int_{f \in (a,b)} \mathcal{N}(f) df. \tag{9.23}$$

References

Freij, G. J., and F. Fallside. 1988. Lexical stress recognition using hidden Markov modeld. In *ICASSP*, 135–138.

HMM-based speech synthesis system (HTS) 2007. http://hts.sp.nitech.ac.jp. Accessed on 1 July, 2008.

Huang, X., A. Acero, and H. Hon. 2001. *Spoken Language Processing*. Upper Saddle River: Prentice Hall PTR.

Imai, S. 1983. Cepstral analysis synthesis on the mel frequency scale. In *Proceedings of ICASSP*, 93–96.

Kawahara, H., I. M. Katsuse, and A.D. Cheveigne.1999a. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27 (3–4): 187–207.

Kawahara, H., H. Katayose, A. D. Cheveigne, and R. D. Patterson. 1999b. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Proceedings of EUROSPEECH*, 2781–2784.

Kawahara, H., J. Estill, and O. Fujimura. 2001. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *Proceedings of MAVEBA*, Firentze, Italy, 13–15.

Kominek, J., and A. Black. 2003. CMU ARCTIC databases for speech synthesis. Language Technology Institute, School of Computer Science, Carnegie Mellon University, Tech. Rep. CMU-LTI-03–177.

Lyche, T., and L. L., Schumaker. 1973. On the convergence of cubic interpolating splines. *Spline functions and approximation theory*. Birkhauser, 169–189.

Masuko, T., K. Tokuda, N. Miyazaki, and T. Kobayashi. 2000. Pitch pattern generation using multi-space probability distribution HMM. *IEICE Transaction* J83-D-II (7): 1600–1609.

Papoulis, A. 1984. *Probability, random rariables, and stochastic processes*. U.S.: McGraw-Hill.

Ross, K. N., and M. Ostendorf. 1999. A dynamical system model for generating fundamental frequency for speech synthesis. *IEEE Transactions on Speech and Audio Processing* 7 (3): 295–309.

Rudin, W. 1987. *Real and complex analysis*, 3rd ed. New York: McGraw-Hill.

Shinoda, K., and T. Watanabe. 1997. Acoustic modeling based on the MDL principle for speech recognition. In *Proceedings of EUROSPEECH*, 99–102.

Talkin, D. 1995. A robust algorithm for pitch tracking (RAPT). In *Speech coding and synthesis*, ed. W. B. Kleijn and K. K. Paliwal, 497–516. Amsterdam: Elsevier.

Toda, T., and K. Tokuda. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems* E90-D (5): 816–824.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proceedings of ICASSP*, 1315–1318.

Tokuda, K., T. Mausko, N. Miyazaki, and T. Kobayashi. 2002. Multi-space probability distribution HMM. *IEICE Transactions on Information and Systems* E85-D (3): 455–464.

Yoshimura, T. 2002. Simultaneous modelling of phonetic and prosodic parameters, and characteristic conversion for HMM based text-to-speech systems, PhD diss, Nagoya Institute of Technology.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1998. Duration modelling in HMM-based speech synthesis system. In *Proceedings of ICSLP*, 29–32.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1999. Simultaneous modelling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proceedings of EUROSPEECH*, 2347–2350.

Yu, K., and S. Young. 2011a. Continuous F0 modelling for HMM based statistical speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing* 19 (5): 1071–1079.

Yu, K., and S. Young. 2011b. Joint modelling of voicing label and continuous f0 for hmm based speech synthesis. In *Proceedings of ICASSP*.

Yu, K., T. Toda, M. Gasic, S. Keizer, F. Mairesse, B. Thomson, and S. Young. 2009. Probablistic modelling of F0 in unvoiced regions in HMM based speech synthesis. In *Proceedings of ICASSP*.

Yu, K., B. Thomson, and S. Young.2010. From discontinuous to continuous F0 modelling in HMM-based speech synthesis. In *Proceedings of ISCA SSW7*.

Zen, H., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 2001. A pitch pattern modeling technique using dynamic features on the border of voiced and unvoiced segments. *Technical report of IEICE* 101 (325): 53–58.

Chapter 10
Use of Generation Process Model for Improved Control of Fundamental Frequency Contours in HMM-Based Speech Synthesis

Keikichi Hirose

Abstract The generation process model of fundamental frequency contours is ideal to represent the global features of prosody. It is a command response model, where the commands have clear relations with linguistic and para/nonlinguistic information conveyed by the utterance. By handling fundamental frequency contours in the framework of the generation process model, flexible prosody control becomes possible for speech synthesis. The model can be used to solve problems resulting from hidden Markov model (HMM)-based speech synthesis, which arise from the frame-by-frame treatment of fundamental frequencies. Methods are developed to add constraints based on the model before HMM training and after the speech synthesis processes. As for controls with increased flexibility, a method is developed to focus on the model differences in command magnitudes between the original and target styles. Prosodic focus is realized in synthetic speech with a small number of parallel speech samples, uttered by a speaker not among the speakers forming the training corpus for the baseline HMM-based speech synthesis. The method is also applied to voice and style conversions.

10.1 Introduction

Synthetic speech close to the quality of human utterances is now available through concatenation-based speech synthesis. However, the method requires a large speech corpus of the speaker and style to be synthesized. Thus, it is necessary to prepare a large speech corpus to realize each new voice quality with a new speaking style. Therefore, hidden Markov model (HMM)-based speech synthesis has garnered special attention from researchers, since it can generate synthetic speech with rather high quality from a smaller sized speech corpus, and can realize flexible control in voice qualities and speech styles. In this method, both segmental and prosodic features of speech are processed together in a frame-by-frame manner, and, therefore, it has

K. Hirose (✉)
Graduate School of Information Science and Technology,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: hirose@gavo.t.u-tokyo.ac.jp

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_10

the advantage that synchronization of both features is kept automatic (Tokuda et al. 2000). However, because of this, the frame-by-frame processing also includes an inherit problem when viewed from the aspect of prosodic features. Although it has the merit that fundamental frequency (F_0) of each frame can be used directly as the training datum, it generally produces oversmoothed F_0 contours with occasional F_0 undulations not observable in human speech, especially when the training data are limited. Moreover, the relationship of the generated F_0 contours with the linguistic (and para/nonlinguistic) information conveyed by them is unclear, preventing further processing, such as adding additional information, changing of speaking styles, etc. Prosodic features cover a wider time span than segmental features, and should be processed differently.

One possible solution to this problem is to apply the generation process model (F_0 model) developed by Fujisaki and his coworkers (Fujisaki and Hirose 1984). Details of the model are given in Chap. 3 of this book along with a method for the automatic extraction of model parameters from observed F_0 contours. The model represents a sentence F_0 contour in a logarithmic scale as the superposition of accent components on phrase components. These components are known to have clear correspondences with linguistic and para/nonlinguistic information, which are conveyed by prosody. Thus, using this model better control of prosody can be realized for F_0 contour generation than with the frame-by-frame control. Because of the clear relationship between generated F_0 contours and linguistic and para/nonlinguistic information of utterances, manipulation of generated F_0 contours is possible, leading to a flexible control of prosody.

We already have developed a corpus-based method of synthesizing F_0 contours in the framework of F_0 model and have combined it with HMM-based speech synthesis to realize speech synthesis in reading and dialogue styles with various emotions (Hirose et al. 2005). In that method, F_0 contours generated by HMM-based speech synthesis were simply substituted with those generated by that method. Although, improved quality is obtained for the synthetic speech generated by the method, the controlling of segmental and prosodic features independently may cause speech quality degradation due to mismatches between the two types of features.

Introducing the F_0 model into HMM-based speech synthesis is not an easy task, since F_0 model commands cannot be well represented in a frame-by-frame manner. An effort has been reported that represents F_0 model in a statistical framework to cope with the above problem, but its implementation into HMM-based speech synthesis requires some additional works (Kameoka et al. 2013). Here, two simple procedures are adopted; one to approximate the F_0 contours of training speech data with the F_0 model, and to use these F_0s for HMM training (Hashimoto et al. 2012), and the other to reshape the generated F_0 contour under the F_0 model framework (Matsuda et al. 2012).

In order to fully reap the benefits of the F_0 model in speech synthesis, a critical problem must be solved, namely how to extract the F_0 model (command) parameters from observed F_0 contours. This process needs to be done at least semiautomatically to avoid the overly time-consuming process of manual extraction. Although several methods have been developed already, their performance is less than satisfactory,

suffering from two major problems: over-extraction of accent components resulting in minor accent components not corresponding to the linguistic content and under-extraction of phrase components resulting in missing phrase components. A new method has been developed for the automatic extraction of F_0 model parameters. It uses the linguistic information of text as constraints during the F_0 model parameter extraction (Hashimoto et al. 2012).

By handling F_0 contours in the framework of F_0 model, "flexible" control of prosodic features becomes possible. A corpus-based method has been developed to predict the differences in F_0 model commands between two versions of utterances containing the same linguistic content (Ochi et al. 2009). By applying the predicted differences to the baseline version of the speech, a new version of the speech can be realized. A large speech corpus is not necessary to train F_0 model command differences. This method is applied to realize prosodic focus (Ochi et al. 2009; Hirose et al. 2012), and speaking style and voice conversions (Hirose et al. 2011).

10.2 Automatic Extraction of F_0 Model Commands

Several methods have already been developed for the automatic extraction of F_0 model commands from given F_0 contours (Narusawa et al. 2002; Mixdorff et al. 2003). The basic idea behind them is as follows: smoothing to avoid microprosodic and erroneous F_0 movements, interpolating to obtain continuous F_0 contours, and taking derivatives of F_0 contours to extract accent command locations and amplitudes. Phrase commands are extracted from the residual F_0 contours (F_0 contour minus extracted accent components) or low-pass filtered F_0 contours. Extracted phrase and accent commands are optimized by an iterative process. These methods, however, are not robust against pitch extraction errors, and produce commands not corresponding to the linguistic information of the utterances to be analyzed. Although attempts have been carried out to reduce extraction errors by introducing constraints (on command locations) induced from the linguistic information, their performance was still not satisfactory.

Interpolation of F_0 contours has a drawback since it relies on F_0s around voiced/unvoiced boundaries, where F_0 extraction is not always precise. This situation leads to the extraction of false commands. Microprosodic F_0 movements during voiced consonants may also degrade the command extraction performance, since they are not expressed in the F_0 model. To avoid false extractions, a new method is developed where F_0 contours only of vowel segments are considered. The downside being that since no F_0 is available between vowels, it becomes difficult to extract accent commands from F_0 contour derivatives. Therefore, the method is designed to take the features of Japanese prosody into account (Hashimoto et al. 2012). In Japanese, F_0s of a syllable take either high or low values corresponding to accent types. The method extracts phrase commands first viewing "Low" parts and then estimates the accent command amplitudes from the "High" parts. The method can extract minor phrase commands that are difficult to be found from the residual

F_0 contours. We can say that the method is motivated from the human process of command extraction. Since it is developed taking Japanese prosody into account, further investigations are necessary to make it applicable to other languages.

10.3 Prosodic Control in HMM-Based Speech Synthesis

10.3.1 Using F_0 Contours Approximated by F_0 Model for HMM Training

F_0 contours usually appear as quasicontinuous curves, since F_0 values are unobservable in the unvoiced parts of speech. To cope with this situation, the multispace probability distribution HMM (MSD-HMM) scheme (Tokuda et al. 1999) is widely used, where discrete HMMs (for voiced/unvoiced signs) and continuous HMMs (for F_0 values and their Δ and Δ^2 values in voiced segments) are combined. When synthesizing, F_0 contours are generated from these HMMs under the maximum likelihood criterion with voiced/unvoiced signs. Using this scheme, efficient processing both in training and synthesis is realized. It is pointed out, however, that the method has a rather limited ability to do pitch tracking and is not robust against F_0 extraction errors (including voiced/unvoiced errors) of the training corpus. Due to F_0 tracking errors or poorly pronounced vowels, a leaf for a state belonging to a vowel may contain more unvoiced occurrences than voiced ones. Thus, if that leaf is chosen, the corresponding state is judged as unvoiced. This leads to the voice quality being degraded not only by the F_0 tracking errors, but also by the *VU* decision errors in HMM training. Due to their larger dynamic F_0 ranges, the problem becomes a serious issue for tonal languages such as Chinese.

To rectify this situation, continuous F_0 HMMs have been proposed (Yu and Young 2011, see Chap. 9). In order to obtain continuous F_0 contours for the training corpus, the method selects the "most probable" F_0 values for unvoiced regions during F_0 extraction processes. Interpolation of F_0 contours can also be used for that purpose. However, the resulting continuous F_0 contours still contain unstable F_0 movements due to microprosody and F_0 extraction errors. By using F_0 contours generated by the F_0 model for HMM training, the F_0 contours generated by the HMM-based speech synthesis can be stabilized.

This idea is first applied to Mandarin speech synthesis, where F_0 extraction often includes serious voiced/unvoiced errors especially for tones with low F_0s (Wang et al. 2010). To evaluate the performance of our method as compared to the MSD-HMM, a Mandarin speech corpus of 300 sentences is divided into 270 sentences for HMM training and 30 sentences for testing. The labels of unvoiced initials attached to the corpus are used as the boundaries of the *VU* switch. The input text to the speech synthesis system includes symbols on pronunciation and prosodic boundaries, which can be obtained from orthographic text using a natural language processing system developed at the University of Science and Technology of China.

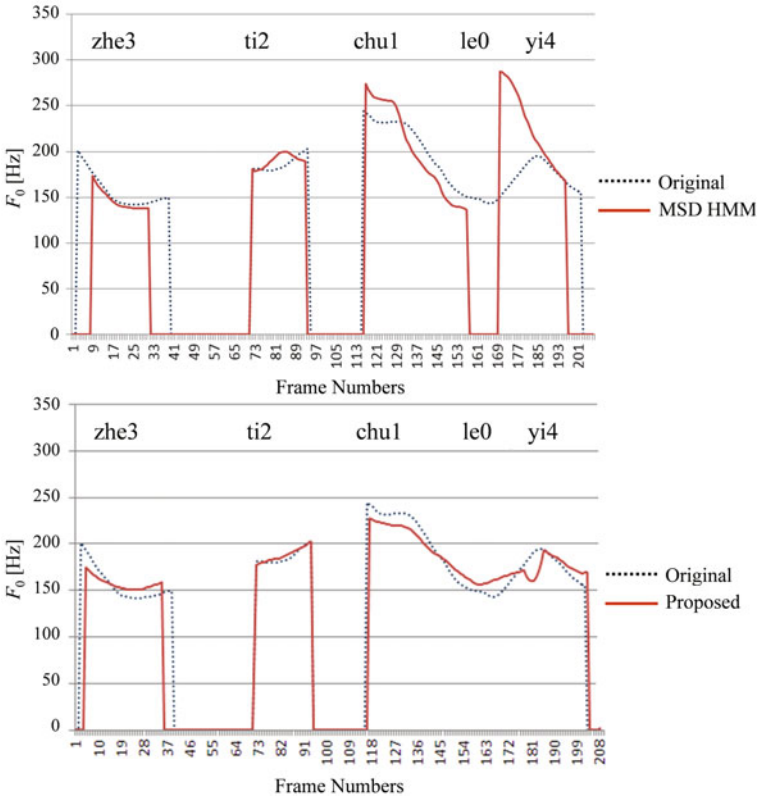


Fig. 10.1 F_0 contours generated by MSD-HMM and by the proposed method, along with corresponding original F_0 contour of natural utterance

Figure 10.1 shows examples of F_0 contours generated by MSD-HMM and by our approach, overlaid onto that of the corresponding original target utterance. The sentence is: "zhe3 + ti2 + chu1 + le0 + yi4." Here, the syllable "zhe3" (Tone 3) is difficult to synthesize correctly because of the large F_0 dynamic range in their contours and occasional creaky phonation. The syllable "le0" (neutral tone) is also hard to be synthesized correctly; reduced power and highly context-dependent F_0 contours make accurate F_0 tracking difficult. On the contrary, our method can generate F_0 contours closer to those of the original utterances with less *VU* decision errors. The validity of the method was demonstrated also through a listening experiment.

The method is then applied to Japanese. Since *VU* decision errors are few in number compared to Chinese, the MSD-HMM may work well in the case of Japanese. So a speech synthesis experiment was done using MSD-HMM. Two cases are compared; using F_0 contours extracted from speech waveforms (the original HMM) and using F_0 contours approximated by the F_0 model (proposed method). The F_0 model approximation is conducted through the automatic model command extraction explained in Sect. 10.2. It should be pointed out that accent phrase boundaries

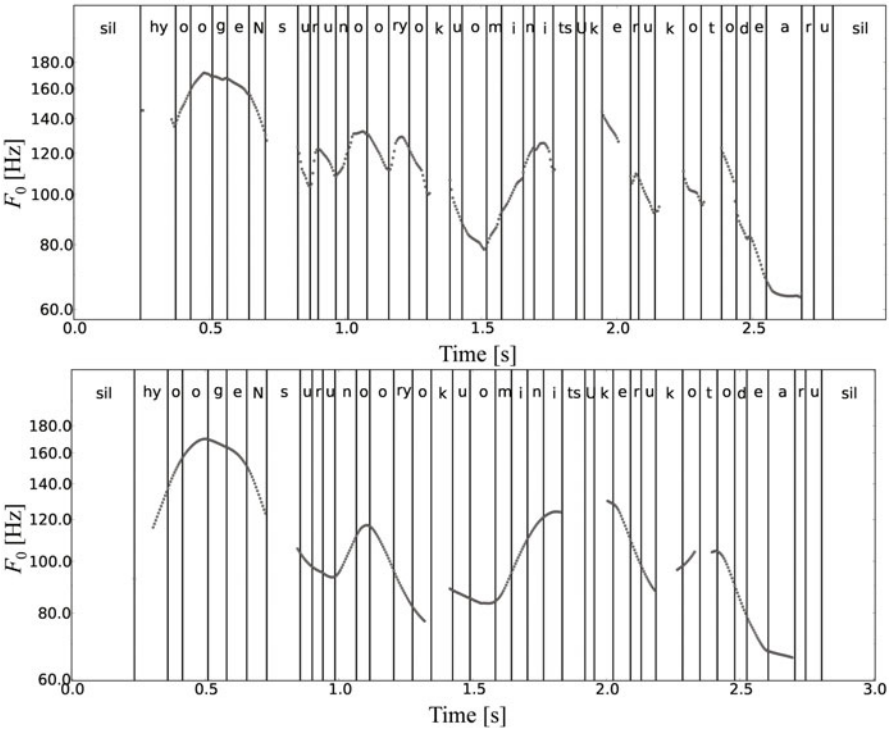


Fig. 10.2 Comparison of F_0 contours for Japanese sentence: "hyoogeNsurunooryokuo minitsukerukotodearu (It is to obtain an ability of expressing)." From *top to bottom*: F_0 contour generated by the original HMM and that generated by the proposed method

and accent types, which are given in the speech corpus, are both used in command extraction and HMM-based speech synthesis processes.

Speech synthesis experiments are conducted using the ATR continuous speech corpus of 503 sentences by speaker MHT. Out of the 503 sentences, 450 sentences are used for HMM training and rest 53 sentences are used for testing. HMM training is conducted using open software HTS-2.11[1]. Two versions of F_0 contours are prepared for the training: F_0 contours extracted from speech waveforms (original HMM), those generated by the F_0 model (proposed method). Speech signals are sampled at 16 kHz sampling rate, and STRAIGHT analysis is used to extract the spectral envelope, F_0, and aperiodicity with a 5 ms frame shift. The spectral envelope is converted to mel-cepstral coefficients using a recursive formula. The feature vector is 138 dimensional and consists of 40 mel-cepstral coefficients including the 0th coefficient, the logarithm of fundamental frequency, five band-aperiodicity (0–1, 1–2, 2–4, 4–6, 6–8 kHz) and their delta and delta–delta coefficients. A five-state left-to-right model topology is used for the HMM. Figure 10.2 compares F_0 contours

[1] http://hts.sp.nitech.ac.jp/.

generated by the original HMM and proposed method. It is clear from the figure that the F_0 contour by the proposed method is smooth.

The quality of synthetic speech from the original HMM and the proposed method is evaluated by means of a listening test with eight native speakers of Japanese. They are asked to listen to pairs of synthetic speech (one by the original HMM and the other by the proposed method) and select one from 5 scores (2: proposed method is clearly better, 1: proposed method is better, 0: no difference, -1: original HMM is better, -2: original HMM is clearly better). The average score over the 53 test sentences is 0.143 with $\pm$ 0.124 confidence interval, significant at a level of 5 %.

10.3.2 Reshaping F_0 Contours

A method was also developed to add an F_0 model constraint on the HMM-based speech synthesis before the speech waveform generation is carried out (Matsuda et al. 2012). The approach is to approximate F_0 contours generated by HMM-based speech synthesis using the F_0 model. The method first makes a decision on the initial positions of the F_0 model commands from the linguistic information and estimates their magnitudes/amplitudes from the F_0 contours generated by the HMM-based speech synthesis. During the optimization process of F_0 model commands, F_0 variance obtained through the HMM-based speech synthesis process is used to normalize the F_0 mismatch between observed F_0s and the F_0s of F_0 model; F_0 mismatch is weighted less for frames with larger variances.

To evaluate the method, speech synthesis was conducted on two Japanese native speakers' utterances (one male and one female) included in the ATR continuous speech corpus. Out of the 503 sentence utterances for each speaker, 450 utterances were used for the HMM training. Two versions of speech were synthesized for the rest of the 53 sentences; one by the original HMM-based speech synthesis and the other by the proposed method. The difference in quality between them was calculated through a listening test with 12 native speakers of Japanese. A 5-point scoring method was employed; 2 (proposed method is much better) and -2 (original HMM-based speech synthesis is much better). The total mean scores are 0.252 with a 95 % confidence interval [0.168, 0.335] and 0.230 with a 95 % confidential interval [0.148, 0.311] for male and female speakers, respectively. Clear improvements in the proposed method are observable especially in the cases when the original HMM-based speech synthesis generates erroneous F_0 contours. Figure 10.3 shows the reshaped F_0 contour compared with one generated by HMM-based synthesis.

10.4 Conversion of Prosody by F_0 Model Command Manipulation

The most significant advantage to adding an F_0 model constraint during speech synthesis is that the resulting F_0 contours are represented by F_0 model commands and can further be adjusted easily to realize flexible controls in speech synthesis. The

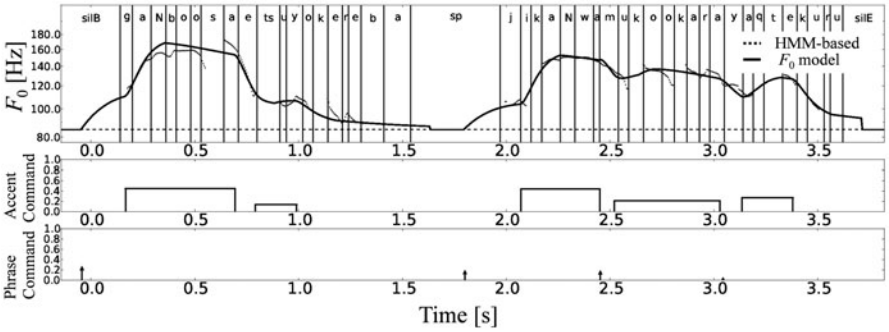


Fig. 10.3 F_0 contour reshaping by the F_0 model approximation for Japanese sentence "gaNboosae tsuyokereba jikaNwa mukookara yattekuru" (time will naturally come if (you) have a strong wish.)

method developed for prosody conversion consists of the following two processes (Hirose et al. 2011):

1. Extract F_0 model commands for the original and target utterances of the same sentence and calculate the differences in magnitudes/amplitudes of corresponding commands. Train binary decision trees (BDTs) to predict these differences.
2. Apply the differences to the phrase command magnitudes of the original utterances and then apply the differences to accent command amplitudes taking the modified phrase commands into account.

10.4.1 Prosodic Focus (Ochi et al. 2009)

Although emphasis of a word(s) is not handled explicitly in most current speech synthesis systems, its control becomes important in many situations, such as when the systems are used for generating response speech in spoken dialogue systems: words conveying information key to the user's question need to be emphasized. Emphasis associated with narrow focus in speech can be achieved by contrasting the F_0s of the word(s) to be focused on from those of neighboring words. This contrast can be achieved by placing a phrase command (or increasing phrase command magnitude, when a command already exists) at the beginning of the word(s), by increasing the accent command amplitudes of the word(s) and by decreasing the accent command amplitudes of the neighboring words. These three controls may manifest differently in each language.

In order to investigate the situation for Japanese, we selected 50 sentences from the 503 sentences of the ATR continuous speech corpus and asked a female speaker to utter each sentence without (specific) focus and with focus assigned on one of the *bunsetsus*[2]. For each sentence, 2–4 *bunsetsus* were assigned depending on the

[2] "*bunsetsu*" is defined as a basic unit of Japanese syntax and pronunciation consisting of content word(s) followed or not followed by particles.

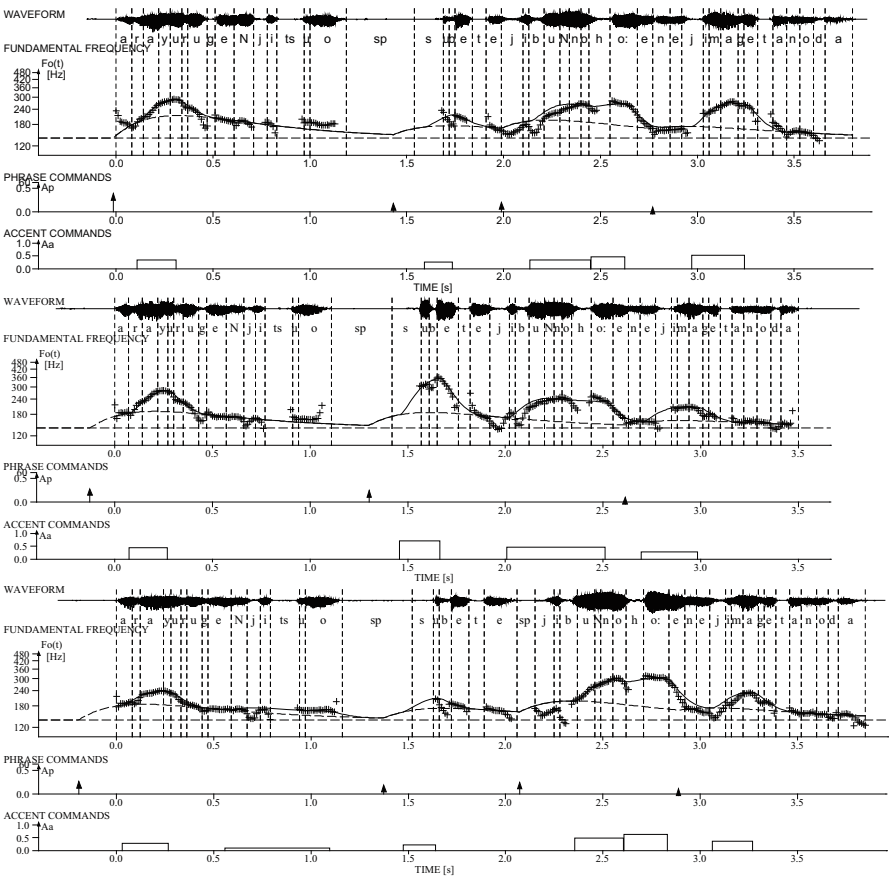


Fig. 10.4 F_0 contours and F_0 model parameters of Japanese sentence "arayuru geNjitsuo subete jibuNno hooe nejimagetanoda" ((He) twisted all the reality to his side.) uttered by a female speaker. The panels from *top* to *bottom*: without specific focus, focus on "subete," and focus on "jibuNno," respectively

sentence length. Figure 10.4 shows F_0 contours together with the results of the F_0 model approximations for utterances of the same sentence in different focal conditions. From the figure it is clear that the above three controls occur in the case of Japanese. It is also clear that there are one-to-one correspondences in phrase and accent command for different focal conditions. (Although "jibuNno hooe" has one accent command when focus is placed on "subete," it can be processed to have two commands with the same amplitude.) This one-to-one correspondence has inspired us to realize focus by controlling command magnitudes/amplitudes.

Tables 10.1 and 10.2 show input parameters for BDTs for predicting command magnitude/amplitude differences between utterances with and without focus. "BDC" in the tables denotes Boundary Depth Code, which represents the depth of syntactic

Table 10.1 Input parameters for the prediction of differences in phrase command magnitudes. The category numbers in parentheses are those for the directly preceding *bunsetsu*

Input parameter	Category
Position in prosodic phrase of current *bunsetsu*	3
Position in prosodic clause of current *bunsetsu*	4
Position in sentence of current *bunsetsu*	5
Distance from focal position (in *bunsetsu* number)	6
Length of *bunsetsu* (in number of *morae*)	4 (5)
Accent type of *bunetsu* (location of accent nucleus)	4 (5)
BDC at the boundary immediately before current *bunsetsu*	9
Existence of pause immediately before current *bunsetsu*	2
Length of pause immediately before current *bunsetsu*	Continuous
Existence of phrase command for the preceding *bunsetsu*	2
Number of *morae* between preceding phrase command and head of current *bunsetsu*	4
Magnitude of current phrase command	Continuous
Magnitude of preceding phrase command	Continuous

Table 10.2 Input parameters for the prediction of differences in accent command amplitudes. The category number in parentheses is those for the directly preceding and following *bunsetsu's*

Input parameter	Category
Position in sentence of current prosodic word	3
Position in prosodic phrase of current prosodic word	3
Position of prosodic phrase to which the current prosodic word belongs	2
Distance from focal position (in number of *bunsetsu*)	5
Accent type of *bunsetsu* (location of accent nucleus)	4 (5)
BDC at the boundary immediately before current *bunsetsu*	2
Amplitude of directly preceding accent command	Continuous
Amplitude of current accent command	Continuous
Magnitude of current phrase command	Continuous
Magnitude of preceding phrase command	Continuous

boundary and is obtainable by analyzing input sentences using the natural language parser KNP[3] (Hirose et al. 2005). The above utterances for investigation on focus control are used to train BDTs. They include 50 utterances without focus and 172 utterances with focus on one of the noun phrases (*bunsetsu* including a noun).

[3] http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html.

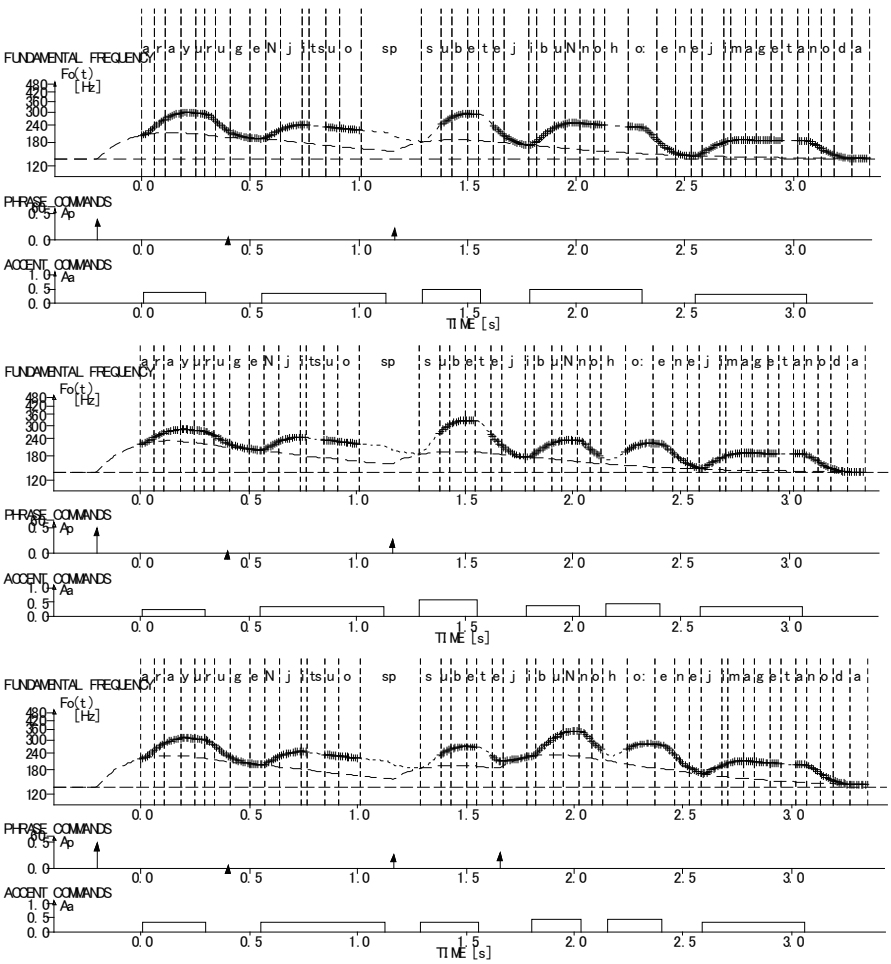


Fig. 10.5 Generated F_0 contours and F_0 model parameters. The sentence and focal conditions are the same with those shown in Fig. 10.2

As for the baseline speech synthesis on which focus control is applied, a combined method is adopted; F_0 model-based generation for F_0s with other acoustic features generated by HMM-based speech synthesis (Hirose et al. 2005). Figure 10.5 shows examples of generated F_0 contours when the predicted changes are applied to F_0 model parameters predicted by the baseline synthesis. Although prosodic focus also involves changes in pause and phone durations, they are not factored into the current experiment to focus on the effect of F_0 contours. The three controls listed above for focus control can be seen in the figure. Here we should note that the speaker used to train the command differences can be different from the one (the narrator) used for training baseline method.

In order to check the effect of the focus control for realizing emphasis, a perceptual experiment was conducted on the synthetic speech. Twenty-six sentences not

Table 10.3 Results of perceptual experiment for synthetic speech with various interpolation/extrapolation levels on the command magnitudes/amplitudes

	r	Naturalness	Emphasis
Extrapolation	1.70	2.91	4.13
	1.50	3.22	3.97
	1.30	3.50	3.89
Interpolation	1.00	3.71	4.06
	0.75	3.19	3.75
	0.50	3.50	3.50
	0.25	3.44	3.47
	0 (without focus)	3.18	2.68

included in the 50 sentences for training command magnitude/amplitude differences are selected from the 503 sentences of the ATR continuous speech corpus and one synthetic utterance is selected for each sentence; 19 utterances with focus and 7 utterances without focus. Eleven native speakers of Japanese were asked to listen to these utterances and check the *bunsetsu* on which they perceived an emphasis. An answer specifying "No emphasis" was also made available. On average, in 76.1 % of the cases, the *bunsetsus* with focus placed by the proposed method were perceived as "with emphasis." If "no emphasis" answers are excluded from the statistics, the rate increases to 83.7 %.

Modification of F_0 contours may cause degradation in synthetic speech quality. In order to investigate this point, the same 11 speakers were also asked to evaluate the synthetic speech from the viewpoint of naturalness in prosody with 5-point scoring (5: very natural, 1: very unnatural). No apparent degradation is observed from the result; 3.03 (standard deviation 1.00) for utterances with focus and 3.12 (standard deviation 0.93) for those without.

Since focus is represented with the changes in the F_0 model command magnitudes/amplitudes, emphasis levels can be controlled easily by interpolating/extrapolating the changes (Ochi et al. 2010). Experiments were conducted by selecting 64 sentences (from the 503 sentences of the ATR continuous speech corpus) not included in the set of 50 sentences for training command magnitude/amplitude differences. The predicted differences in command magnitudes/amplitudes were multiplied by the scale factor r before applying it to the command magnitudes/amplitudes predicted by the baseline method. For each sentence, a scale factor r was selected from eight levels ranging from 0 (baseline) to 1.7 as shown in Table 10.2, so that the same sentence would not appear in a series of perceptual experiment. Speech synthesis was conducted for each generated F_0 contour and in total 64 speech samples were prepared (Eight speech samples for each scale factor). Four native speakers of Japanese were asked to evaluate the naturalness andjudge the emphasis levels for the synthetic speech. The evaluation/judgment was done in this case as well with 5-point scoring. As for the emphasis levels, a score of five means strong emphasis and score of one means no emphasis. Scoring for naturalness

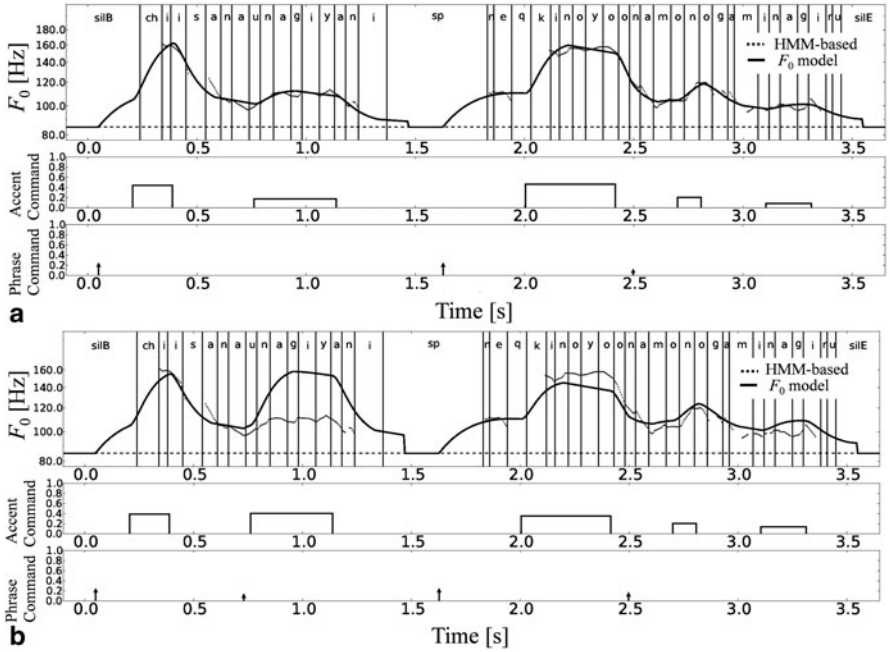


Fig. 10.6 F_0 contours and F_0 model parameters for Japanese sentence "chiisana unagiyani nekkino yoonamonoga minagiru (A small eel shop is filled with a kind of hot air)." **a** without specific focus and **b** focus on "unagiyani." F_0 contour by HMM-based speech synthesis (without specific focus) is shown for comparison

was done the same as in the former experiment. As shown in Table 10.3, emphasis levels can be changed by the interpolation/extrapolation without serious degradation in naturalness. The emphasis level is perceived as 2.68 in the case $r = 0$ (no focus). This may be due to the default focus, for which the phrase-initial word/*bunsetsu* is usually perceived as focused.

Prosodic focuses can be added in a similar way to F_0 contours reshaped by the F_0 model in Sect. 10.3.2. Figure 10.6 shows examples of (a) reshaped F_0 contour and (b) F_0 contour with prosodic focus on "*unagiyani*." It is assumed that prosodic focus can also be added to F_0 contours generated by HMM-based speech synthesis trained using F_0 model-based F_0 contours (Sect. 10.3.1). Although the F_0 model command extraction process is necessary for F_0 contours generated by the HMM-based speech synthesis before the command manipulation, from Fig. 10.2, it is expected to be achieved easily.

10.4.2 Voice Conversion (Hirose et al. 2011)

Voice conversion is a technique used to convert one voice to another without altering the linguistic (and para/nonlinguistic) contents of utterances, despite no knowledge

of these contents. Among various methods for voice conversion, those based on Gaussian mixture modeling (GMM) are widely used. In this Chapter, we take the method by Kain et al. (Kain et al. 2002) as the baseline method, where the cepstral features of original and target speakers' utterances of the same contents are tied to form joint feature vectors. Time synchrony between feature vectors is maintained through DP matching. In the method, F_0s are linearly converted using the means and standard deviations of the original and target speakers.

We replace this method with ours, which makes use of the differences in the F_0 model commands. Pause and phone durations are left unchanged. Although better prediction is possible by taking into account the linguistic information of the utterances, such as part of speech, syntactic structure, and so on, it is not included here to determine how the proposed method works with only parameters obtainable from the acoustic features of utterances.

Speech synthesis experiments were conducted using ATR continuous speech corpus of 503 sentences. Utterances by male narrator MHT are used as original utterances and those by female narrator FKS are used as target utterances. Out of the 503 sentences, 200 sentences and 53 sentences are selected, and used for training and testing (evaluation), respectively.

Ten native speakers of Japanese are asked to select the one (A or B) which is closer to X in AB-X test. A and B are synthetic speech produced by the baseline and proposed methods respectively, while X is the target speech. In order to avoid order effect, both cases with "A: original and B: proposed" and "A: proposed and B: original" are included in the stimuli. A score "1" or "$-$ 1" is assigned when speech by the proposed method is judged as being closer or farther to the target speech, respectively. When a subject cannot judge, a score of 0 is allowed. The average score over the 53 test sentences is 0.419 with $\pm$ 0.09 confidence interval at a significance level of 5 %.

10.5 Conclusions

Two methods are developed to improve the naturalness of prosody in HMM-based synthetic speech. Both are based on the F_0 model: one is to use F_0 contours approximated by the F_0 model for HMM training and the other is to reshape F_0 contours generated by the HMMs using the F_0 model. Prosodic focus is realized by manipulating the F_0 model command magnitudes/amplitudes, indicating that the F_0 model can add flexibility in prosody control. Voice conversion is also realized in the same framework.

Although the model constraint provides an improved control of F_0 contours in HMM-based speech synthesis, it has a drawback that F_0 movements not represented by the model are missing in synthetic speech. Effects of these fractional movements, such as microprosody, etc., on synthetic speech quality are assumed to be minor, but a scheme still needs to be developed to handle them properly. One possible solution is to predict these movements also in HMM-based speech synthesis.

References

Fujisaki, H., and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustic Society of Japan. (E)* 5 (4): 233–242.

Hashimoto, H., K. Hirose, and N. Minematsu. 2012. Improved automatic extraction of generation process model commands and its use for generating fundamental frequency contours for training HMM-based speech synthesis. *Proceedings of the INTERSPEECH*, 4.

Hirose, K., K. Sato, Y. Asano, and N. Minematsu. 2005. Synthesis of F_0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis. *Speech Communication* 46 (3–4): 385–404.

Hirose, K., K. Ochi, R. Mihara, H. Hashimoto, D. Saito, and N. Minematsu. 2011. Adaptation of prosody in speech synthesis by changing command values of the generation process model of fundamental frequency. *Proceedings of the INTERSPEECH*, 2793–2796.

Hirose, K., H. Hashimoto, J. Ikeshima, and N. Minematsu. 2012. Fundamental frequency contour reshaping in HMM-based speech synthesis and realization of prosodic focus using generation process model. *Proceedings of the International Conference on Speech Prosody*, 171–174.

Kain, A., and M. W. Macon. 2002. Spectral voice conversion for text-to-speech synthesis. *Proceedings of the IEEE ICASSP*, 285–288.

Kameoka, H., K. Yoshizato, T. Ishihara, Y. Ohishi, K. Kashino, and S. Sagayama. 2013. Generative modeling of speech F_0 contours. *Proceedings of the INTERSPEECH*, 1826–1830.

Kawahara, H., M. Morise, T. Takahashi, R. Nishimura, T. Irino, and H. Banno. 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F_0 and aperiodicity estimation. *Proceedings of the IEEE ICASSP*, 3933–3936.

Kurematsu, A., K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. 1990. ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech Communication* 9 (4): 357–363.

Matsuda, T., K. Hirose, and N. Minematsu. 2012. Applying generation process model constraint to fundamental frequency contours generated by hidden-Markov-model-based speech synthesis. *Acoustical Science and Technology, Acoustics Society of Japan* 33 (4): 221–228.

Mixdorff, H., Y. Hu, and G. Chen. 2003. Towards the automatic extraction of Fujisaki model parameters for Mandarin. *Proceedings of the INTERSPEECH*, 873–876.

Narusawa, S., N. Minematsu, K. Hirose, and H. Fujiaski. 2002. A method for automatic extraction of model parameters from fundamental frequency contours of speech. *Proceedings of the IEEE ICASSP*, 509–512.

Ochi, K., K. Hirose, and N. Minematsu. 2009. Control of prosodic focus in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model. *Proceedings of the IEEE ICASSP*, 4485–4488.

Ochi, K., K. Hirose, and N. Minematsu. 2010. Realization of prosodic focuses in corpus-based generation of fundamental frequency contours of Japanese based on the generation process model. *Proceedings of the International Conference on Speech Prosody*, 4.

Tokuda, K., T. Masuko, N. Miyazaki, and T. Kobayashi. 1999. Hidden Markov models based on multispace probability distribution for pitch pattern modeling. *Proceedings of the IEEE ICASSP*, 229–232.

Tokuda, K., T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. *Proceedings of the IEEE ICASSP*, 1315–1318.

Tokuda, K., T. Mausko, N. Miyazaki, and T. Kobayashi. 2002 Multispace probability distribution HMM. *IEICE Transactions on Information and Systems* E85-D (3): 455–464.

Wang, M., M. Wen, K. Hirose, and N. Minematsu. 2010. Improved generation of fundamental frequency in HMM-based speech synthesis using generation process model. *Proceedings of the INTERSPEECH*, 2166–2169.

Yu, K., and Steve Young. 2011. Continuous F_0 modeling for HMM based statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing* 19 (5): 1071–1079.

Chapter 11
Tone Nucleus Model for Emotional Mandarin Speech Synthesis

Miaomiao Wang

Abstract Variations in fundamental frequency (F_0) contours of a lexical tone make it difficult for emotional Mandarin speech synthesis. In order to better capture Mandarin tonal and emotional information, a tone nucleus model is used to carry the most important information of tones and represent the F_0 contour for Mandarin speech. After automatically estimating tone nucleus parameters from the F_0 contour, the tone nucleus part is converted to emotional speech from neutral speech. The tone nuclei variations are modeled by the classification and regression tree (CART) and dynamic programming. Compared with previous prosody transforming methods, the proposed method can avoid the data sparseness problems in emotion conversion. In addition, using only a modest amount of training data, the perceptual accuracy was shown to be comparable to that obtained by a professional speaker.

11.1 Introduction

With the intelligibility of synthetic speech approaching that of human speech, the need for increased naturalness and expressiveness becomes more palpable. However, there has been a lack of emotional affect in the synthetic speech of the state-of-art text-to-speech (TTS) systems. This is largely due to the fact that the prosodic modules in these systems are unable to predict prosody from text accurately for emotional speech. Emotional speech synthesis consists of formant synthesis, diphone concatenation, unit selection, and HMM-based methods (Schröder 2001; Yamagishi 2003). The quality of those data-driven methods heavily relies on the size of the emotional speech corpus, which takes great effort to build. Another emotional speech synthesis approach is to obtain prosodic variations between neutral speech and emotional speech, and then make the synthesized emotional speech acquire these prosodic variations. As the prosody prediction model for neutral speech has been extensively studied and implemented as robust prosodic modules in current state-of-the-art TTS systems, it would be beneficial to build the prosody prediction model for emotional

M. Wang (✉)
Institute of Data Science and Technologies, Alibaba Group, Beijing, China
e-mail: miaomiao.wmm@alibaba-inc.com

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_11

speech upon these existing systems, such as prosody conversion systems. In Tao et al., (2006) the Gaussian mixture model (GMM) and CART-based F_0 conversion methods are used for mapping neutral prosody to emotional Mandarin prosody. In Tang et al., (2008) a difference approach is adopted to predict the prosody of emotional speech, where the prosody variation parameters are predicted for each phoneme. The GMM-based spectral conversion techniques were applied to emotion conversion, but it was found that spectral transformation alone is not sufficient for conveying the required target emotion. All the research depends on correct understanding and modeling of prosodic features including F_0 contours, duration, and intensity.

F_0 modeling is important for emotional voice conversion in any language, but it is critical to Mandarin. Mandarin, the standard Chinese, is a well-known tonal language which means that the meaning of the word depends crucially on shape and register distinctions among four highly stylized syllables F_0 contour types. However, at the same time, Mandarin also allows some range of F_0 variations to express emotion, mood, and attention. Thus, F_0 modeling will be more complex than non-tonal languages such as English and Japanese. In the case of Mandarin, F_0 variations show larger undulations than those in nontonal languages. The lexical tones show consistent tonal F_0 patterns when uttered in isolation, but show complex variations in continuous speech (Chen and Wang 1995; Xu 1997). The invariance problem is the difficulty of giving a physical phonetic definition of a given linguistic category that is constant and always free of context (Lindblom 1990).

The tone nucleus model suggests that a syllable F_0 contours can be divided into three segments: onset course, tone nucleus, and offset course. The tone nucleus of a syllable is assumed to be the target F_0 of the associated lexical tone. The other two are optional and nondeliberately produced articulatory transition F_0 loci. The tone nucleus usually conforms more to the standard tone pattern than the articulatory transitions. This model has improved the tone recognition rate in Zhang and Hirose (2004) to show that the tone nucleus keeps the important discriminant information between tonal F_0 patterns and underlying tone type. Those findings lead us to the idea that we can apply the tone nucleus model to emotional speech synthesis for Mandarin. The basic assumption is that we only model the prosodic features of the tone nucleus part of each syllable, instead of directly converting the whole syllable F_0 contour, which will contain too much redundancy and cause the data sparseness problem.

11.2 F_0 Conversion Using Tone Nucleus Model (Wen et al. 2011)

11.2.1 Tone Nucleus Model

There are four basic lexical tones (referred to as T1, 2, 3, 4) and a neutral tone (T0) for each Mandarin syllable. The four basic lexical tones are characterized by their perceptually distinctive pitch patterns which are conventionally called by linguists as: high-level (T1), high-rising (T2), low-dipping (T3), and high-falling (T4) tones

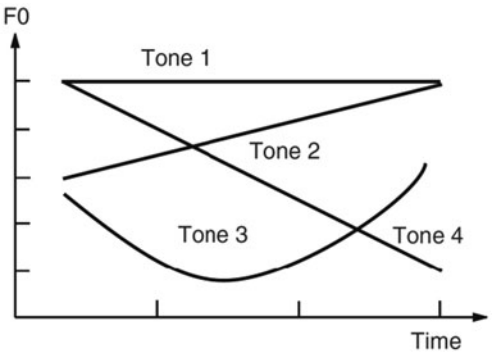


Fig. 11.1 Standard distinctive F_0 patterns of the four basic lexical tones

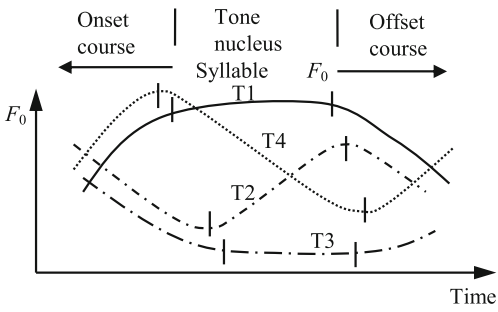


Fig. 11.2 Tonal F_0 contours with possible articulatory transitions and their tone nucleus

(Chao 1968), as shown in Fig. 11.1. The neutral tone, according to Chao (1968) does not have any specific pitch pattern, and is highly dependent on the preceding tone and usually perceived to be temporally short and zero F_0 range.

For a syllable F_0 contour, as pointed out in Zhang and Hirose (2004), lexical tone is not evenly distributed in a syllable because of F_0 variations in a syllable F_0 contour in various phonetic contexts. Only its later portion, approximately corresponding to the final vocalic part, is regarded to bear tonal information, whereas the early portion is regarded as a physiological transition period from the previous tone. It was also found that there are often cases where the voicing period in the ending portion of a syllable also forms a transition period of vocal vibration and contributes nothing to the tonality. From these considerations, we can classify a syllable F_0 contour into tone nucleus, onset course, and offset course:

1. Tone nucleus represents the target F_0 and serves as the main acoustic cue for tone perception.
2. Onset course and offset course are the F_0 variations occurring as the transitions to or from the target F_0.

Figure 11.2 illustrates some typically observed tonal F_0 variations in continuous speech and their tone nuclei notations. The tone nucleus part will conform more likely to the standard tone pattern. These F_0 values serve as distinctive features characterizing the four basic lexical tones.

Table 11.1 Describing parameters of the predictor for each tone type

Tone type	Feature of tone nucleus	Parameters to be predicted
T1	Flat F_0 with high level	Average F_0
T2	Rising F_0	Average F_0, F_0 range, template identity
T3	Very low F_0	Average F_0
T4	Falling F_0	Average F_0, F_0 range, template identity
T0	No specific feature	Average F_0

Figure 11.3 shows an example of extracted tone nucleus parts of syllables of one sentence read in different emotions. As compared in Fig. 11.3, we observe that the shape or pattern of the tone nucleus part of the syllable remains rather stable while the emotional status changes. However, the features (parameters) of the tone nucleus vary with the emotional status. For example, all T4 syllables have a similar falling tone nucleus F_0 contour, but it will have a bigger average F_0 and F_0 range in the angry and happy utterances, and a smaller average F_0 and F_0 range in a sad utterance. Moreover, it is argued in Busso et al. (2009) that F_0 contour statistics such as mean, maximum, minimum, and range are more emotionally prominent than features describing the F_0 shape. So instead of predicting the exact F_0 shape parameters like in Tao et al. (2006), we use a few F_0 contour templates to represent the tone nucleus shape. These considerations lead us to an idea of generating F_0 contours only for tone nuclei, and to concatenate them to produce the whole sentence F_0 contour Sun et al. (2012). The tone nucleus parameters are defined for each tone type as shown in Table 11.1. We use several tone nucleus templates to represent different T2 and T4 nucleus contour shapes. For T0, T1, and T3, tone nuclei are defined as a flat F_0, which is represented by a single parameter, i.e. average F_0 value.

11.2.2 Automatic Extraction of Tone Nucleus

To apply the tone nucleus model for speech synthesis and emotion conversion, it is necessary to automatically estimate tone nucleus parameters from the F_0 contour. For each syllable F_0, we use a robust tone nucleus segmentation and location method based on statistical means.

The method has two steps: the first step is F_0 contour segmentation based on the iterative segmental; K means segmentation procedure, with which a T-Test based decision of segment amalgamation is combined (Zhang and Hirose 2004). When the segmentation becomes available, which segment is tone nucleus is decided according to the following rules in the second step:

1. For T1, the segment with the biggest average F_0
2. For T2, the segment with the largest average delta F_0
3. For T3, the segment with the lowest average F_0
4. For T4, the segment with the lowest average delta F_0

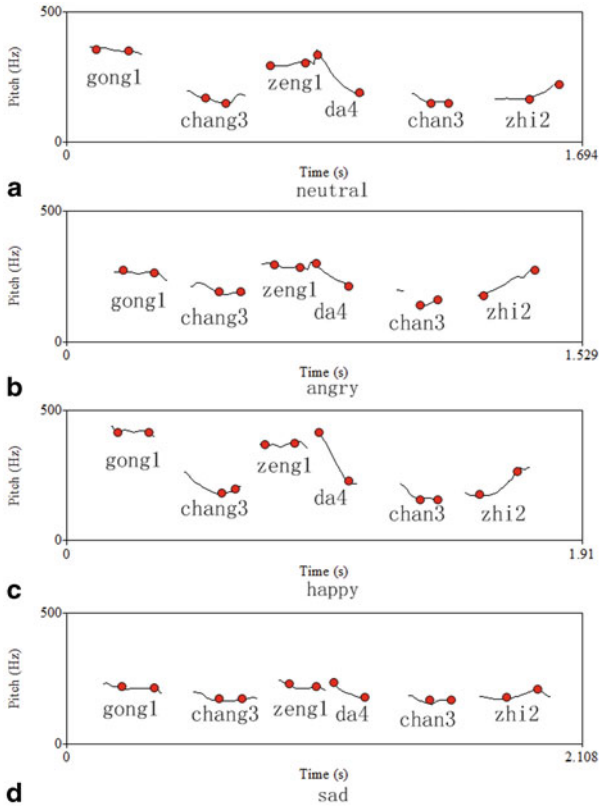


Fig. 11.3 An example sentence in different emotions: the segment between the dots indicates the tone nucleus part of the syllable

Considering that the syllable's maximum and minimum F_0 points carry important information in expressing emotions, if the chosen segment fails to cover the maximum or the minimum F_0 point, it will be expanded to include these two critical points. Since T0 shows no inherent F_0 contour, a stable definition of tone nucleus for T0 is difficult. We assume the entire voiced segment of the syllable as the tone nucleus for T0.

After extraction, average F_0 is estimated for each T0, T1, T2, T3, and T4 tone nucleus. F_0 range is estimated for each T2 and T4 tone nucleus. To obtain the tone nucleus template identity for T2 and T4, we normalize the extracted tone nucleus F_0 contours in time and frequency. Let the nucleus part of the syllable represented by $\boldsymbol{O} = (o_1, o_2 \ldots o_{11})$, the vector o_i is a two component vector (log $F_{0\mathrm{j}}$, delta log $F_{0\mathrm{j}}$). Then for T2 and T4, all $\boldsymbol{O}$s are clustered into a few (less than 10) groups using X-Means clustering method (Duda et al. 2001). For each group, an F_0 template is calculated as the average of the samples in the group. Figure 11.4a shows the templates for T4 nuclei in angry utterances. The width of the line represents the percentage of

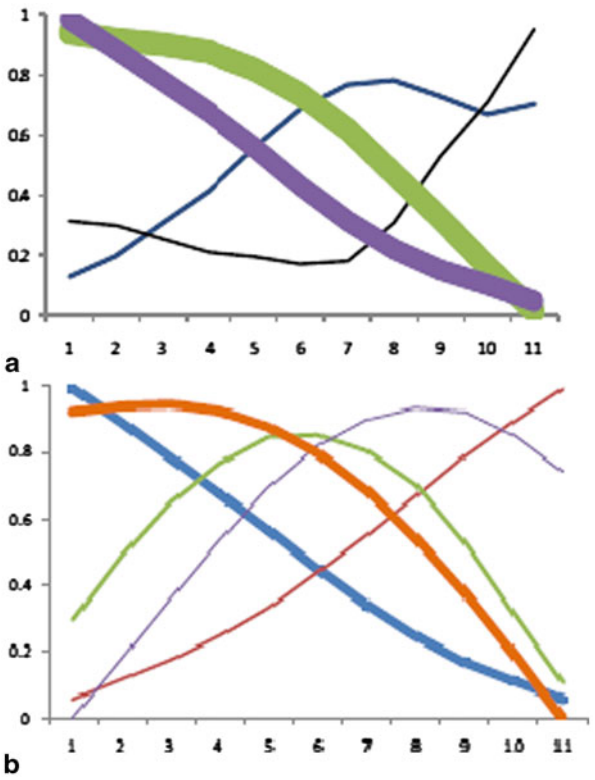


Fig. 11.4 F_0 templates for **a** angry T4 by tone nucleus **b** angry T4 by center segment

this cluster out of all the angry T4 syllables. For comparison, we averagely divide each syllable into three segments, then normalize the center segment. The clustered templates for these T4 center segments in angry utterances are shown in Fig. 11.4b. It is clear that F_0 templates of the center segment are scattered and thus hard to predict. The extracted tone nucleus templates could better capture the tone F_0 shape (e.g. a falling shape for T4) and are easier to predict. It should be noticed that 12 % of the extracted T4 nuclei have a rising shape. This may be due to several possible reasons. First, when expressing anger, the speaker sometimes adopts a rhetorical mood. Then the ending part of the utterance will have a rising F_0. Also, these rising T4 nuclei might be caused by tone co-articulation (Xu 1993) and tone nucleus extraction error.

11.2.3 F_0 Conversion

F_0 conversion is to convert a neutral F_0 contour into an emotional F_0 contour using a mapping function. The mapping function is automatically learned from the parallel speech corpus. Instead of directly mapping surface F_0 contour, tone nucleus

model parameters estimated from the F_0 contours are employed to build the mapping rules. The differences between the neutral and emotional tone nucleus parameters are modeled by classification and regression trees (CART). The input parameters of the CART contain the following:

Tone identity, including current, previous, and following tones, each with five categories

Initial identity, including current and following syllables' initial types, each with five categories

Final identity, including current and previous syllables' final types, each with two categories

Position of the current word in the current word foot/prosodic word/sentence

Part of speech (POS), including the current, previous, and following words, each with 30 categories.

11.3 Spectrum Conversion Based on GMM

Spectrum conversion can be thought of as just another form of voice conversion. The neutral speech could be regarded as the source speaker, while the target emotion speech could be regarded as the target speaker. In practice, voice conversion techniques have focused on the transformation of the vocal tract spectra, as it has been pointed out that strong feelings often literally distort the physical vocal tract. For example, "anger" often involves a physical tension which can be felt throughout the body and certainly has an effect on the tenseness of the speech organs, which in turn creates a distinct acoustic effect. Similarly, "happiness" might involve a less total physical change, often just a smile which is "talked through" (Tao et al. 2006). In Inanoglu and Young (2009) and Kawanami et al. (1999), GMM-based spectral conversion techniques were applied to emotion conversion but it was found that spectral transformation alone is not sufficient for conveying the required target emotion. In Tao et al. (2006), an integrated method based on GMM and codebook mapping is used.

GMM based voice conversion was first introduced by Stylianou et al. (1998). A GMM of the joint probability density of source and target features is employed for performing spectral conversion between speakers. Stylianou's method is to convert spectral parameters frame by frame based on the minimum mean square error. Although this method is reasonably effective, the deterioration of speech quality is caused by some problems: (1) appropriate spectral movements are not always caused by the frame-based conversion process, and (2) the converted spectra are excessively smoothed by statistical modeling. To address these problems, we use the Toda'01 method (Toda et al. 2001), which is based on the maximum likelihood estimation of a spectral parameter trajectory. Not only static, but also dynamic feature statistics are used for realizing the appropriate converted spectrum sequence.

Table 11.2 Tone error rate for emotion conversion using whole syllable F_0 contour

	Angry	Happy	Sad
Tone error rate	9.71 %	4.37 %	5.34 %

11.4 Experiments and Results

An emotional corpus contains 300 sentences with no obvious textual bias towards any of the emotional styles. A professional actor read each sentence in four basic emotional states: neutral, anger, happy, and sadness, and then each sentence was automatically segmented at the syllable level by a forced alignment procedure. Two hundred and seventy sentences, including about 1700 syllables, are used to train transforming functions and the rest are employed to test our conversion method. Our experiment uses 40 order cepstrum feature, while the number of the Gaussian mixture is 64. The STRAIGHT analysis and synthesis methods (Kawahara et al. 2008) were employed for spectral extraction and speech synthesis, respectively.

In the training procedure, neutral and other three emotional F_0 contours from the parallel corpus are first aligned according to syllable boundaries. Then tone nucleus parameters are extracted from each syllable's F_0 contour and mapping functions of the parameters are obtained. As for duration conversion, we use relative prediction which predicts a scaling factor to be applied to the neutral phone duration. The same feature set is used to train a relative regression tree. After that, the converted tone nucleus parameters are used to generate the emotional F_0 contours.

As for comparison, in the listening test, the one converted using the original syllable F_0 will have tone errors as some of the syllable will sound like other tones. Thus it greatly changes the sentence meanings. Two native speakers checked the converted sentences using original syllable F_0 contours and found that the syllable tone error rate as shown in Table 11.2, while using the tone nucleus model, doesn't have these kinds of errors.

Figures 11.4, 11.5, 11.6 respectively show the perceptual results of the synthesized emotional speech utterances with the neutral utterances. The four groups of utterances are listed below.

Natural speech (NS + NP) the original recorded utterance

Converted emotional spectrum with linearly converted prosody (CS + LCP) Spectrum is converted by GMM-based method. F_0 linearly converted from neutral F_0 contour using the following equation, where $p_t\ (x)$ and $p_t\ (y)$ are input and converted F_0 values, respectively. $\mu(\hat{A}\cdot)$ and $\sigma(\hat{A}\cdot)$ are the mean and the standard deviation of F_0, respectively.

$$p_t(Y) = \frac{p_t(X) - \mu(X)}{\sigma(X)} \times \sigma(Y) + \mu(Y) \tag{11.1}$$

Natural spectrum with converted emotional prosody (NS + CP) Converted emotional prosody using tone nucleus model is given to original recorded speech.

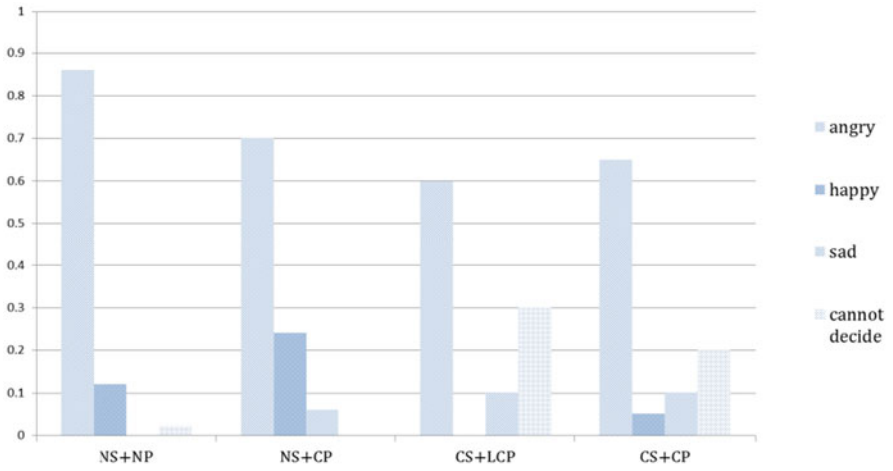


Fig. 11.5 Subjective evaluation of angry speech

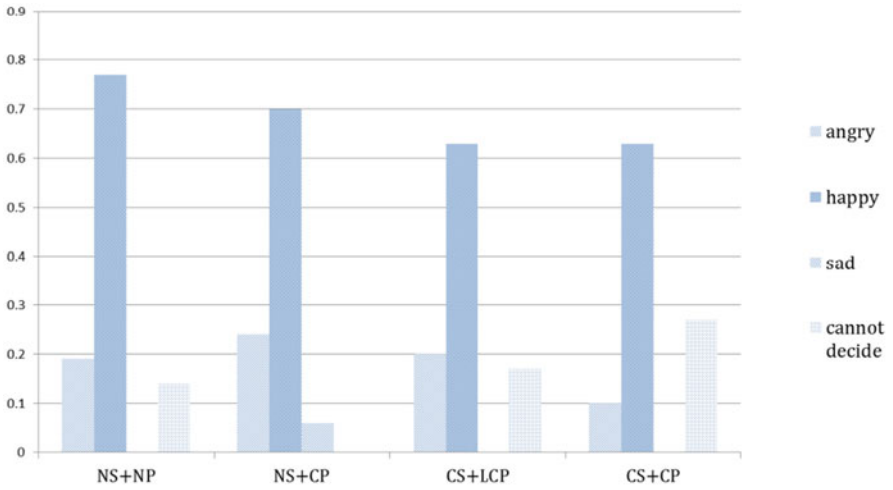


Fig. 11.6 Subjective evaluation of happy speech

Converted spectrum with converted prosody (CS + CP) Spectrum is converted to that of target emotion from neutral speech and converted emotional prosody using tone nucleus model is given to it.

As the result of the subjective experiment clearly shows, prosodic features mainly dominate emotional expression. The prosody converted using the tone nucleus model performs better than the ones using linear conversions from the neutral F_0 contour. The happy and sad emotions indicate that spectrum conversion will lower the perception rate. This may be due to a lot of *unvoiced/voiced* errors and unnaturalness caused by spectrum conversion (Fig. 11.7).

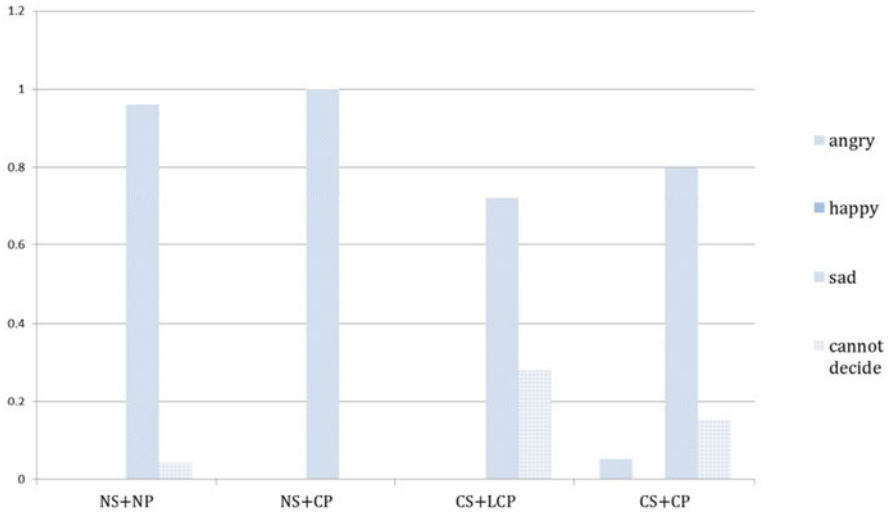


Fig. 11.7 Subjective evaluation of sad speech

11.5 Conclusions

This work mainly focuses on how to employ the tone nucleus model to implement F_0 conversion for emotional Mandarin speech synthesis. Advantages of the proposed method are that parametric F_0 models such as the tone nucleus model can provide an underlying linguistic or physiological description for the surface F_0 contour and can furnish several compact parameters to represent a long pitch contour. The CART mapping method is employed to generate transforming functions of tone nucleus model parameters. GMM-based spectral conversion techniques were also adapted to spectrum conversion. The subjective listening test shows that synthesized speech using predicted prosody parameters is able to present specific emotion.

Acknowledgment The authors of this paper would like to thank Prof. Jianhua Tao from the Chinese Academy of Sciences for offering us the emotional speech corpus and his kind advice; and Prof. Jinsong Zhang in BLCU for his useful suggestions.

References

Busso, C., S. Lee, and S. Narayanan. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech and Language Processing* 17 (4): 582–596.

Chao, Y. 1968. *A grammar of spoken Chinese*. Berkeley: University of California Press.

Chen, S., and Y. Wang. 1995. Tone recognition of continuous Mandarin speech based on neural networks. *IEEE Transaction on Speech and Audio Processing* 3 (2): 146–150.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern classification*. 2nd ed. New York: Wiley.

Inanoglu, Z., and S. Young. 2009. Data-driven emotion conversion in spoken English. *Speech Communication* 51:268–283.

Kawahara, H., M. Morise, T. Takahashi, R. Nishimura, T. Irino, and H. Banno. 2008. Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F_0 and aperiodicity estimation. *Proceedings of the IEEE ICASSP*, 3933–3936.

Kawanami, H., Y. Iwami, T. Toda, H. Saruwatari, and K. Shikamo. 1999. GMM-based voice conversion applied to emotional speech synthesis. *IEEE Transaction on Speech and Audio Processing* 7 (6): 697–708.

Lindblom, B. 1990. Explaining phonetic variation: A sketch of the H & H theory. In *Speech production and speech modelling,* ed. W.Hardcastle and A. Marchal, pp. 403–439. Dordrecht: Kluwer.

Schröder, M. 2001. Emotional speech synthesis: A review. *Proceedings of the Eurospeech* 1:561–564.

Stylianou, Y., O. Cappé, and E. Moulines. 1998. Continuous probabilistic transform for voice conversion. *IEEE Transaction on Speech and Audio Processing* 6 (2): 131–142.

Sun, Q., K. Hirose, and N. Minematsu. 2012. A method for generation of Mandarin F_0 contours based on tone nucleus model and superpositional model. *Speech Communication*, 54 (8): 932–945.

Tang, H., X. Zhou, M. Odisio, M. Hasegawa-Johnson, and T. Huang. 2008. Two-Stage prosody prediction for emotional text-to-speech synthesis. *Proceedings of the INTERSPEECH*. 2138–2141.

Tao, J., Y. Kang, and A. Li. 2006. Prosody conversion from neutral speech to emotional speech. *IEEE Transaction Audio, Speech and Language Processing* 14:1145–1153.

Toda, T., H. Saruwatari, and K. Shikano. 2001. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. *Proceedings of the IEEE ICASSP* 2: 841–844.

Wen, M., M. Wang, K. Hirose, and N. Minematsu. 2011. Prosody conversion for emotional Mandarin speech synthesis using the tone nucleus model. *Proceedings of INTERSPEECH*. 2797–2800.

Xu, Y. 1993. Contextual tonal variation in Mandarin Chinese. PhD Diss., The University of Connecticut.

Xu, Y. 1997. Contextual tonal variations in Mandarin. *J. Phonetics* 25:61–83.

Yamagishi, J., K. Onishi, T. Masuko, and T. Kobayashi. 2003. Modeling of various speaking styles and emotions for HMM-based speech synthesis. *Proceedings of the Eurospeech*. 2461–2464.

Zhang, J., and K. Hirose. 2004. Tone nucleus modeling for Chinese lexical tone recognition. *Speech Communication* 42 (3–4): 447–466.

Chapter 12
Emphasis, Word Prominence, and Continuous Wavelet Transform in the Control of HMM-Based Synthesis

Martti Vainio, Antti Suni and Daniel Aalto

Abstract Speech prosody, especially intonation, is hierarchical in nature. That is, the temporal changes in, e.g., fundamental frequency are caused by different factors in the production of an utterance. The small changes due to segmental articulation—consonants and vowels—are different both in their temporal scope and magnitude when compared to word, phrase, and utterance level changes. Words represent perhaps the most important prosodic level in terms of signaling the utterance internal information structure as well as the information structure that relates an utterance to the discourse background and other utterances in the discourse. In this chapter, we present a modeling scheme for hidden Markov model (HMM)-based parametric speech synthesis using word prominence and continuous wavelet transform (CWT). In this scheme emphasis is treated as one of the extrema in the word prominence scale, which is modeled separately from other temporal scales (segmental, syllabic, phrasal, etc.) using a hierarchical decomposition and superpositional modeling based on CWT. In this chapter, we present results on both automatic labeling of word prominences and pitch contour modeling with an HMM-based synthesis system.

12.1 Introduction

Modern statistical parametric synthesizers are capable of very high quality speech in terms of intelligibility while suffering from a lack of naturalness, whereas the situation is reversed with the best systems based on stored waveforms. Parametric speech synthesis allows for modeling prosody in a very flexible way and it could be argued that at least some of the added intelligibility of the best systems is due to the lack of discontinuities in the parametric tracks. The systems also tend to produce average values with small variances diluting the impact of possible phonological

M. Vainio (✉) · A. Suni · D. Aalto
Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland
e-mail: martti.vainio@helsinki.fi

A. Suni
e-mail: antti.suni@helsinki.fi

D. Aalto
e-mail: daniel.aalto@helsinki.fi

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_12

errors caused by faulty predictions. On the other hand, the averaging, which relates to the general redundancy in speech, has hidden the fact that most synthesis systems use a minimal amount of linguistic information for predicting prosody and rely on the HMM framework's ability to model, e.g., the pitch contour in a natural and controlled manner. The speech parameter track generation in HMM synthesis is based on both static and dynamic features (delta and delta-delta), which constrain the track so that it is free of artificial discontinuities (Tokuda et al. 1995, 2000). Thus, in a phonetic sense, the HMMs work very well. That is, they are capable of modeling any parameter track in a natural fashion that reflects the articulatory constraints in the modeled speech data; the phonetic output of the system is generally *appropriate* in terms of parameter movement. Therefore, the HMM framework can best be seen as a statistical mapping between discrete symbolic input and continuous phonetic output. Because of the appropriateness of the phonetic output, it makes no sense to use anything directly related to phonetic form as input to the mapper. For instance, HMM systems are capable of producing extremely good formant tracks from contextually determined phone or phoneme input—instructing such systems directly how the formants should behave would only confuse the mapping. Similarly, with respect to controlling prosody, the systems should not get formal input with respect to the relevant parameters such as f_0.

In traditional synthesis systems, prosody is controlled with distinct modules, which translate some form of linguistic input into quantitative values. They may, for instance, produce timings and magnitudes for further processing by quantitative models, which in turn produce the desired parameter tracks (e.g., Fujisaki and Hirose 1984; Taylor 2000; Bailly and Holm 2005a). Traditionally this was done by rules, which have in most instances been replaced today by statistical models. That is, the models are trained from data to produce a mapping between the discrete linguistic input and a continuous acoustic output. This calls for a priori knowledge concerning the forms that the system should be able to handle. In systems based on phonological representations of prosody (e.g., ToBI), this requires a two-phase approach, where the predicted label sequence is interpreted quantitatively by a phonetic component. In a trainable system, this calls for two separate statistical models to have an effect on any given phonetic form; one that predicts the symbolic labels and one that predicts the contours. It is easy to see how such a system could be in internal formal conflict due to, for instance, inadequately labeled training data or bad predictions that will not fit the training data. In order to minimize the possibility of such internal conflicts regarding form, one should separate the formal descriptions from functional ones as much as possible. That is, the control of prosody in an HMM-based system should reflect the control of segmental output. To that end, the training data for the synthesis need to be annotated at a suitable level of abstraction. We argue, that word prominence is perhaps the best at such a level. It is both functional—thus avoiding the phonetic form problem—and theory neutral. That is, prominences can be automatically estimated from speech signals and the estimation methods can be evaluated against human prominence judgements.

In this chapter, we inspect the use of continuous wavelet transform (CWT) as a method for analysing, estimating, and modeling prosody in text-to-speech (TTS)

synthesis. Although, the presented methods are applicable to all continuous prosodic signals, we concentrate on the modeling of pitch contours.

The fundamental frequency (f_0) contour of speech contains information about different linguistic units at several distinct temporal scales. Likewise prosody in general, f_0 is inherently hierarchical in nature. The hierarchy can be viewed in phonetic terms as ranging from segmental perturbation (i.e., microprosody) to levels that signal phrasal structure and beyond (e.g., utterance level downtrends). In between, there are levels that signal relations between syllables and words (e.g., tones and pitch accents). Consequently, the pitch movements happen on different temporal scales: the segmental perturbations are faster than typical pitch accents, which are faster than phrasal movements and so on. These temporal scales range between several magnitudes from a few milliseconds to several seconds and beyond.

Traditionally, the hierarchy has been approached by using superpositional models that separate syllable and word level accents from phrases (Fujisaki and Sudo 1971; Öhman 1967). Although superpositional models, such as the command-response model (Fujisaki and Sudo 1971; Fujisaki and Hirose 1984), can in principle model any contour, estimating the parameters for the model is difficult. Several superpositional models with a varying degree of levels have been proposed since Fujisaki (Bailly and Holm 2005b; Anumanchipalli et al. 2011; Kochanski and Shih 2000; Kochanski and Shih 2003). Superpositional models attempt to capture both the chunking of speech into phrases as well as the highlighting of words within an utterance. Typically smaller scale changes, caused by, e.g., the modulation of the airflow (and consequently the f_0) by the closing of the vocal tract during certain consonants, are not modeled.

In HMM-based speech synthesis paradigm, all modeling is based on phone-sized units. In principle, slower intonation patterns are more difficult to model than segmentally determined ones. Moreover, the statistical procedure of decision-tree (DT) clustering highlights instances that are more common, resulting in a good reproduction of microprosody and overall trends (such as general downtrends) and relatively poor reproduction of prosody at the level of words and phrases. This shortcoming calls for methods that take into account the inherent hierarchical nature of prosody.

In this chapter, we propose a unified approach for annotation, modeling, and manipulation of prosody, based on CWT. The CWT is used to decompose the f_0 contour into several temporal scales ranging from microprosody to utterance levels of prosodic hierarchy. In subsequent chapters, we sketch how this decomposition can be utilized for word prominence annotation, HMM-modeling, and production of emphasis.

12.2 Word Prominence and Emphasis

Prominence can be seen as a functional phonological phenomenon that signals syntagmatic relations of words within an utterance by highlighting some parts of the speech signal while attenuating others. Thus, for instance, more prominent syllables

within a word stand out as stressed (Eriksson et al. 2001). At word level, prominence relations can signal how important the speaker considers each word in relation to others in the same utterance. These, often information-based relations, range from simple phrasal internal structures (e.g., prime minister, yellow car) to relating utterances to each other in discourse as in the case of contrastive focus (e.g., "Was the bus full? No, we WALKED here."). Although prominence probably works in a continuous fashion, it is relatively easily categorized in, e.g, four levels where the first level stands for words that are not stressed in any fashion prosodically to moderately stressed and stressed and finally words that are emphasized (as the word WALKED in the example above). These four categories are fairly easily and consistently labeled even by nonexpert listeners (Cole et al. 2010; Vainio et al. 2009; Arnold et al. 2012). In sum, prominence functions structure utterances in a hierarchical fashion that directs the listener's attention in a way that enables the understanding of the message in an optimal manner.

Thus, the prominence of a given word has a very complex relationship with the actual prosody of the utterance as well as other factors that are more linguistic in nature: while fundamental frequency is the main determinant of prominence, it can be systematically modified by other prosodic variables (segmental duration and intensity) as well as the linguistic structure of the utterance (Vainio and Järvikivi 2006; Cole et al. 2010). For instance, the first word in a read paragraph often has a higher f_0 than the other words unless some of them are emphasized for some reason. In absolute terms, it would be easy to equate the two. Listeners are, however, able to separate the effect of what could be called a *paratone* from the utterance level prominences and the initial word will not be equated in prominence with the, say, contrastively focused one in the same paragraph.

In terms of a continuous scale, word prominence ranges from a totally nonprominent to emphatic, where nonprominent can be interpreted as unaccented, and emphatic as narrow prosodic focus. Thus, emphasis is simply one end in the prominence scale that is determined to reflect the linguistic contrasts in the speech signal. To this end, intonation in TTS can be modeled by assigning appropriate prominence levels to each word in an utterance and mapping those values to the parameter tracks with HMMs or any other suitable statistical method (e.g., artificial neural networks).

TTS systems are typically trained for neutral speech that has a fairly narrow range of prosodic variation and such phenomena as contrastive (narrow) focus do not often occur in the data. Reciprocally, these phenomena are also extremely difficult to predict on the basis of linguistic analysis of the text. Thus, modeling emphasis has not been at the center of TTS research. With respect to the phonetics of synthetic speech, the phenomena are extremely interesting and relevant and should be modeled as part of the general dynamics of speech production.

The continuous nature of prominence and the hierarchical nature of prosody call for methods that allow both for the analysis and synthesis of speech in a coherent framework. The CWT offers such a possibility by being hierarchical (in terms of temporal scope) and the signal can be approximately recovered even after being reduced to a few number of scales. That is, it allows for prosodic analysis and

synthesis on linguistically and phonetically relevant scales that can be further mapped with linguistic analysis of text.

In what follows, we will describe the CWT-based hierarchical prosodic analysis, the prominence estimation methods based on the analyses, as well as the synthesis of pitch contours from the inverse of the CWT. In particular, we will describe two methods for estimating word level prominences based on (1) a single temporal scale at the level of words and (2) a more temporally aligned method that takes into account the cumulative effect of different scales; lines of maximal amplitude (LoMA).

12.3 Continuous Wavelet Transform

The time evolution of the fundamental frequency consists of several components that are produced by different mechanisms and are separated by the auditory system. Time-frequency analysis can be used to disentangle rapid (segmental, syllabic; accent, tone) and slow (phrase, utterance; declination) components. In time-frequency analysis, the temporally varying signal (here f_0) is decomposed in signal tiers that have a restricted spectral range. The earliest of such tools is the (windowed) Fourier analysis where the strengths of (localized) sinusoidal oscillations at different frequencies are gathered to represent the signal. More recently, wavelets have been used to complement the Fourier analysis. In contrast to the windowed Fourier analysis, the wavelet window size varies as a function of the scale (corresponding to the Fourier frequency). Hence, the scale corresponding to fast oscillations uses a short window, whereas the slow oscillations are treated with a longer window (See Fig. 12.1 for an example.).

It should be remarked that there are several wavelet techniques available. Discrete wavelet transforms are often used for compressing signals since they offer almost exact recovery of a signal that is efficient and robust against noise. CWTs are computationally less efficient but offer a possibility to follow the time-scale structure more accurately.

12.3.1 Definition of the Continuous Wavelet Transform

Any signal, including the f_0 contour or the gain, can be decomposed by CWT and this process can be inverted. The CWT of a signal describes the content of the signal at various scales. Indeed, the CWT has an additional dimension scale. To define the CWT, let s be a real-valued time-varying signal with finite energy, i.e.,

$$\int_{-\infty}^{\infty} |s(t)|^2 dt < \infty.$$

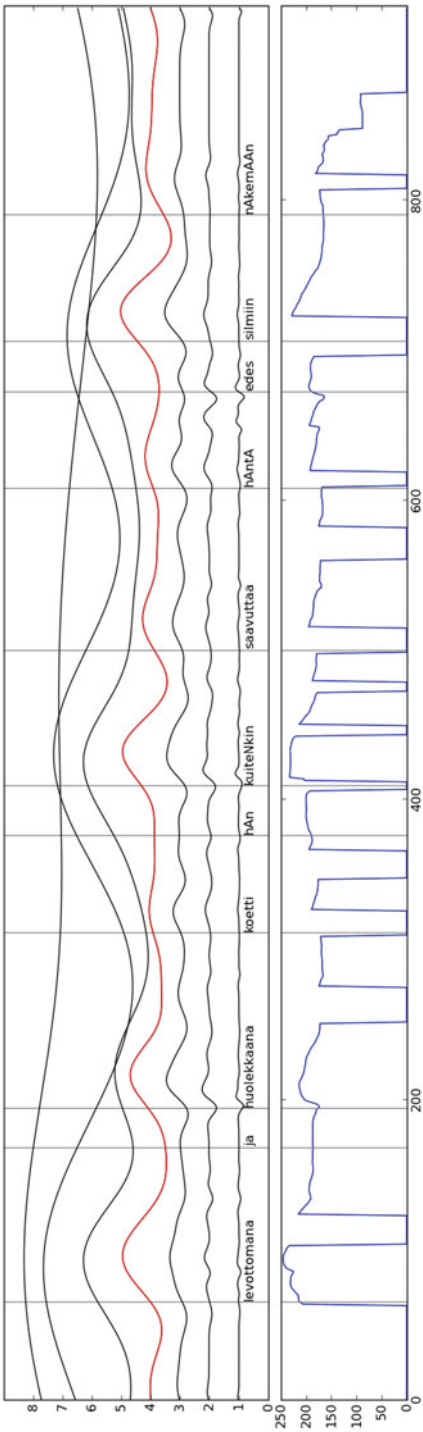


Fig. 12.1 The word prosody scale is chosen from a discrete set of scales with ratio 2 between ascending scales as the one with the number of local maxima as close to the number of words in the corpus as possible. The *upper panel* shows the representations of f_0 at different scales. The word level (4.2 Hz; see text) is drawn in *red*. The *lower panel* shows the f_0 curve. The *abscissa* shows the frame count from the beginning of the utterance (5 ms frame duration)

In practice, all the signals considered here have both finite size and finite duration leading to a bounded s with finite support whence the finite energy condition is satisfied. The CWT Ws of the signal s is defined for scale $\sigma > 0$ and translation τ through a convolution as

$$Ws(\sigma, \tau) = \sigma^{-1/2} \int_{-\infty}^{\infty} s(t)\psi\left(\frac{t-\tau}{\sigma}\right) dt \tag{12.1}$$

where ψ is the mother wavelet. In what follows, we will use the Mexican hat mother wavelet defined as

$$\psi(t) = \frac{2}{\sqrt{3\sqrt{\pi}}}(1-t^2)\exp(-t^2/2).$$

Integration by parts of the formula (12.1) shows that, for regular signals (s two times differentiable), $Ws(\sigma, \tau)$ can be interpreted as the second derivative of the signal convolved with a Gaussian filter (depending on the scale σ).

12.3.2 Approximate Recovery of a Signal from Few Wavelet Scales

In order to be applicable to synthesis, the wavelet transformed signal has to be recoverable from the analysis. An efficient reconstruction of the signal starting from the wavelet transform Ws has been established by Grossmann and Morlet (see also Kronland-Martinet et al. 1987, Eq. 7, p. 13):

$$s(t) = \frac{1}{C_\psi} \int_0^{\infty} Ws(\sigma, t)\sigma^{-3/2} d\sigma \tag{12.2}$$

where

$$C_\psi = \int_0^{\infty} \frac{\left|\hat{\psi}(\omega)\right|}{\omega} d\omega$$

and $\hat{\psi}$ is the Fourier transform of the mother wavelet ψ. For the proof of the identity (12.2), see the original argument using Fourier analysis techniques (Grossman and Morlet 1985, p. 160). For other inversion formulas and historical background, see Mallat's or Daubechies's book (Mallat 1999, p. 122, Theorem 4.3 or Daubechies et al. 1992, p. 24, Proposition 2.4.1).

The CWT requires infinitely many scales to be included in the representation of the decomposed signal. We would like to reduce the integral in (12.2) to a finite sum with a small number of terms. The above derivations assume a continuous signal $s(t)$ that is infinitely long. The results are easily adapted to digital signals with finite duration. Assuming that the signal starts at $t = 0$ and has a duration $T > 0$ (i.e., the support of s is the interval $[0, T]$), the coarsest scales above $\sigma = T$ do not reflect the inherent behavior of the signal but rather its relation to the padded zeros

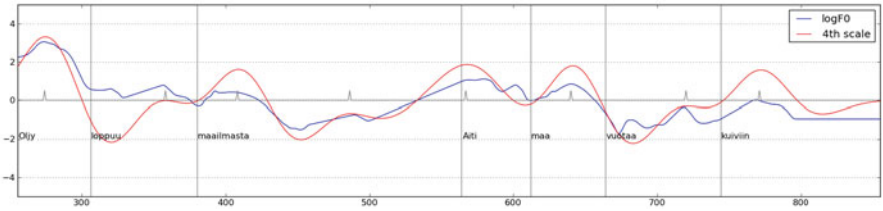


Fig. 12.2 Comparison of selected word scale and original f_0 contour with detected peaks marked with gray triangles. Observe that the wavelet contour is free of noise and declination trend.

before the onset and after the offset. These border problems are often solved by force periodizing the signal by copying the signal to the intervals $(T, 2T]$, $[-T, 0)$ etc., as is adopted here. In the context of intonation contour, mirroring has been used (Kruschke and Lenz 2003). Putting aside for a moment the potential artifacts caused by the borders, we continue now the discretization of the scale parameter. Let $a > 1$ and let us compute the convolutions $Ws(a^j, .)$ for $j = 1, 2, \ldots, N$, only. Following the reasoning of Mallat (1999, p. 135), we can discretize the reconstruction formula (12.2) to arrive at

$$s(t) \approx \frac{\log a}{C_\psi} \sum_{j=1}^{N} a^{-j/2} Ws(a^j, t) =: \sum_{j=1}^{N} W_j s(t)$$

where log stands for the natural logarithm and $W_j s$ is the summand for each $j = 1, \ldots, N$. The above formula coincides with the widely used numerical approach (see Torrence and Compo 1998 for more details and efficient numerical implementation using fast Fourier transform). The choice of a and the number of included scales N will be $a = 2$ and $N = 5$ unless otherwise stated.

12.3.3 Lines of Maximum Amplitude (LoMA)

CWT is a highly redundant way of representing a signal. For coarse scales, there are slow, smooth variations in the wavelet image, Ws, whereas for finer scales, the variations are faster. At any fixed scale, the rough shape is captured by the size and the location of the peaks. For wavelet images based on a Gaussian mother wavelet, like the Mexican hat, the maxima of consecutive scales form continuous lines that always reach the finest scales (for a proof, see Mallat 1999, Proposition 6.1, p. 246). The lines of maximum amplitude (LoMA) contain somewhat surprisingly all the information necessary to reconstruct a signal (although for some wavelet families only approximate reconstruction is possible). Moreover, these reconstructions can be performed numerically in an efficient way using, e.g., Mallat's algorithm *à trous* (see Mallat 1999, pp. 253–259 for more details). For time-frequency analysis of speech signals with a similar approach, see Riley (1989).

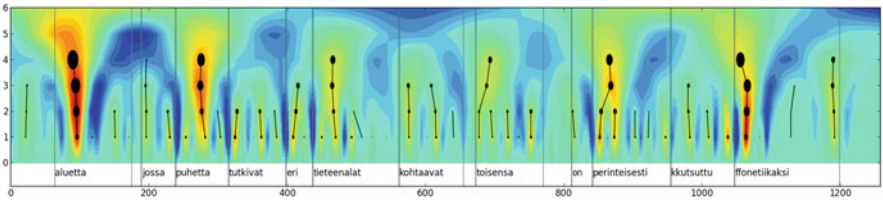


Fig. 12.3 CWT and LoMA-based analysis of an utterance. The word prominences are based on the strength of the accumulated LoMA values that reach the hierarchical level corresponding to words. See text for more details

12.3.4 Prominence Estimation Based on a Fixed Wavelet Scale

Since CWT decomposes the f_0 into oscillations of varying frequency, a natural candidate for the f_0-based word prominence would be the contribution of the scale corresponding to the average word occurrence rate. Let T be the total duration of the speech database and let W be the number of words in the data. Then the integer j_W closest to

$$\tilde{j}_W = \frac{W}{T \log a}$$

gives the desired scale.

Then $W_{j_W} s$ is normalized to have zero mean and unit variance. The prominence of the word w is defined as $P_W(w) = 0$ if there is no local maximum occurring during the word and $P_W(w) = W_{j_W} s(t^*)$ if there is one. If there are several local maxima, then the largest maximum is used to estimate the prominence value. Figure 12.2 shows a segment of speech with both the original f_0 contour (blue line) and the temporal CTW scale corresponding to words (red line). The prominence is estimated simply by the height of the peak of the CTW signal. The correct temporal scale can be estimated in a number of ways; here it is estimated by comparing the number of peaks of a given CWT level with the number of words in an utterance. See Suni et al. (2013) for a comparison of different methods of prominence estimation from the CWT at the level of word.

12.3.5 LoMA-Based Prominence Tagging

Inspired by the fact that LoMA can represent the essentials of a signal, we introduce a way to estimate the contribution of the f_0 across the scales to the word prominence. To this end, we define $P_L(w)$ (prominence by LoMA analysis of f_0). Let $W_j s$, $j = 1, \ldots, N$, be a coarse representation of f_0. We build a tree T connecting the local maxima $t^*_{j,k}$ of the level $j = 1, \ldots, N-1$ to the consecutive level (i.e., $j+1$). More precisely, t^* is a local maximum if $f_0(t^*) > f_0(t^* \pm \Delta t)$, where $\Delta t > 0$ is the size of the time step, here $\Delta t = 5$ ms.

Now the (local) maximum t^*_{j,k_j} is connected to the first $t^*_{j+1,k_{j+1}} > t^*_{j,k_j}$ if $W'_{j+1}s(t^*_{j,k_j}) > 0$, and to the last $t^*_{j+1,k_{j+1}} \leq t^*_{j,k_j}$ otherwise (i.e., if the time derivative of $W_{j+1}(s)$ is not positive at t^*_{j,k_j}). Two maxima that are connected are called child (scale j) and parent (scale $j+1$).

Next, we prune the tree to arrive at separate lines. Let

$$\tilde{W}_j s(t) = \frac{W_j s(t) - m}{s}$$

where m is the temporal mean of $W_j s$ and s is the standard deviation of $W_j s$. Let $A_{j,1}, \ldots, A_{j,N_j}$ be the cumulative sums defined iteratively as follows. First set

$$A_{1,k_1} = W_1 s(t^*_{1,k_1}).$$

Assume now that A_{j,k_j} have been defined and let us define $A_{j+1,k_{j+1}}$ for $k_{j+1} = 1, \ldots, N_{j+1}$. For $t^*_{j+1,k_{j+1}}$, let t^*_{j,k_j} be the child with largest A_{j,k_j} and define

$$A_{j+1,k_{j+1}} = W_{j+1} s(t^*_{j+1,k_{j+1}}) + A_{j,k_j}.$$

In other words, in a branching point of the tree, the branch with greatest accumulated sum is selected. Finally, the LoMA prominence $P_L(w)$ of a word w is defined as the largest A_{j,k_j} over all t^*_{j,k_j} occurring during the word.

Figure 12.3 shows a CWT analysis of f_0 with a superimposed LoMA analysis for a Finnish utterance "Aluetta, jossa puhetta tutkivat eri tieteenalat kohtaavat toisensa on perinteisesti kutsuttu fonetiikaksi" (The area, where different sciences working on speech, has traditionally been called phonetics). In the figure, the red side of the color spectrum depicts positive values and the blue side depicts the negative values. The inherent hierarchy is clearly visible in the figure.

12.4 Comparison of Automatic Labeling with Expert Judgements

As mentioned above, we operationalized word prominence as an abstract feature of speech prosody that can be used to simplify the process of predicting prosody from text. It can be further argued that the automatically estimated values are good per se and need not be related to how human listeners perceive them. Nevertheless, in order to assess the relative goodness of the different methods, the estimates need to be compared to some baseline. Human listeners are known to be fairly reliable in estimating prominences in a simple scale. Although several different scales can be used (see Arnold et al. 2012), many studies have found that the most reliable scale has four levels that correspond to prosody in roughly the following way: unaccented, moderately accented, accented, and emphatic (as in, e.g., narrow focus).

We compared the two automatic f_0-based prominence estimation schemes against a database of expert labels produced by ten participants (all students of phonetics). The database was constructed using iterative training, where syllable-aligned

Table 12.1 Confusion table for CWT-estimated tags (*horizontal axis*) vs. manually estimated prominences (*y-axis*). Seventeen outliers where both humans and the algorithm had used a value greater than three were removed

Level	0	1	2	3
0	2473	621	61	1
1	744	1873	430	13
2	52	587	779	72
3	1	4	7	0

prosodic differences between a simple HMM synthesis model (i.e., synthesis without prominence-related lexical information) and training data (as well as normalized prosodic features of the training data) were used for preliminary tags, which were then corrected by the expert labelers. Altogether 900 sentences spoken by a female speaker were used. The utterances were chosen from both read prose and nonfiction, as well as a set of phonetically rich sentences. The set contained over 7600 separate words, which were all labeled by three separate labelers. In summary, the raw agreement between labelers was as follows: 44.6 % of the time all three labelers agreed, 77.4 % of the time two out of three agreed—1.25 % of the disagreements were between two categories or more.

By analyzing the expert estimates we noticed that some of the labelers were, however, inconsistent (three used a finer scale) and two had shifted all their estimates. Thus, we had to disregard the data from three labelers. The data from the two labelers were shifted to the right scale so that their estimates corresponded to the estimates for the five other labelers. This reduced our data so that most of the words had only two estimates and only 2263 words had estimates from three labelers. With these data, 59.1 % times all three were in agreement, and 99 % of the time two out of three agreed. Single wavelet scale achieved slightly better correspondence with expert labels than the LoMA-based method, while both were clearly superior to our baseline method, raw f_0 maximum within a word. In order to make the estimates comparable with the four-level scale used in the manual task, the corresponding automatic estimates were scaled by multiplying the values by 0.7 and rounding them. The correspondences are summarized as a confusion matrix in Table 12.1. The correctly estimated categories are as follows: 75.6 % for level 0, 60.7 % for level 1, and 61 % for level 2. There were no estimations for category 3. In contrast, the best category for the raw f_0 peak value was 61.0 % for category 1. It is also clearly visible from the confusion matrix that the great majority of errors are within one category; majority of them are due to rounding. In practice, depending on prominence prediction methods, it is often suitable to treat prominence as a continuous scale, with the discrete four categories merely as a convenience for human annotators.

12.5 CWT and Prominence in Synthesis

The implementation of the CWT-based prosody in HMM TTS is described in Suni et al. (2013). In summary, the f_0 contour is modeled on five distinct scales ranging from phone (microprosody) to utterance level. All five scales are treated separately as different streams with their corresponding DTs.

12.5.1 CWT for Intonation Modeling in HMM-Based Speech Synthesis

In addition to providing useful features for prosody annotation, the CWT can provide attractive features for hierarchical modeling of prosody. This chapter reviews intonation modeling with CWT in HMM-based speech synthesis, described in more detail in Suni et al. (2013).

In standard HMM-based speech synthesis, f_0 is modeled jointly with voicing decision. The unit of modeling is typically a phone HMM with five states. For each state, predefined contextual questions concerning phones, syllables, words, and phrases are used to form a set of possible splits in a DT. The splitting decisions are made in a greedy fashion based on likelihood increase. Thus the hierarchical nature of intonation is only implicitly addressed by questions on different levels of hierarchy. With multiple levels, including voicing decision, modeled by a single set of trees, the rare or slow events can not be modeled robustly due to fragmentation of the training data by previous, more urgent splits for the short time scale of the model.

Previously in the HMM framework, decomposition of f_0 to its explicit hierarchical components during acoustic modeling has been investigated in Lei et al. (2010) and Zen and Braunschweiler (2009). These approaches rely on exposing the training data to a level-dependent subset of questions for separating the layers of the prosody hierarchy. The layers can then be modeled separately as individual streams (Lei et al. 2010), or jointly with adaptive training methods (Zen and Braunschweiler 2009).

With the CWT method, f_0 decomposition is performed in analysis phase, analogous to spectral feature extraction. Extracted scales can then be modeled either jointly or separately, or combined selectively prior to training.

In the best performing method in (Suni et al. 2013), the CWT with the Mexican hat mother wavelet is applied on the interpolated logF0 contours of the training utterances with ten scales, one octave apart. Then reconstruction formula (2) is applied to individual scales and finally each two adjacent scales are summed to form the five scales for training, corresponding to phone, syllable, word, phrase, and utterance levels.

During training, all five scales are treated separately as different streams with their corresponding DTs. In synthesis, parameters of each scale are generated independently and then summed to form the final contour, with unvoiced parts chopped based on the voicing decision of the standard multi space fundamental frequency probability distribution (MSD-F0) stream.

Table 12.2 A summary of the performance results of the syntheses. The means of the performance measures for each of the two data sets (female, male)

	CWT	Baseline
corr (F)	0.76	0.68
corr (M)	0.85	0.81
RMSE (F)	1.38	1.53
RMSE (M)	1.57	1.76

As each DT can now focus only on the respective scale with associated contextual features, the hypothesis was that the training data could be utilized more effectively than in the standard MSD method, and the resulting pitch contours would be closer to natural ones. The objective results supported this, with substantial improvements on both correlation and root mean square error (RMSE) on two Finnish voices, male and a female (Table 12.2).

12.5.2 Selective f_0 Enhancement and Emphasis

In HMM-based speech synthesis, DT clustering and maximum likelihood parameter generation cause the generated parameter trajectories to be oversmooth. From perceptual point of view, the resulting speech sounds muffled with monotonous prosody. To alleviate this problem, the original variance of the speech parameters can be taken into account via global variance parameter (GV). For spectral parameters, this method provides improved clarity, but for pitch, no perceptual improvement with restored variance has been observed (Toda and Tokuda 2007). We suspect that the reason for this is the aforementioned tendency of DT clustering to focus on microprosodic aspects of the fundamental frequency contour, and the nonselective nature of variance expansion; the subword variation is expanded too much, and higher levels not enough, causing an unnaturally jumpy prosody.

In contrast, when using CWT-based f_0 modeling, GV can be applied selectively for each scale, prior to reconstruction. Informally, we have observed that enhancing the variance of only the word and phrase scales provides the most natural results. Importantly, the selective enhancement opens up a method for production of emphasis, even if the model has not properly learned the highest prominence category, due to limited number of training samples. Figure 12.4 presents a comparison between baseline and CWT f_0 contours. The prominence category of the first and final word is kept constant (one), while the prominence of the middle word is varied from zero to two. The CWT version appears to have learned the prominence categories better, with clearly separated contours on and around the middle word.

In an attempt to produce emphasis, the variance of the baseline contour is expanded by a factor of 1.5, whereas in CWT method, the same expansion is placed on the word scale only, with the variances of both versions subsequently set identical. The results are shown in red in the figure. As can be seen, the baseline contour is affected adversely, with an unnaturally strong declination trend and enhancing the already

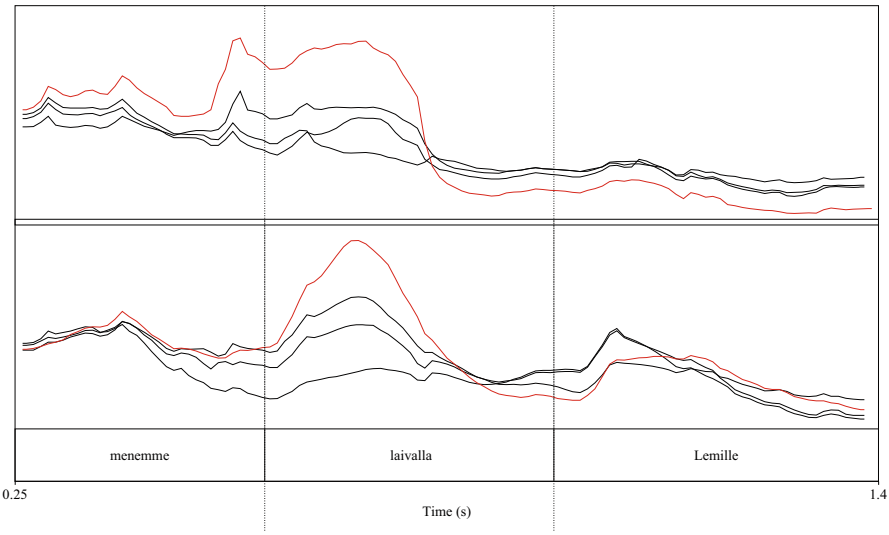


Fig. 12.4 An example of emphasis by postprocessing prominence for contrastive focus. See text for more details

noisy fluctuations, while the CWT contour appears quite natural with emphasized word standing out with acceptable transitions on adjacent words.

As described in Suni et al. (2013), the selective enhancement can also be used to model speaking styles. Emphasizing phrase level at the cost of syllable and word levels could provide a suitable style for reading prose or sped-up synthesis, and vice versa.

12.6 Conclusion

In this chapter we have described a novel way of both analyzing and synthesizing speech prosody concentrating on modeling the fundamental frequency. The system is based on hierarchical time-frequency analysis and reconstruction based on CWT. The CWT is further used for analyzing word prominences that are used to simplify the symbolic linguistic representation of an HMM-based synthesis system for producing natural sounding prosody in a TTS context. There are several advantages for using both the CWT and word prominence in TTS. The CWT can be used to model all variability in the f_0 contours; in addition to accentuation (prominence), the changes due to, e.g., utterance mode and phrasing are easily captured and can be straightforwardly modeled in a similar fashion with prominence. The CTW decomposition of the f_0 signal has a further advantage in that it allows for access to the relevant temporal scales in both a priori and post hoc fashions. One potential use is shown here, where an emphasis has been added to a word to render it narrowly focused by

adding variance to the word level wavelet signal. Such post hoc manipulations can be used for many other purposes as well. For instance, the compression of pitch in shouting can be selectively contested in order to maintain intelligibility. All in all, the system allows for capturing and representing all of the dynamics of speech in a natural fashion, and can in principle be trained in a fully unsupervised way.

Acknowledgements The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement nº 287678 (Simple4All) and the Academy of Finland (projects 135003 LASTU programme, 1128204, 128204, 125940). We would also like to thank Heini Kallio for collecting the prominence data.

References

Anumanchipalli, Gopala Krishna, Luis C. Oliveira, and Alan W. Black. 2011. A statistical phrase/accent model for intonation modeling. In *INTERSPEECH*, 1813–1816.

Arnold, Denis, Petra Wagner, and Bernd Möbius. 2012. Obtaining prominence judgments from naïve listeners—Influence of rating scales, linguistic levels and normalisation. In *Proceedings of INTERSPEECH* 2012.

Bailly, G., and B. Holm. 2005a. SFC: A trainable prosodic model. *Speech Communication* 46 (3–4): 348–364 (Cited by (since 1996) 15).

Bailly, Gérard, and Bleicke Holm. 2005b. SFC: A trainable prosodic model. *Speech Communication* 46 (3): 348–364.

Cole, Jennifer, Yoonsook Mo, and Mark Hasegawa-Johnson. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology* 1 (2): 425–452.

Daubechies, Ingrid, et al. 1992. *Ten lectures on wavelets*. vol. 61. Philadelphia: SIAM.

Eriksson, Anders, Gunilla C. Thunberg, and Hartmut Traunmüller. 2001. Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. In *Proceedings of European Conference on Speech Communication and Technology Aalborg*, vol. 1, 399–402. September 2001.

Fujisaki, H., and K. Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)* 5 (4): 233–241.

Fujisaki, Hiroya, and Hiroshi Sudo. 1971. A generative model for the prosody of connected speech in Japanese. *Annual Report of Engineering Research Institute* 30:75–80.

Grossman, A., and Jean Morlet. 1985. Decomposition of functions into wavelets of constant shape, and related transforms. *Mathematics and Physics: Lectures on Recent Results* 11:135–165.

Kochanski, Greg, and Chilin Shih. 2000. Stem-ML: Language-independent prosody description. In *INTERSPEECH*, 239–242.

Kochanski, Greg, and Chilin Shih. 2003. Prosody modeling with soft templates. *Speech Communication* 39 (3): 311–352.

Kronland-Martinet, Richard, Jean Morlet, and Alexander Grossmann. 1987. Analysis of sound patterns through wavelet transforms. *International Journal of Pattern Recognition and Artificial Intelligence* 1 (02): 273–302.

Kruschke, Hans, and Michael Lenz. 2003. Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis. In *INTERSPEECH*.

Lei, Ming, Yi-Jian Wu, Frank K. Soong, Zhen-Hua Ling, and Li-Rong Dai. 2010. A hierarchical F0 modeling method for HMM-based speech synthesis. In *INTERSPEECH*, 2170–2173.

Mallat, S. 1999. *A wavelet tour of signal processing*. Academic press.

Martti, Vainio, Antti Suni, Tuomo Raitio, Jani Nurminen, Juhani Järvikivi, and Paavo Alku. 2009. New method for delexicalization and its application to prosodic tagging for text-to-speech synthesis. In *INTERSPEECH*, 1703–1706. Brighton, UK, September 2009.

Öhman, Sven. 1967. *Word and sentence intonation: A quantitative model*. Stockholm: Speech Transmission Laboratory, Department of Speech Communication, Royal Institute of Technology.

Riley, Michael D. 1989. *Speech time-frequency representation*, vol. 63. Berlin: Springer.

Suni, Antti Santeri, Aalto Daniel, Raitio Tuomo, Alku Paavo, Vainio Martti, et al. 2013. Wavelets for intonation modeling in HMM speech synthesis. In *8th ISCA Workshop on Speech Synthesis, Proceedings*. Barcelona, 31 August–2 September 2013.

Taylor, Paul. 2000. Analysis and synthesis of intonation using the tilt model. *The Journal of the Acoustical Society of America* 107 (3): 1697–1714.

Toda, Tomoki, and Keiichi Tokuda. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems* 90 (5): 816–824.

Tokuda, Keiichi, Takao Kobayashi, and Satoshi Imai. 1995. Speech parameter generation from HMM using dynamic features. In *Acoustics, Speech, and Signal Processing, ICASSP-95, 1995 International Conference on*, vol. 1, 660–663. IEEE.

Tokuda, Keiichi, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. 2000. Speech parameter generation algorithms for HMM-based speech synthesis. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3, 1315–1318. IEEE.

Torrence, Christopher, and Gilbert P. Compo. 1998. A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society* 79 (1): 61–78.

Vainio, Martti, and Järvikivi Juhani. 2006. Tonal features, intensity, and word order in the perception of prominence. *Journal of Phonetics* 34 (3): 319–342.

Zen, Heiga, and Norbert Braunschweiler. 2009. Context-dependent additive log F_0 model for HMM-based speech synthesis. In *INTERSPEECH*, 2091–2094.

Chapter 13
Exploiting Alternatives for Text-To-Speech Synthesis: From Machine to Human

Nicolas Obin, Christophe Veaux and Pierre Lanchantin

Abstract The absence of alternatives/variants is a dramatical limitation of text-to-speech (TTS) synthesis compared to the variety of human speech. This chapter introduces the use of speech alternatives/variants in order to improve TTS synthesis systems. Speech alternatives denote the variety of possibilities that a speaker has to pronounce a sentence—depending on linguistic constraints, specific strategies of the speaker, speaking style, and pragmatic constraints. During the training, symbolic and acoustic characteristics of a unit-selection speech synthesis system are statistically modelled with context-dependent parametric models (Gaussian mixture models (GMMs)/hidden Markov models (HMMs)). During the synthesis, symbolic and acoustic alternatives are exploited using a Generalized Viterbi Algorithm (GVA) to determine the sequence of speech units used for the synthesis. Objective and subjective evaluations support evidence that the use of speech alternatives significantly improves speech synthesis over conventional speech synthesis systems. Moreover, speech alternatives can also be used to vary the speech synthesis for a given text. The proposed method can easily be extended to HMM-based speech synthesis.

13.1 Introduction

Today, speech synthesis systems (unit selection (Hunt and Black 1996), HMM-based (Zen et al. 2009)) are able to produce natural synthetic speech from text. Over the last decade, research has mainly focused on the modelling of speech

N. Obin (✉)
IRCAM, UMR STMS IRCAM-CNRS-UPMC,
Paris, France
e-mail: Nicolas.Obin@ircam.fr

C. Veaux
Centre for Speech Technology Research, The University of Edinburgh, Edinburgh, UK
e-mail: cveaux@inf.ed.ac.uk

P. Lanchantin
Department of Engineering, Cambridge University, Cambridge, UK
e-mail: pkl27@cam.ac.uk

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,* Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_13

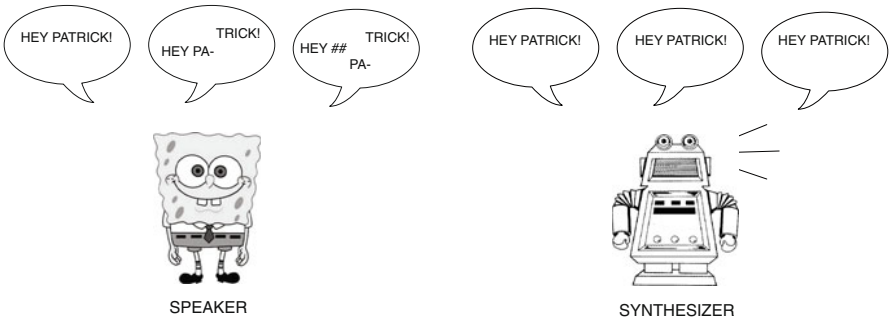


Fig. 13.1 Illustration of speech alternatives: human vs. machine

prosody—"the music of speech" (accent/phrasing, intonation/rhythm)—for text-to-speech (TTS) synthesis. Among them, GMMs/HMMs (Gaussian mixture models and hidden Markov models) are today the most popular methods used to model speech prosody. In particular, the modelling of speech prosody has gradually and durably moved from short-time representations ("frame-by-frame": Yoshimura et al. 1999; Zen et al. 2004; Tokuda et al. 2003; Toda and Tokuda 2007; Yan et al. 2009) to the use of large-time representations (Gao et al. 2008; Latorre and Akamine 2008; Qian et al. 2009; Obin et al. 2011b)). Also, recent researches tend to introduce deep architecture systems to model more efficiently the complexity of speech (deep neural networks (Zen et al. 2013)). However, current speech synthesis systems still suffer from a number of limitations, which consequence into the fact that the synthetic speech does not totally sound as "human". In particular, the absence of alternatives/variants in the synthetic speech is a dramatical limitation compared to the variety of human speech (see Fig. 13.1 for illustration): for a given text, the speech synthesis system will always produce exactly the same synthetic speech.

A human speaker can use a variety of alternatives/variants to pronounce a text. This variety may induce variations in the symbolic (prosodic event: accent, phrasing) and acoustic (prosody: prosodic contour; segmental: articulation, co-articulation) speech characteristics. These alternatives depend on linguistic constraints, specific strategies of the speaker, speaking style, and pragmatic constraints. Current speech synthesis systems do not exploit this variety during statistical modelling or synthesis. During the training, the symbolic and acoustic speech characteristics are usually estimated with a single normal distribution which is assumed to correspond with a single strategy of the speaker. During the synthesis, the sequence of symbolic and acoustic speech characteristics are entirely determined by the sequence of linguistic characteristics associated with the sentence—the *most-likely* sequence.

In real-world speech synthesis applications (e.g. announcement, storytelling, or interactive speech systems), expressive speech is required (Obin et al. 2011a; Obin 2011). The use of speech alternatives in speech synthesis may substantially improve speech synthesis (Bulyko and Ostendorf 2001), and fill the gap of the machine to the human. First, alternatives can be used to provide a variety of speech candidates

that may be exploited to vary the speech synthesized for a given sentence. Second, alternatives can also be advantageously used as a relaxed constraint for the determination of the sequence of speech units to improve the quality of the synthesized speech. For instance, the use of a symbolic alternative (e.g. insertion/deletion of a pause) may conduct to a significantly improved sequence of speech units.

This chapter addresses the use of speech alternatives to improve the quality and the variety of speech synthesis. The proposed speech synthesis system (IRCAMTTS) is based on unit selection, and uses various context-dependent parametric models to represent the symbolic/acoustic characteristics of speech prosody (GMMs/HMMs). During the synthesis, symbolic and acoustic alternatives are exploited using a generalized Viterbi algorithm (GVA) (Hashimoto 1987). First, a GVA is used to determine a set of symbolic candidates, corresponding to the $K_{\text{symb.}}$ sequences of symbolic characteristics, in order to enrich the further selection of speech units. For each symbolic candidate, a GVA is then used to determine the $K_{\text{acou.}}$ sequences of speech units under the joint constraint of segmental and speech prosody characteristics. Finally, the optimal sequence of speech units is determined so as to maximize the cumulative symbolic/acoustic likelihood. Alternatively, the introduction of alternatives allows to vary the speech synthesis by selecting one of the K most likely speech sequences instead of the most likely one. The proposed method can easily be extended to HMM-based speech synthesis.

The speech synthesis system used for the study is presented in Sect. 13.2. The use of speech alternatives during the synthesis, and the GVA are introduced in Sect. 13.3. The proposed method is compared to various configurations of the speech synthesis system (modelling of speech prosody, use of speech alternatives), and validated with objective and subjective experiments in Sect. 13.4.

13.2 Speech Synthesis System

Unit-selection speech synthesis is based on the optimal selection of a sequence of speech units that corresponds to the sequence of linguistics characteristics derived from the text to synthesize. The optimal sequence of speech units is generally determined so as to minimize an objective function usually defined in terms of concatenation and target acoustic costs. Additional information (e.g. prosodic events—ToBI labels) can also be derived from the text to enrich the description used for unit selection.

The optimal sequence of speech units $\overline{\mathbf{u}}$ can be determined by jointly maximizing the symbolic/acoustic likelihood of the sequence of speech units $\mathbf{u} = [u_1, \ldots, u_N]$ conditionally to the sequence of linguistic characteristics $\mathbf{c} = [c_1, \ldots, c_N]$:

$$\overline{\mathbf{u}} = \underset{\mathbf{u}}{\text{argmax}}\ \text{p}(\mathbf{O}(\mathbf{u})|\mathbf{c}) \tag{13.1}$$

where $\mathbf{O}(\mathbf{u}) = [\mathbf{O}_{\text{symb.}}(\mathbf{u}),\ \mathbf{O}_{\text{acou.}}(\mathbf{u})]$ denotes the symbolic and acoustic characteristics associated with the sequence of speech units $\mathbf{u}$.

A suboptimal solution to this equation is usually obtained by factorizing the symbolic/acoustic characteristics:

$$\overline{\mathbf{u}}_{\text{symb.}} = \underset{\mathbf{u}_{\text{symb.}}}{\text{argmax}}\ \text{p}(\mathbf{O}_{\text{symb.}}(\mathbf{u}_{\text{symb.}})|\mathbf{c}) \tag{13.2}$$

$$\overline{\mathbf{u}}_{\text{acou.}} = \underset{\mathbf{u}}{\text{argmax}}\ \text{p}(\mathbf{O}_{\text{acou.}}(\mathbf{u}_{\text{acou.}})|\mathbf{c}, \overline{\mathbf{u}}_{\text{symb.}}) \tag{13.3}$$

where $\mathbf{u}_{\text{symb.}}$ is the symbolic sequence of speech units (typically, a sequence of prosodic events, e.g. accent and phrasing), and $\mathbf{u}_{\text{acou.}}$ is the acoustic sequence of speech units (i.e. a sequence of speech units for unit-selection and a sequence of speech parameters for HMM-based speech synthesis). This acoustic sequence of speech units represents the short- (source/filter) and long-term (prosody: F0, duration) variations of speech over various units (e.g. phone, syllable, and phrase).

In other words, the symbolic sequence of speech units $\overline{\mathbf{u}}_{\text{symb.}}$ is first determined, and then used for the selection of acoustic speech units $\overline{\mathbf{u}}_{\text{acou.}}$. This conventional approach suffers from the following limitations:

1. Symbolic and acoustic modelling are processed separately during training and synthesis, which remain suboptimal and may degrade the quality of the synthesized speech.
2. A single sequence of speech units is determined during synthesis, while the use of alternatives enlarges the number of speech candidates available, and then improves the quality of the synthesized speech.

To overcome these limitations, the ideal solution is: the joint symbolic/acoustic modelling in order to determine the sequence of speech units that is globally optimal (Eq. 13.1); and the exploitation of speech alternatives in order to enrich the search for the optimal sequence of speech units. The present study only addresses the use of symbolic/acoustic alternatives for speech synthesis. In the present study, symbolic alternatives are used to determine a set of symbolic candidates $\overline{\mathbf{u}}_{\text{symb.}}$ so as to enrich the further selection of speech units (Eq. 13.2). For each symbolic candidate, the sequence of acoustic speech units $\overline{\mathbf{u}}_{\text{acou.}}$ is determined based on a relaxed-constraint search using acoustic alternatives (Eq. 13.3). Finally, the optimal sequence of speech units $\overline{\mathbf{u}}$ is determined so as to maximize the cumulative likelihood of the symbolic/acoustic sequences.

The use of symbolic/acoustic alternatives requires adequate statistical models that explicitly describe alternatives, and a dynamic selection algorithm that can manage these alternatives during speech synthesis. Symbolic and acoustic models used for this study are briefly introduced in Sects. 13.2.1 and 13.2.2. Then, the dynamic selection algorithm used for unit selection is described in Sect. 13.3.

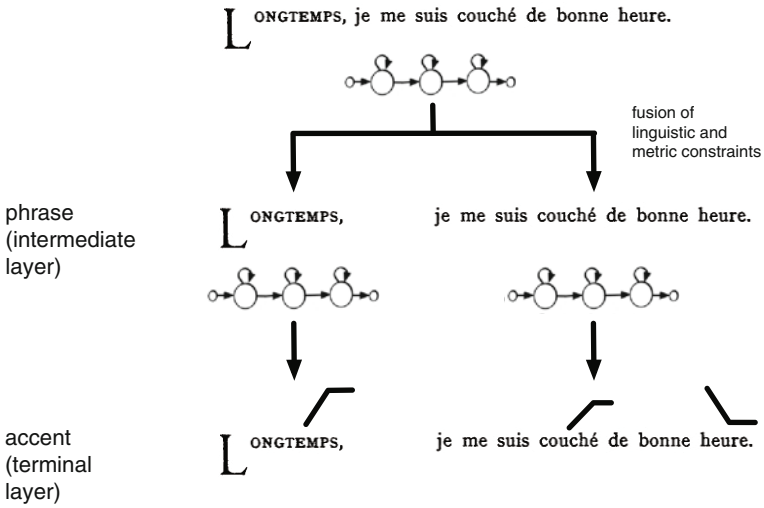


Fig. 13.2 Illustration of the HHMM symbolic modelling of speech prosody for the sentence: "*Longtemps, je me suis couché de bonne heure*" ("*For a long time I used to go to bed early*"). The *intermediate layer* illustrates the segmentation of a text into phrases. The *terminal layer* illustrates the assignment of accents

13.2.1 Symbolic Modelling

The prosodic events (accent and phrasing) are modelled by a statistical model based on HMMs (Black and Taylor 1994; Atterer and Klein 2002; Ingulfen et al. 2005; Obin et al. 2010a, 2010b; Parlikar and Black 2012; Parlikar and Black 2013). A hierarchical HMM (HHMM) is used to assign the prosodic structure of a text: the root layer represents the text, each intermediate layer a phrase (here, intermediate phrase and phrase), and the final layer the sequence of accents. For each intermediate layer, a segmental HMM and information fusion are used to combine the linguistic and metric constraints (length of a phrase) for the segmentation of a text into phrases (Ostendorf and Veilleux 1994; Schmid and Atterer 2004; Bell et al. 2006; Obin et al. 2011c). An illustration of the HHMM for the symbolic modelling of speech prosody is presented in Fig. 13.2.

13.2.2 Acoustic Modelling

The acoustic (short- and long-term) models are based on context-dependent GMMs (cf. Veaux et al. 2010; Veaux and Rodet 2011, for a detailed description). Three different observation units (phone, syllable, and phrase) are considered, and separate GMMs are trained for each of these units. The model associated with the phone unit is merely a reformulation of the target and concatenation costs traditionally used in

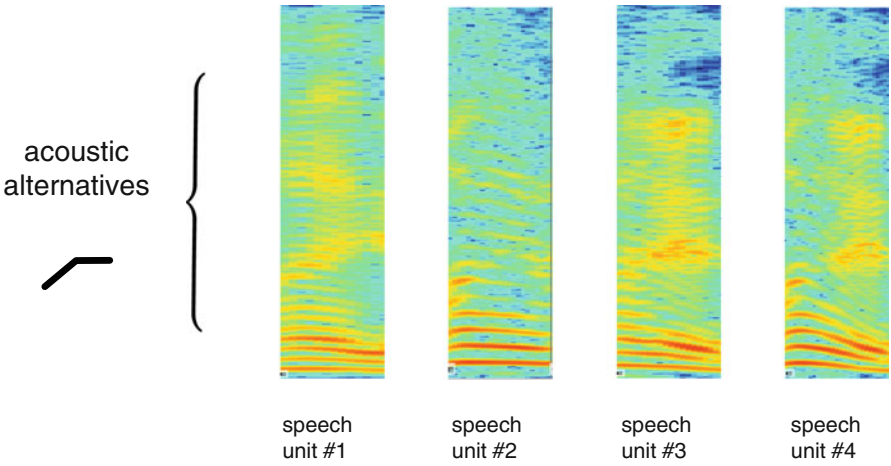


Fig. 13.3 Illustration of acoustic alternatives for a given symbolic unit

unit-selection speech synthesis (Hunt and Black 1996). The other models are used to represent the local variation of prosodic contours (*F0* and durations) over the syllables and the major prosodic phrases, respectively. The use of GMMs allows to capture prosodic alternatives associated with each of the considered units (Fig. 13.3).

13.3 Exploiting Alternatives

The main idea of the contribution is to exploit the symbolic/acoustic alternatives observed in human speech. Fig. 13.4 illustrates the integration of symbolic/acoustic alternatives for speech synthesis. The remainder of this section presents the details of the generalized Viterbi search to exploit symbolic/acoustic alternatives for TTS synthesis.

In a conventional synthesizer, the search for the optimal sequence of speech units (Eq. 13.1) is decomposed in two separate optimisation problems (Eqs. 13.2 and 13.3). These two equations are generally solved using the Viterbi algorithm. This algorithm defines a lattice whose states at each time t are the N candidate units. At each time t, the Viterbi algorithm considers N lists of competing paths, each list being associated to one of the N states. Then, for each list, only one survivor path is selected for further extension. Therefore the Viterbi algorithm can be described as a N-list 1-survivor (N,1) algorithm. The GVA (Hashimoto 1987) consists in a twofold relaxation of the path selection.

- First, more than one survivor path can be retained for each list.
- Second, a list of competing paths can encompass more than one state.

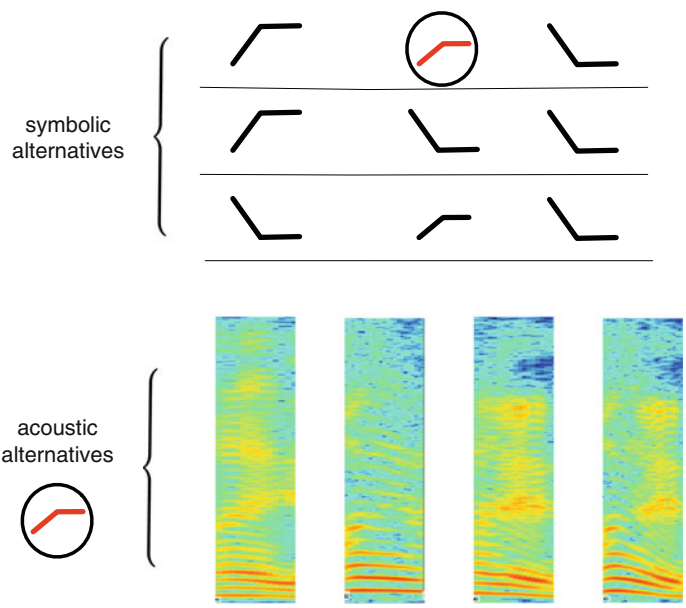


Fig. 13.4 Illustration of symbolic/acoustic alternatives for text-to-speech synthesis. The *top* of the figure presents three symbolic alternative sequences to a given input text. The *bottom* of the figure presents four acoustic alternatives to the symbolic event *circled* on *top*. Fundamentally, each text has symbolic alternative sequence, and each symbolic alternative sequence has acoustic alternative sequences

An illustration of this approach is given in Fig. 13.5, which shows that the GVA can retain survivor paths that would otherwise be merged by the classical Viterbi algorithm. Thus, the GVA can keep track of several symbolic/prosodic alternatives until the final decision is made.

In this study, the GVA is first used to determine a set of symbolic candidates corresponding to the $K_{\text{symb.}}$ most-likely sequences of symbolic characteristics, in order to enrich the further selection of speech units. For each symbolic candidate, a GVA is then used to determine the $K_{\text{acou.}}$ most-likely sequences of speech units under the joint constraint of segmental characteristics (phone model) and prosody (syllable and phrase models). Finally, the optimal sequence of speech units is determined so as to maximize the cumulative symbolic/acoustic likelihood.

13.4 Experiments

Objective and subjective experiments were conducted to address the use of speech alternatives in speech synthesis, with comparison to a BASELINE (no explicit modelling of speech prosody, no use of speech alternatives) and CONVENTIONAL (explicit

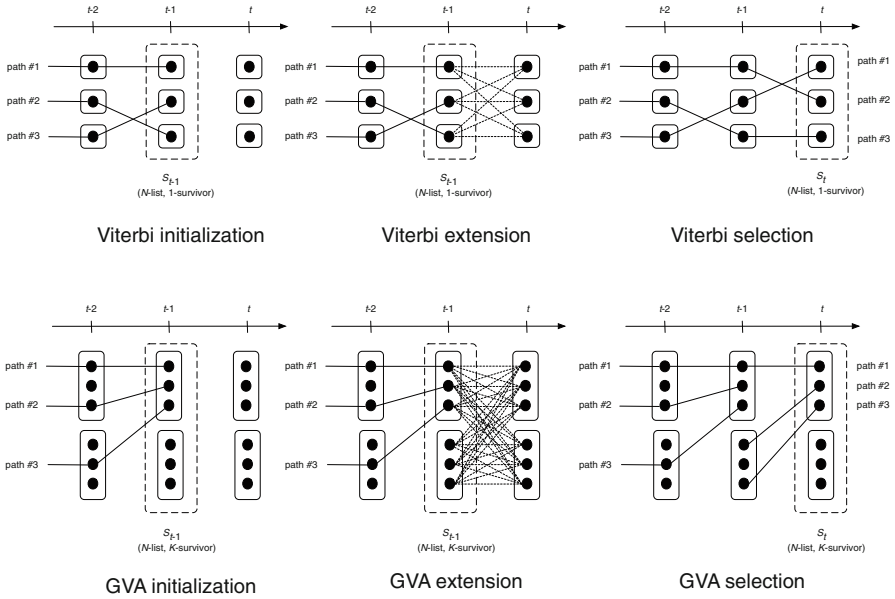


Fig. 13.5 Illustration of VITERBI SEARCH and GENERALIZED VITERBI SEARCH. The *boxes* represent the list of states among which the best S path is selected. For the VITERBI SEARCH, only one path is retained at all time, and only one survivor is retained during selection. For the GENERALIZED VITERBI SEARCH, K paths are retained at all time, and K survivors are retained during selection (alternative candidates, here $K = 3$). At all time, the GENERALIZED VITERBI SEARCH has a larger memory than the VITERBI SEARCH

Table 13.1 Description of TTS systems used for the evaluation. Parentheses denote the optional use of symbolic alternatives in the TTS system

	Symbolic	Acoustic	
	Alternatives	Prosody	Alternatives
BASELINE	(✓)	–	–
CONVENTIONAL	(✓)	Syllable/phrase	–
PROPOSED	(✓)	Syllable/phrase	✓

modelling of speech prosody, no use of speech alternatives) speech synthesis systems (Table 13.1). In addition, symbolic alternatives have been optionally used for each compared method to assess the relevancy of symbolic and acoustic alternatives separately.

13.4.1 Speech Material

The speech material used for the experiment is a 5-h French storytelling database interpreted by a professional actor, which was designed for expressive speech synthesis. The speech database comes with the following linguistic processing: orthographical transcription; surface syntactic parsing (POS and word class); manual speech segmentation into phonemes and syllables, and automatic labelling/segmentation of prosodic events/units (cf. Obin et al. 2010b for more details).

13.4.2 Objective Experiment

An objective experiment has been conducted to assess the relative contribution of speech prosody and symbolic/acoustic alternatives to the overall quality of the TTS system. In particular, a specific focus will be made on the use of symbolic/acoustic alternatives.

13.4.2.1 Procedure

The objective experiment has been conducted with 173 sentences of the fairy tale "*Le Petit Poucet*" ("*Tom Thumb*").
For this purpose, a *cumulative* log-likelihood has been defined as a weighted integration of the *partial* log-likelihoods (symbolic, acoustic). First, each partial log-likelihood is averaged over the utterance to be synthesized so as to normalize the variable number of observations used for the computation (e.g. phonemes, syllable, and prosodic phrase). Then, log-likehoods are normalized to ensure comparable contribution of each partial log-likelihood during the speech synthesis. Finally, the cumulative log-likelihood of a synthesized speech utterance is defined as follows:

$$LL = w_{\text{symbolic}} LL_{\text{symbolic}} + w_{\text{acoustic}} LL_{\text{acoustic}} \tag{13.4}$$

where LL_{symbolic} and LL_{acoustic} denote the partial log-likelihood associated with the sequence of symbolic and acoustic characteristics; and w_{symbolic} and w_{acoustic}, corresponding weights.
Finally, the optimal sequence of speech units is determined so as to maximize the cumulative log-likelihood of the symbolic/acoustic characteristics. In this study, weights were heuristically chosen as $w_{\text{symbolic}} = 1$, $w_{\text{phone}} = 1$, $w_{\text{syllable}} = 5$, and $w_{\text{phrase}} = 1$; 10 alternatives have been considered for the symbolic characteristics, and 50 alternatives for the selection of speech units.

13.4.2.2 Discussion

Cumulative likelihood obtained for the compared methods is presented in Fig. 13.6, with and without the use of symbolic alternatives. The PROPOSED method (modelling

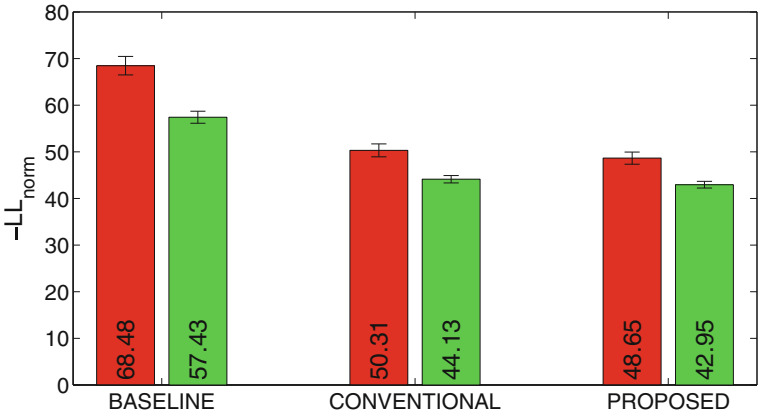


Fig. 13.6 Cumulative negative log-likelihood (mean and 95 % confidence interval) obtained for the compared TTS, without (*left*) and with (*right*) use of symbolic alternatives

of prosody, use of acoustic alternatives) moderately but significantly outperforms the CONVENTIONAL method (modelling of prosody, no use of acoustic alternatives); and dramatically outperforms the BASELINE method. In addition, the use of symbolic alternatives conducts to a significant improvement regardless of the method considered. Finally, the optimal synthesis is obtained for the combination of symbolic/acoustic alternatives with the modelling of speech prosody.

For further investigation, partial likelihoods obtained for the compared methods are presented in Fig. 13.7, with and without the use of symbolic alternatives. Not surprisingly, the modelling of speech prosody (syllable/phrase) successfully constraints the selection of speech units with adequate prosody, while this improvement comes with a slight degradation of the segmental characteristics (phone). The use of acoustic alternatives conducts to an improved speech prosody (significant over the syllable, not significant over the phrase) that comes with a slight degradation of the segmental characteristics (nonsignificant). This suggests that the phrase modelling (as described by Veaux and Rodet 2011) has partially failed to capture relevant variations, and that this model remains to be improved. Finally, symbolic alternatives are advantageously used to improve the prosody of the selected speech units, without a significant change in the segmental characteristics.

13.4.3 Subjective Experiment

A subjective experiment has been conducted to compare the quality of the BASELINE, CONVENTIONAL, and PROPOSED speech synthesis systems.

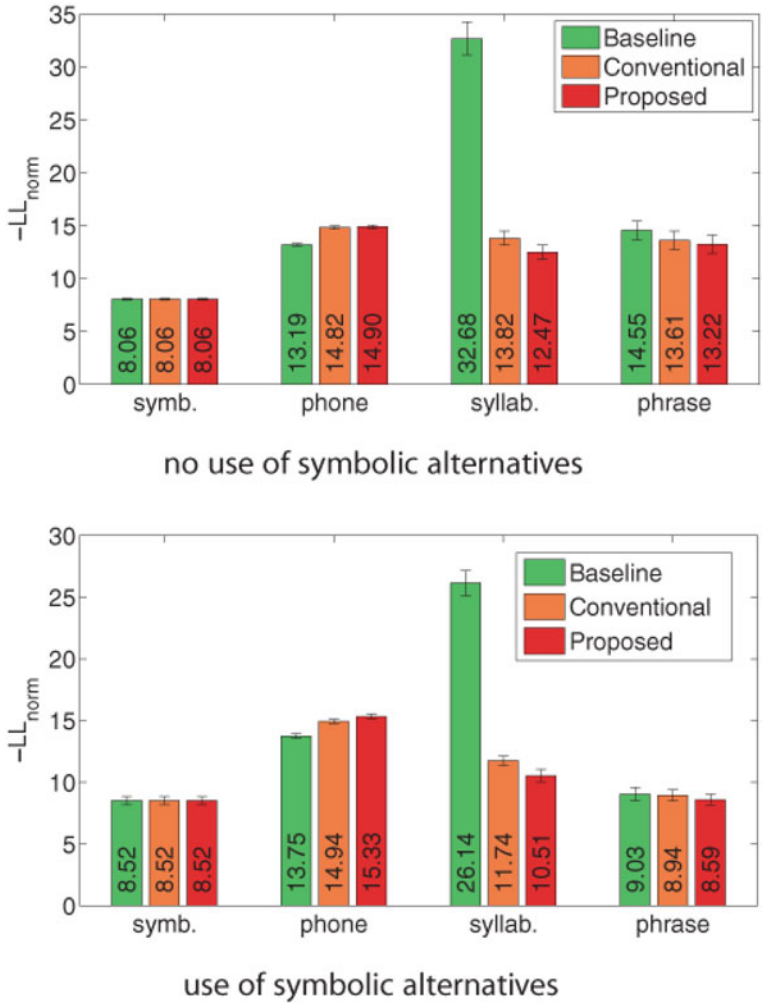


Fig. 13.7 Partial negative log-likelihoods (mean and 95 % confidence intervals) for the compared methods, with and without use of symbolic alternatives

13.4.3.1 Procedure

For this purpose, 11 sentences have been randomly selected from the fairy tale, and used to synthesize speech utterances with respect to the considered systems. Fifteen native French speakers have participated in the experiment. The experiment has been conducted according to a *crowdsourcing* technique using social networks. Pairs of synthesized speech utterances were randomly presented to the participants who were asked to attribute a preference score according to the *naturalness* of the

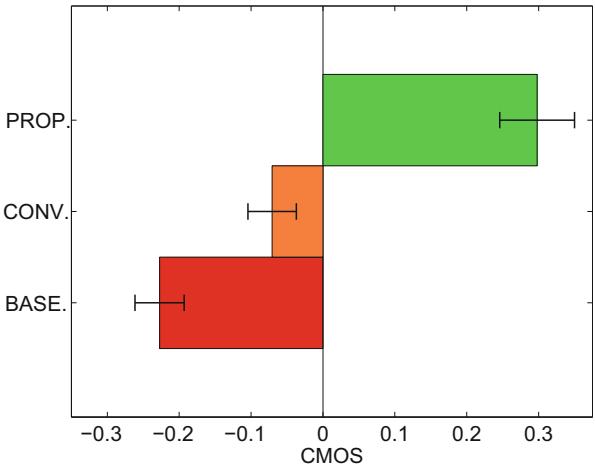


Fig. 13.8 CMOS (mean and 95 % confidence interval) obtained for the compared methods

speech utterances on the comparison mean opinion score (CMOS) scale. Participants were encouraged to use headphones.

13.4.3.2 Discussion

Figure 13.8 presents the CMOS obtained for the compared methods. The PROPOSED method is substantially preferred to other methods, which indicates that the use of symbolic/acoustic alternatives conducts to a qualitative improvement of the speech synthesized over all other systems. Then, CONVENTIONAL method is fairly preferred to the BASELINE method, which confirms that the integration of speech prosody also improves the quality of speech synthesis over the BASELINE system (cf. observation partially reported in Veaux and Rodet 2011).

13.5 Conclusion

In this chapter, the use of speech alternatives/variants in the unit-selection speech synthesis has been introduced. Objective and subjective experiments support the evidence that the use of speech alternatives qualitatively improves speech synthesis over conventional speech synthesis systems. The proposed method can easily be extended to HMM-based speech synthesis. In further studies, the use of speech alternatives will be integrated into a joint modelling of symbolic/acoustic characteristics so as to improve the consistency of the selected symbolic/acoustic sequence of speech units. Moreover, speech alternatives will further be used to vary the speech synthesis for a given text.

References

Atterer, M., and E. Klein. 2002. Integrating linguistic and performance-based constraints for assigning phrase breaks. In *International Conference on Computational Linguistics*, Taipei, Taiwan, 995–998.

Bell, P., T. Burrows, and P. Taylor. 2006. Adaptation of prosodic phrasing models. In *Speech Prosody*, Dresden, Germany.

Black, A., and P. Taylor. 1994. Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. In *International Conference on Spoken Language Processing*, Yokohama, Japan, 715–718.

Bulyko, I., and M. Ostendorf. 2001. Joint prosody prediction and unit selection for concatenative speech synthesis. In *International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, 781–784.

Gao, B., Y. Qian, Z. Wu, and F. Soong. 2008. Duration refinement by jointly optimizing state and longer unit likelihood. In *Interspeech*, Brisbane, Australia, 2266–2269.

Hashimoto, T. 1987. A list-type reduced-constraint generalization of the Viterbi algorithm. *IEEE Transactions on Information Theory* 33 (6): 866–876.

Hunt, A., and A. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Audio, Speech, and Signal Processing*, 373–376.

Ingulfen, T., T. Burrows, and S. Buchholz. 2005. Influence of syntax on prosodic boundary prediction. In *Interspeech*, Lisboa, Portugal, 1817–1820.

Latorre, J., and M. Akamine. 2008. Multilevel parametric-base F0 model for speech synthesis. In *Interspeech*, Brisbane, Australia, 2274–2277.

Obin, N. 2011. MeLos: Analysis and modelling of speech prosody and speaking style. PhD Thesis, Ircam - UPMC.

Obin, N., P. Lanchantin, A. Lacheret, and X. Rodet. 2010a. Towards improved HMM-based speech synthesis using high-level syntactical features. In *Speech Prosody*, Chicago, USA

Obin, N., A. Lacheret, and X. Rodet. 2010b. HMM-based prosodic structure model using rich linguistic context. In *Interspeech*, Makuhari, Japan, 1133–1136.

Obin, N., P. Lanchantin, A. Lacheret, and X. Rodet. 2011a. Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation. In *Interspeech*, Florence, Italy, 2785–2788.

Obin, N., A. Lacheret, and X. Rodet. 2011b. Stylization and trajectory modelling of short and long term speech prosody variations. In *Interspeech*, Florence, Italy, 2029–2032.

Obin, N., P. Lanchantin, A. Lacheret, and X. Rodet. 2011c. Reformulating prosodic break model into segmental HMMs and information fusion. In *Interspeech*, Florence, Italy, 1829–1832.

Ostendorf, M., and N. Veilleux. 1994. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Journal of Computational Linguistics* 20 (1): 27–54.

Parlikar, A., and A. W. Black. 2012. Modeling pause-duration for style-specific speech synthesis. In *Interspeech*, Portland, Oregon, USA, 446–449.

Parlikar, A., and A. W. Black. 2013. Minimum error rate training for phrasing in speech synthesis. In *Speech Synthesis Workshop (SSW)*, Barcelona, Spain, 13–17.

Qian, Y., Z. Wu, and F. K. Soong. 2009. Improved prosody generation by maximizing joint likelihood of state and longer units. In *International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 3781–3784.

Schmid, H., and M. Atterer. 2004. New statistical methods for phrase break prediction. In *International Conference on Computational Linguistics*, Geneva, Switzerland, 659–665.

Toda, T., and K. Tokuda. 2007. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Transactions on Information and Systems* 90 (5): 816–824.

Tokuda, K., H. Zen, and T. Kitamura. 2003. Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features. In *European Conference on Speech Communication and Technology*, Geneva, Switzerland, 865–868.

Veaux, C., and X. Rodet. 2011. Prosodic control of unit-selection speech synthesis: A probabilistic approach. In *International Conference on Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, 5360–5363.

Veaux, C., P. Lanchantin, and X. Rodet. 2010. Joint prosodic and segmental unit selection for expressive speech synthesis. In *Speech Synthesis Workshop (SSW7)*, Kyoto, Japan, 323–327.

Yan, Z.-J., Y. Qian, and F. K. Soong. 2009. Rich context modeling for high quality HMM-based TTS. In *Interspeech*, Brighton, UK, 4025–4028.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *European Conference on Speech Communication and Technology*, Budapest, Hungary, 2347–2350.

Zen, H., K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. 2004. Hidden semi-Markov model based speech synthesis. In *International Conference on Spoken Language Processing*, Jeju Island, Korea, 1397–1400.

Zen, H., K. Tokuda, and A. Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51 (11): 1039–1064.

Zen, A., A. Senior, and M. Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, 7962–7966.

Chapter 14
Prosody Control and Variation Enhancement Techniques for HMM-Based Expressive Speech Synthesis

Takao Kobayashi

Abstract Natural speech has diverse forms of expressiveness including emotions, speaking styles, and voice characteristics. Moreover, the expressivity changes depending on many factors at the phrase level, such as the speaker's temporal emotional state, focus, feelings, and intention. Thus taking into account such variations in modeling of speech synthesis units is crucial to generating natural-sounding expressive speech. In this context, two approaches to HMM-based expressive speech synthesis are described: a technique for intuitively controlling style expressivity appearing in synthetic speech by incorporating subjective intensity scores in the model training and a technique for enhancing prosodic variations of synthetic speech using a newly defined phrase-level context for HMM-based speech synthesis and its unsupervised annotation for training data consisting of expressive speech.

14.1 Introduction

Synthesizing natural-sounding speech with diverse forms of expressiveness including emotions, speaking styles, voice characteristics, focuses, and emphases, is a key to achieving more natural human–computer interactions. In this regard, the promising results of HMM-based speech synthesis (Erickson 2005; Nose and Kobayashi 2011a; Schröder 2009; Zen et al. 2009) have recently led to a number of attempts at applying this approach to expressive speech synthesis using speech corpora recorded under realistic conditions (Eyben et al. 2012; Koriyama et al. 2011; Maeno et al. 2011; Obin et al. 2011a, b).

When the target expressiveness consistently appears in every utterance in a corpus, it can be modeled properly and synthetic speech can be generated with expressiveness similar to that of the corpus (Yamagishi et al. 2003). Moreover, a style interpolation

T. Kobayashi (✉)
Department of Information Processing, Tokyo Institute of Technology,
Yokohama 226-8502, Japan
e-mail: takao.kobayashi@ip.titech.ac.jp

K. Hirose, J. Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis,*Prosody, Phonology and Phonetics,
DOI 10.1007/978-3-662-45258-5_14

technique (Tachibana et al. 2005) or a multiple regression HMM-based speech synthesis approach (Miyanaga et al. 2004; Nose et al. 2006), called the style control technique, can be used to strengthen or weaken the intensity of target expressivity of the synthetic speech. However, the prosodic variation of real expressive speech is generally much larger than that of simulated speech and is not consistent. In other words, expressivity changes depending on many factors, such as the speaker's temporal emotional state, focus, feelings, and intention at the phrase level. In addition, its intensity is not constant within a sentence (Cowie and Cornelius 2003; Doukhan et al. 2011). Thus, incorporating such variations into the modeling of the HMM-based speech synthesis units will be crucial to generating natural-sounding expressive speech.

This paper addresses two issues related to modeling and synthesizing expressive speech in the HMM-based framework. In the following, the expressions of emotions, speaking styles, prominences, etc., which may appear singly or simultaneously, will be referred to simply as *styles* (Miyanaga et al. 2004; Yamagishi et al. 2003).

The first issue is that the original style control technique did not take account of the style intensities of the respective utterances during the model training phase. The style intensity is a subjective score of a certain style expressivity given by listeners. In Miyanaga et al. (2004) and Nose et al. (2006), the style intensity, represented by a style vector, was assumed to be a fixed value regardless of the variations in style intensity appearing in respective training samples. This assumption may result in synthetic speech with consistently lower expressivity than what the user expects, if the average intensity of the target style in the training data is much lower than what the user expects. This problem can be alleviated by using subjective style intensity scores of respective utterances and taking them into account in the model training of the style control technique. Note that this technique is similar to other adaptive training approaches (Anastasakos et al. 1996; Gales 2000; Yu et al. 2001).

The second issue is an inevitable impediment to natural-sounding expressive speech synthesis. In the HMM-based speech synthesis approach, it is possible to reproduce locally appearing styles or prosodic variations only if the corpus for model training has been properly annotated and appropriate context labels have been given (Koriyama et al. 2011; Maeno et al. 2011; Yu et al. 2010). While manually annotating a corpus might work for this purpose, it is time-consuming and tends to be expensive and impractical on a large corpus. In addition, even if the cost is acceptable, another difficulty arises in that consistent annotation of styles, such as emotional expressions, is not always possible. Alternatively, unsupervised clustering of styles for expressive speech synthesis has been examined as a way of avoiding manual annotation and categorization of styles (Eyben et al. 2012; Székely et al. 2011). However, since there is not always an explicit expressiveness in the resultant clusters, users may have difficulty choosing an appropriate cluster to output the desired expressive speech in the speech synthesis phase.

In this context, there is an alternative technique for enhancing the prosodic variations of synthetic expressive speech without requiring the manual annotation of style information in the model training. An additional context, called the phrase-level F0

context, can be introduced, and it is defined by the average difference in prosodic features between the original and synthetic speech of the training sentences. The advantage of using this newly defined context is that proper annotation can be done fully automatically without any heuristics. Moreover, the obtained context has an intuitive prosodic meaning of higher or lower pitch at a certain accent phrase.

14.2 Prosody Control Based on Style Control Technique

14.2.1 Style Control Using Multiple-Regression HMM

Style control is an approach that enables us to intuitively change the style expressivity, i.e., emotional expressions and/or speaking styles and their intensities appearing in synthetic speech (Nose and Kobayashi 2011a). The style control technique is based on the idea of representing the variations of styles by using multiple-regression HMMs (MRHMMs) (Miyanaga et al. 2004) or multiple-regression hidden semi-Markov models (MRHSMMs) (Nose et al. 2006).

In the case of using MRHSMM, probability density functions (pdfs) for the output of the states, i.e., spectral and pitch features, and durations of the states are expressed using Gaussian pdfs with mean parameters that are assumed to be a multiple regression of a low-dimensional vector $\mathbf{s}$, i.e.,

$$\boldsymbol{\mu}_i = \mathbf{H}_{b_i} \boldsymbol{\xi} \tag{14.1}$$

$$m_i = \mathbf{H}_{p_i} \boldsymbol{\xi} \tag{14.2}$$

$$\boldsymbol{\xi} = [1, s_1, s_2, \ldots, s_L]^\top = [1, \mathbf{s}^\top]^\top \tag{14.3}$$

where $\boldsymbol{\mu}_i$ and m_i are the mean parameters of the state-output and state-duration Gaussian pdfs at state i, respectively, and $\mathbf{s} = [1, s_1, s_2, \ldots, s_L]^\top$ is a style vector in a low-dimensional style space. As shown in Fig. 14.1, each axis of the style space represents a certain style, such as joyful, sad, appealing, or storytelling, and each component of the style vector represents the expressivity intensity of a specific style. In addition, $\mathbf{H}_{b_i}$ and $\mathbf{H}_{p_i}$ are respectively $M \times (L+1)$- and $1 \times (L+1)$-dimensional regression matrices, where M is the dimensionality of the mean vector $\boldsymbol{\mu}_i$. These regression matrices are determined with maximum likelihood (ML) estimation.

In the speech synthesis phase, for an arbitrarily given style vector $\mathbf{s}$, the mean parameters of each synthesis unit are determined using (14.1) and (14.2). The speech signal is generated in the same manner as in ordinary HMM-based speech synthesis. Synthetic speech with a corresponding style intensity can be generated by setting the style vector to a desired point in the style space. Moreover, we can continuously and intuitively change the style and expressivity intensity by varying the style vector gradually along the state or phone transition.

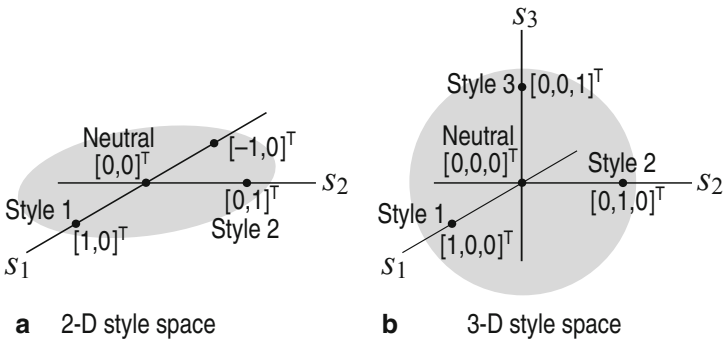


Fig. 14.1 Example of style spaces and style vectors (indicated by *dots*) for training data. **a** 2-D style space. **b** 3-D style space

14.2.2 Model Training with Perceptual Style Expressivity

The model training of MRHSMM requires style vectors for respective training speech samples. A simple choice is using a fixed style vector for each style as is done in Nose et al. (2006). As shown in Fig. 14.1, one specific vector is set as the style vector for each style independently of the intensity of expressivity appearing in respective speech samples, and it is used during the model training. Although it has been shown that this works well for expressivity control, it may cause a problem wherein we cannot always obtain synthetic speech with the desired style expressivity when the perceptual style intensity of the training data is biased, i.e., weaker or stronger than expected.

An alternative way of setting the style vector is to add the subjective style intensities into the model training (Nose and Kobayashi 2011b; Nose and Kobayashi 2013). Specifically, the style intensities perceived for the respective training utterances are quantified in listening tests and the obtained subjective scores are used as the style vectors in the model training. This leads to an additional advantage of requiring only the speech samples of the target style in the modeling of the style and intensity. In contrast, the technique using fixed style vectors (Nose et al. 2006) needs two or more styles for the model training.

14.2.3 Example of Style Expressivity Control

Figure 14.2 shows an evaluation of the controllability of style expressivity using the proposed style control technique (Nose and Kobayashi 2011b). The evaluation used emotional speech data uttered by a professional female narrator. The speech samples consisted of 503 phonetically balanced sentences with joyful and sad styles. Nine participants listened to each utterance in random order and rated the intensity of the

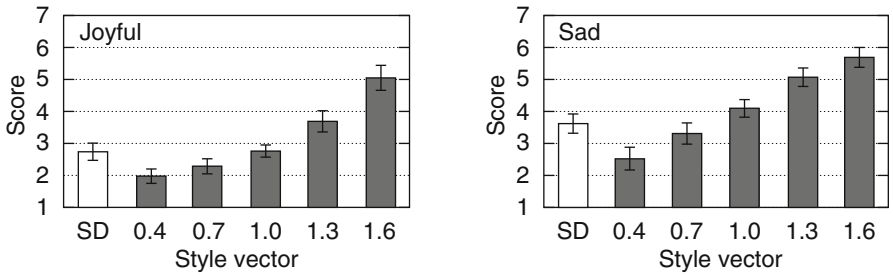


Fig. 14.2 Average style intensities of synthetic joyful and sad style speech for different style vectors. *SD* represents the result for the case of *style-dependent* HSMM trained without using style intensity information

style expressivity as "1.5" for strong, "1.0" for standard, "0.5" for weak, and "0" for not perceivable. The average score of the nine participants was taken to be the style intensity score. After the subjective scoring, 40 sentences with a perceptual score of 1.0 were chosen as the test data, since this score was expected to represent the standard expressivity of the target styles. The other 463 sentences were used as training data. Speech signals were sampled at a rate of 16 kHz, and STRAIGHT analysis (Kawahara et al. 1999) with a 5-ms shift was used to extract the spectral features. A five-state left-to-right model topology was used for the MRHSMM. The other conditions are detailed in Nose and Kobayashi (2011b).

In the parameter generation, the style vector was changed from 0.4 to 1.6 with an increment of 0.3. Seven participants listened to a speech sample chosen randomly from test sentences and rated its style expressivity by comparing it to that of a reference speech sample whose perceptual scores were 1.0. The reference samples were vocoded speech in the target style. The rating was a seven-point scale with "1" for very weak, "4" for similar to that of the reference, and "7" for very strong. The figure plots average subjective scores against given style vectors with a confidence interval of 95 %. It can be seen from the figure that the style control technique enables us to control the style intensity in accordance with the value of the style vector. Further evaluation results and detailed discussions can be found in Nose and Kobayashi (2011b, 2013).

14.3 Prosodic Variation Enhancement Using Phrase-Level F0 Context

14.3.1 Phrase-Level Prosodic Contexts

The global and local characteristics of the prosodic features of expressive speech often differ from those of neutral or reading style speech. Global prosodic features are generally well modeled using conventional statistical approaches by simply adding

a global context (Yamagishi et al. 2003). In contrast, it is not easy to model local variations, typically phrase-level ones, because they are rather diverse, depending on a variety of factors such as speakers, styles, and other paralinguistic contexts. Although we can reflect such variations in synthetic speech by using manually annotated speech samples for model training, manual annotation is time-consuming and impractical for large speech corpora. Moreover, consistent annotation is especially difficult for expressive speech. As a way of solving this problem, additional contexts for HMM-based speech synthesis, called phrase-level prosodic contexts, are defined, and they enables us to annotate training data automatically and enrich the prosodic variations of synthetic speech (Maeno et al. 2013, 2014). While the phrase-level prosodic contexts are defined for F0, duration, and power features, the following deals with only the phrase-level F0 context.

Consider a context labeling process for given training data. Let us assume that the ordinary context labels including accent phrase boundary information are available for the HMM-based speech synthesis and that conventional context-dependent HMMs using those labels are trained in advance. By using the F0s extracted from the training speech sample and the synthetic speech sample generated from the obtained HMMs, the average log F0 difference at each accent phrase is expressed as

$$d = f_o - f_s \tag{14.4}$$

where f_o and f_s are average log F0 values of the original and synthetic speech within each accent phrase. Then, for a prescribed positive value α, the phrase-level F0 context is defined as

- "Low" for $d < -\alpha$
- "Neutral" for $-\alpha \leq d < \alpha$
- "High" for $d \geq \alpha$.

Finally, every accent phrase is labeled with the above phrase-level F0 context associated with the value of d.

14.3.2 Automatic Labeling of Phrase-Level F0 Context

In the phrase-level F0 context labeling, an appropriate classification threshold α should be determined before labeling. For this purpose, an optimal α that minimizes the F0 error can be chosen using a simple grid search approach. The algorithm for obtaining the optimal threshold for the phrase-level F0 context is summarized as follows:

1. Specify a value of α between the possible lower and upper bounds, α_s and α_e.
2. Perform phrase-level F0 context labeling with the specified α for all training samples.
3. Train new HMMs using the context labels including the obtained phrase-level F0 context, and generate F0 sequences for all training samples using the newly trained HMMs and the context labels.

4. Calculate the root mean square (RMS) error E_α of log F0s between all the original and newly synthesized speech samples.
5. Specify a new value of α ($\alpha_s \leq \alpha \leq \alpha_e$) different from the value used in the previous iteration and repeat steps 2 to 4.
6. Finally, choose the optimal threshold α^* that minimizes E_α as

$$\alpha^* = \arg\min_{\alpha} E_\alpha. \tag{14.5}$$

A simple way of performing a grid search is specifying $\alpha = \alpha_s$ in the first iteration and in the nth iteration

$$\alpha = \alpha_s + (n-1)\Delta\alpha \quad (\alpha \leq \alpha_e) \tag{14.6}$$

where α is the increment in each iteration. In addition, the algorithm may use another F0 error, e.g., the maximum log F0 error, instead of the RMS error.

Note that there is a similar approach to prosodic tagging that uses the difference between the prosodic features of the generated and original speech (Suni et al. 2012; Vainio et al. 2005). However, this approach requires manually labeled training data or empirically determined weights. In contrast, the algorithm described here has the advantage that the classification threshold for the phrase-level F0 context is automatically optimized depending on the target expressive corpus without using any heuristics.

14.3.3 Model Training Example with Optimum Prosodic Labeling

The proposed phrase-level F0 context labeling was applied to two types of Japanese expressive speech data: appealing speech in sales talk and fairytale speech in storytelling. Speech samples were recorded under realistic circumstances in which no speech styles were specified to the speakers and only the target domain (situation) was made known to them (Nakajima et al. 2010). Appealing style samples were uttered by a female professional narrator, whereas fairytale style samples were uttered by a male professional narrator. The amounts of speech data of appealing and fairytale speech were approximately 33 and 52 min, respectively. Speech signals were sampled at a rate of 16 kHz, and STRAIGHT analysis with a frame shift of 5 ms was applied. A five-state left-to-right HSMM with a no-skip topology was used for the modeling of acoustic features including F0. The optimal threshold for the phrase-level F0 context was determined on the basis of fourfold crossvalidation for each style and by setting $\alpha_s = 0$, $\alpha_e = 0.3$ (519 cent), which is the maximum value of d for all accent phrases of the training data, and $\Delta\alpha = 0.01$ (17 cent). In addition, since the prosodic variation could affect the current phrase as well as the adjacent phrases, the phrase-level F0 context labels for the preceding and succeeding accent phrases as well as the current one were taken into account in the context clustering process. The other conditions are described in Maeno et al. (2014).

Table 14.1 RMS log F0 error (cent) between original and synthetic speech for test data

Style	BASELINE	PROPOSED
Appealing	254	**201**
Fairy tale	359	**273**

Table 14.1 compares RMS log F0 errors between the original and synthetic speech for test samples that were not included in the training samples. The entries for BASELINE represent the results for the model without using the phrase-level F0 context, whereas those for PROPOSED represent the case of using the phrase-level F0 context with the optimum labeling. It can be seen that the F0 distortion significantly decreased as a result of using the proposed phrase-level F0 context. Fig. 14.3 shows the F0 contours of the original and synthetic speech for a fairytale style utterance. The figure shows that the F0 reproducibility is much improved by using phrase-level F0 context labels in the HMM-based speech synthesis.

14.3.4 Prosodic Variation Enhancement for Expressive Speech Synthesis

Although the proposed phrase-level F0 context labeling solves the annotation problems for the training data and improves the F0 reproducibility of expressive synthetic speech, the technique faces a problem when it is to be applied to text-to-speech (TTS) systems. That is, the proposed phrase-level F0 context labels are not obtained from the input text automatically, because they are determined using real utterances of the target expressive speech, which are not available for arbitrarily input text.

Instead, let us consider the usage of the proposed technique when the phrase-level F0 context information for the input text is unknown. A typical example of such usage is when users want to create synthetic speech samples of voice actresses/actors with higher prosodic variability for audiobook and movie content. In this case, users first synthesize speech for the target sentence using F0 context labels whose values are all

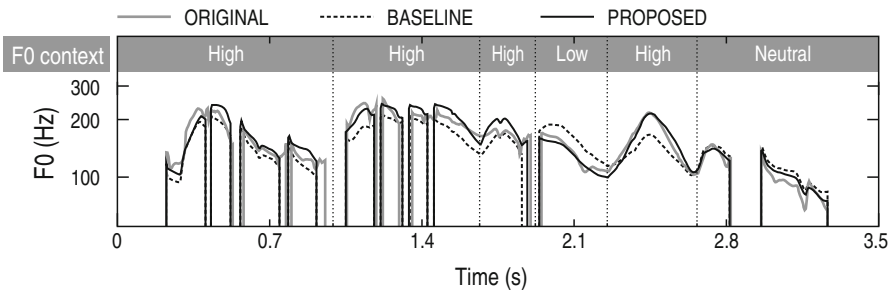


Fig. 14.3 Example of F0 contours generated with the phrase-level F0 context for fairly tale style speech sample. *BASELINE* shows the result without using the phrase-level F0 context

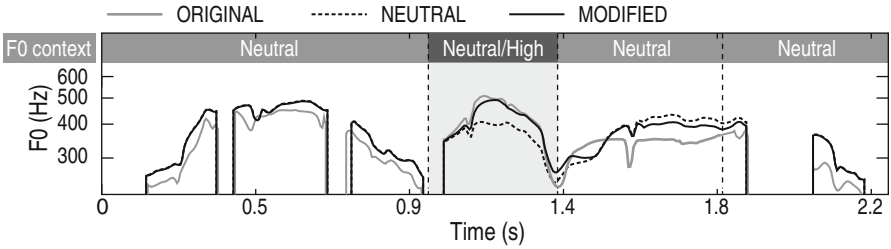


Fig. 14.4 Example of F0 contours before and after changing the F0 context for an appealing style speech sample. The F0 context in the accent phrase indicated by the shaded region was modified from "Neutral" to "High"

set to "Neutral." Synthetic speech obtained under this condition would sometimes result in poor prosodic variability compared with the real expressive speech. Next, the users listen to the synthetic speech sample and modify the F0 context of a certain accent phrase to "High" or "Low" if they want to enhance the F0 variation in that phrase. Then they synthesize speech again with the modified context labels and check the resultant synthetic speech. By repeating this procedure, users can obtain better synthetic speech in terms of F0 variations.

Figure 14.4 illustrates an example of this F0 variation enhancement process. Conditions for the phrase-level F0 context labeling and model training are the same as described in Sect. 14.3.3. A user first listened to an appealing style synthetic speech sample generated with the phrase-level F0 context labels being set all "Neutral" (the F0 contour is denoted as NEUTRAL in the figure). Since the user felt a lack of prominence in the second accent phrase, its phrase-level context label was changed to "High" and the speech sample was resynthesized. The figure shows that the resultant F0 contour denoted by MODIFIED became closer to that of the real speech denoted by ORIGINAL. Further experimental results and detailed discussions are provided in Maeno et al. (2013, 2014).

14.4 Conclusions

Techniques of prosody control and prosodic variation enhancement were discussed for HMM-based expressed speech synthesis. First, a brief review of the prosody control technique based on multiple regression HSMMs (MRHSMM) was given. Then subjective style intensities was incorporated into the technique to achieve more intuitive control of the styles. The use of subjective style intensities in the training of MRHSMMs normalizes the variation of style intensities appearing in the training data, and this results in an intensity that users would deem normal for the style of speech (Nose and Kobayashi 2013). Moreover, the training of the MRHSMMs can be done using only data for a single style by introducing style intensity scores for the respective training samples.

Next, an unsupervised labeling technique for phrase-level prosodic variations was described. The technique can be used to enhance the prosodic variation of synthetic expressive speech. A new prosodic context, the phrase-level F0 context, for HMM-based speech synthesis was defined and a fully automatic labeling algorithm for the newly defined context was described. Experiments on the prosodic context labeling revealed that the variations in the F0 feature appearing in the training samples were effectively captured with the proposed technique. Although phrase-level F0 context labels are unknown for arbitrary input text in practical TTS situations, the technique enables users to intuitively enhance the prosodic characteristics of a target accent phrase by manually changing the proposed context label from "Neutral" to "High" or "Low."

Acknowledgement The author would like to thank T. Nose, Y. Maeno, and T. Koriyama for their contributions to this study at Tokyo Tech. He would also like to thank O. Yoshioka, H. Mizuno, H. Nakajima, and Y. Ijima for their helpful discussions and providing expressive speech materials.

References

Anastasakos, T., J. McDonough, R. Schwartz, and J. Makhoul. 1996. A compact model for speaker adaptive training. Proceedings of ICSLP, 1137–1140.

Cowie, R., and R. R. Cornelius. 2003. Describing the emotional states that are expressed in speech. *Speech Communication* 40 (1–2): 5–32.

Doukhan, D., A. Rilliard, S. Rosset, M. Adda-Decker, and C. d'Alessandro. 2011. Prosodic analysis of a corpus of tales. Proceedings of INTERSPEECH, 3129–3132.

Erickson, D. 2005. Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology* 26 (4): 317–325.

Eyben, F., S. Buchholz, N. Braunschweiler, J. Latore, V. Wan, M. J. F. Gales, and K. Knill. 2012. Unsupervised clustering of emotion and voice styles for expressive TTS. Proceedings of ICASSP, pp. 4009–4012.

Gales, M. J. F. 2000. Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing* 8 (4): 417–428.

Kawahara, H., I. Masuda-Katsuse, and A. de Cheveigne. 1999. Restructuring speech representations using a pitch-adaptive time–Frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication* 27 (3–4): 187–207.

Koriyama, T., T. Nose, and T. Kobayashi. 2011. On the use of extended context for HMM-based spontaneous conversational speech synthesis. Proceedings of INTERSPEECH, 2657–2660.

Maeno, Y., T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka. 2011. HMM-based emphatic speech synthesis using unsupervised context labeling. Proceedings of INTERSPEECH, 1849–1852.

Maeno, Y., T. Nose, T. Kobayashi, T. Koriyama, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka. 2013. HMM-based expressive speech synthesis based on phrase-level F0 context labeling. Proceedings of ICASSP, pp. 7859–7863.

Maeno, Y., T. Nose, T. Kobayashi, T. Koriyama, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka. 2014. Prosodic variation enhancement using unsupervised context labeling for HMM-based expressive speech synthesis. *Speech Communication* 57:144–154.

Miyanaga, K., T. Masuko, and T. Kobayashi. 2004. A style control technique for HMM-based speech synthesis. Proceedings of INTERSPEECH-ICSLP, 1437–1440.

Nakajima, H., N. Miyazaki, A. Yoshida, T. Nakamura, and H. Mizuno. 2010. Creation and analysis of a Japanese speaking style parallel database for expressive speech synthesis. http://desceco.org/O-COCOSDA2010/proceedings/paper_30.pdf. Accessed 6 Dec 2014.

Nose, T., and T. Kobayashi. 2011a. Recent development of HMM-based expressive speech synthesis and its applications. Proceedings of APSIPA ASC. http://www.apsipa.org/proceedings_2011/pdf/APSIPA189.pdf. Accessed 6 Dec 2014.

Nose, T., and T. Kobayashi. 2011b. A perceptual expressivity modeling technique for speech synthesis based on multiple-regression HSMM. Proceedings of INTERSPEECH, 109–112.

Nose, T., and T. Kobayashi. 2013. An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. *Speech Communication* 55 (2): 347–357.

Nose, T., J. Yamagishi, and T. Kobayashi. 2006. A style control technique for speech synthesis using multiple-regression HSMM. Proceedings of INTERSPEECH-ICSLP, 1324–1327.

Obin, N., A. Lacheret, and X. Rodet. 2011a. Stylization and trajectory modelling of short and long term speech prosody variations. Proceedings of INTERSPEECH, 2029–2032.

Obin, N., P. Lanchantin, A. Lacheret, and X. Rodet. 2011b. Discrete/continuous modelling of speaking style in HMM-based speech synthesis: Design and evaluation. Proceedings of INTERSPEECH, 2785–2788.

Schröder, M. 2009. Expressive speech synthesis: Past, present, and possible futures. In: *Affective information processing*, ed. J. H. Tao and T. N. Tan, 111–126. London: Springer.

Suni, A., T. Raitio, M. Vainio, and P. Alku. 2012. The GlottHMM Entry for Blizzard Challenge 2012: Hybrid Approach. Proceedings of Blizzard Challenge Workshop. http://festvox.org/blizzard/bc2012/HELSINKI_Blizzard2012.pdf. Accessed 6 Dec 2014.

Székely, E., J. Cabral, P. Cahill, and J. Carson-Berndsen. 2011. Clustering expressive speech styles in audiobooks using glottal source parameters. Proceedings of INTERSPEECH, 2409–2412.

Tachibana, M., J. Yamagishi, T. Masuko, and T. Kobayashi. 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE Transactions on Information and Systems* E88-D (11): 2484–2491.

Vainio, M., A. Suni, and P. Sirjola. 2005. Accent and prominence in Finnish speech synthesis. Proceedings of International Conference on Speech and Computer (SPECOM), 309–312.

Yamagishi, J., K. Onishi, T. Masuko, and T. Kobayashi. 2003. Modeling of various speaking styles and emotions for HMM-based speech synthesis. Proceedings of INTERSPEECH, 2461–2464.

Yu, K., H. Zen, F. Mairesse, and S. Young. 2001. Context adaptive training with factorized decision trees for HMM-based statistical parametric speech synthesis. *Speech Communication* 53 (6): 914–923.

Yu, K., F. Mairesse, and S. Young. 2010. Word-level emphasis modelling in HMM-based speech synthesis. Proceedings of ICASSP, 4238–4241.

Zen, H., K. Tokuda, and A. Black. 2009. Statistical parametric speech synthesis. *Speech Communication* 51 (11): 1039–1064.