Kinematic Modelling of Dipthong Articulation

Cristina Carmona-Duarte, Jesús B. Alonso, Moises Diaz, Miguel A. Ferrer, Pedro Gómez-Vilda and Réjean Plamondon

Abstract The Sigma-Lognormal model has successfully been applied to handwriting modeling but never to acoustic-phonetic articulation. The hypothesis of this paper is that vocal tract dynamics, which includes jaw and tongue can be approached by the Kinematic Theory. In speech analysis, the movement of the tongue and jaw has been linked to the variation on first and second formants. In this paper, we explore the Kinematic hypothesis, based on diphthong pronunciation, which invoke the most extreme tongue gestures in the vowel triangle, estimation of their formants, and transformation of these to space for evaluating the speed profile. The estimated speed profile is modelled by the sigma lognormal model of the Kinematic Theory. An average reconstruction error of 20 dB has been obtained in the experiments carried out with 20 different volunteers. This result validates the work hypothesis, opening a new research line in speech processing.

Keywords Sigma-Lognormal · Formant · Kinematic · Diphthong · Articulation

C. Carmona-Duarte (✉) · J.B. Alonso · M. Diaz · M.A. Ferrer
Instituto Universitario para el Desarrollo Tecnológico y la Innovación en Comunicaciones,
Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain
e-mail: ccarmona@idetic.eu

J.B. Alonso
e-mail: jalonso@idetic.eu

M. Diaz
e-mail: mdiaz@idetic.eu

M.A. Ferrer
e-mail: mferrer@idetic.eu

P. Gómez-Vilda
Facultad de Informática, Universidad Politécnica de Madrid,
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, Spain
e-mail: pedro@fi.upm.es

R. Plamondon
Laboratoire Scribens, Déartement de Génie Électrique,
École Polytechnique de Montréal, Montreal, Canada
e-mail: rejean.plamondon@polymtl.ca

A. Esposito et al. (eds.), *Recent Advances in Nonlinear Speech Processing*,
Smart Innovation, Systems and Technologies 48,
DOI 10.1007/978-3-319-28109-4_6

1 Introduction

It is well known in speech analysis that a specific vowel phonation depends on the velo-pharyngeal switch, the tongue neuromotor system, the mandibular system and the laryngeal system. In the case of vowels, a space representation can be obtained from the two first formants (F1 and F2) [1]. For instance, in [2] the jaw-tongue dynamics is considered as the basis to explain the space formant distribution in running speech. Besides, [2] shows a simple model to estimate jaw-tongue dynamics from the two first formants.

The Kinematic Theory of rapid human movements [3] describes the way in which neuromuscular systems are involved in the production of muscular movements. This theory has been applied successfully to handwriting, analyzing the neuromuscular system involved in the production of rapid movements [4], the variations of handwriting with time [5, 6], the prevention of brain strokes [7], the specification of a diagnostic system for neuro-muscular disorder [9], etc.

In the case of handwriting, the arm and trunk muscles act to generate the handwriting. Similarly, the tongue and jaw muscles are moving to generate different vowels sounds during the speech articulation. So, our hypothesis includes that both speech and handwriting signals could be studied as a human movement. Furthermore, a parallelism can be established between both signals.

In the present paper, a method to estimate the speed profile from the two first formants is introduced, using the Sigma-Lognormal model to foresee the possibility of its applicability to model speech dynamics.

The paper is organized as follows: in Sect. 2 an introduction to the physiology of phonation and the method used to transform formants into distance is presented. In Sect. 3, a brief description of the Sigma-Lognormal model is introduced. In Sect. 4, we present the method followed to estimate the speed signal from the two first formants. The results of the experiments carried out over 20 subjects are presented. Finally, conclusions are commented in Sect. 5.

2 Physiology of Phonation

The articulatory organs and nasal cavity allow focusing the energy of the speech signal at certain frequencies (formants), due to oropharyngeal tract resonators. Estimating the resonance or formant structure of voiced speech is possible from a digital inverse filter formulation. There are numerous techniques to perform the inverse filtering of a speech signal as the Iterative Adaptive Inverse Filtering algorithm [9], which can provide an adequate estimation of the glottal excitation. However, a linear prediction model based on an autoregressive process (AR) [10] is enough to determine the formants (in non-nasal phonations). In this case, speech signal $s(n)$ can be modelled as follows:

$$s(n) = \sum_{\mathrm{i=1}}^{\mathrm{N_{LP}}} a_{LP}(i)s(n-i) + e(n) \tag{1}$$

where N_{LP} represents the order of the predictor, $\{a_{LP}\}$ are the coefficients of linear prediction (LPC) and $e(n)$ represents the error in the model. The LPC coefficients are calculated by least squared error algorithms and define the transfer function of the vocal tract, V(z), assumed to be given as an all-pole function:

$$V(z) = \frac{G}{1 - \sum_{i=1}^{N_{LP}} a_{LP}(i)z^{-i}} = \frac{G}{\prod_{i=1}^{N_{LP}} (1 - p_{LP}(i)z^{-1})} \tag{2}$$

The poles characterize the formants which are the local maxima of the spectrum, where the first (F1) and second (F2) formants correspond to the two first maximum values in the LPC spectrum. The estimation of the predictor order is based on the sample frequency: for an fs of 22050 samples/s a compromise is to use 15 coefficients.

Also, it is well known that in the vowel phonation the formants F1 and F2 vary for each vocals creating a vowel triangle [1] (Fig. 1).

The acoustic representation spaces are associated to jaw position, articulation place and lip rounding. Assuming a simplification in the acoustic representation, the closed-open gesture and a back-front gesture can be defined. The first is produced by the muscles involved in the jaw movements and the second by the tongue movements. These movements can be correlated with the formants positions in plane F1 versus F2 [2] as:

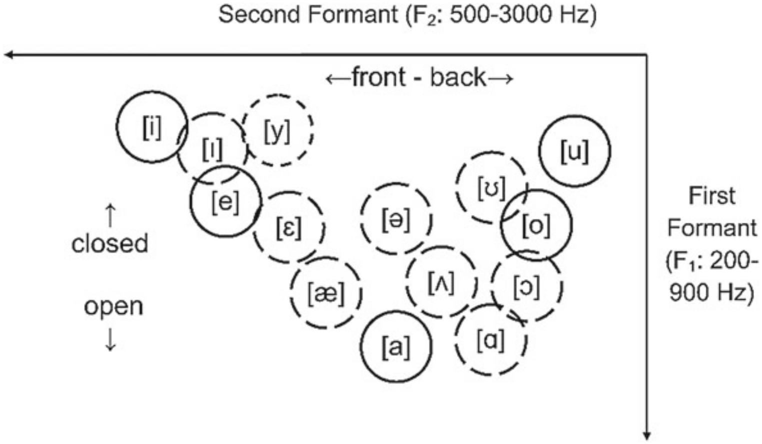


Fig. 1 Vowel representation spaces adapted from [1]. Spanish (*full circle*) and American English (*long-dash circle*)

$$\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} = \begin{bmatrix}c_{11} & c_{12}\\ c_{21} & c_{22}\end{bmatrix}\begin{bmatrix}\Delta F1\\ \Delta F2\end{bmatrix} \tag{3}$$

where Δx and Δy are the relative displacement from an initial position of the jaw and tongue (x_0, y_0). c_{ij} are the weights of the combination matrix.

3 Overview of the Sigma-Lognormal Model

In this work, we applied a Sigma-Lognormal model [11] in order to parameterize the speed profile of the tongue movement. This model considers the resulting speed of a neuromuscular system action or stroke describing a lognormal function scaled by a command (D) and time-shifted by the time occurrence of the command (t_0) [11]. The complex pattern is produced by summing each resulting lognormal function, given by equation:

$$\overrightarrow{v_n}(t) = \overrightarrow{\sum} \bigwedge(t) = \sum\nolimits_{i=1}^{M} \vec{v}_j(t) \tag{4}$$

where M represents the number of simple movements involved in the generation of a given pattern, $\bigwedge$(t) is the Sigma-Lognormal and $\vec{v}_j(t)$ is the velocity profile of the jth stroke.

The speed in the Cartesian space can be calculated as:

$$v_{n\_x}(t) = \sum\nolimits_{i=1}^{M} |\vec{v}_j(t)| \cos(\emptyset_j) \tag{5}$$

$$v_{n\_y}(t) = \sum\nolimits_{i=1}^{M} |\vec{v}_j(t)| \sin(\emptyset_j) \tag{6}$$

where $\emptyset_j$ is the direction angle in the jth stroke.

Given v_{n_x} and v_{n_y}, the goodness of their reconstruction from the Sigma-Lognormal domain, is given by the error between the original and its reconstructed signal, which must be as minimum as possible. This criterion can be evaluated using the Signal-to-Noise-Ratio (SNR) between the reconstructed speed profile ($\vec{v_v}(t)$) and the original one ($\vec{v_v}(t)$). In this way, the SNR is defined as:

$$SNR = 20\log\left(\frac{\int_{t_s}^{t_n}[v_{x\_n}^2(t) + v_{y\_n}^2(t)]dt}{\int_{t_s}^{t_n}[\left(v_{x\_n}^2(t) + v_x^2(t)\right)^2 + (v_{yn}^2(t) - v_y^2(t))^2]dt}\right) \tag{7}$$

The higher is the SNR the better is the reconstruction. Generally speaking, a SNR greater than 20 dB provides excellent signal reconstruction.

4 Methods

4.1 Subjects

Twenty mid age healthy subjects (fifteen males and five females) participated in the experiment. They were recorded during the utterance of three different diphthong phonations pronounced by Spanish speakers (/au/, /iu/, and /ai/).

4.2 Formant Estimation

Formants F1 and F2 for each sample (see Fig. 2) were estimated. Then, Eq. 3 was used to transform the two formants into a space representation. For this purpose, the study in [12] was taken as a reference to estimate a first approximation to the coefficients. In Fig. 3, we can observe the special representation of the Δx and Δy obtained from the F1 and F2 from one subject and the three analyzed diphthongs.

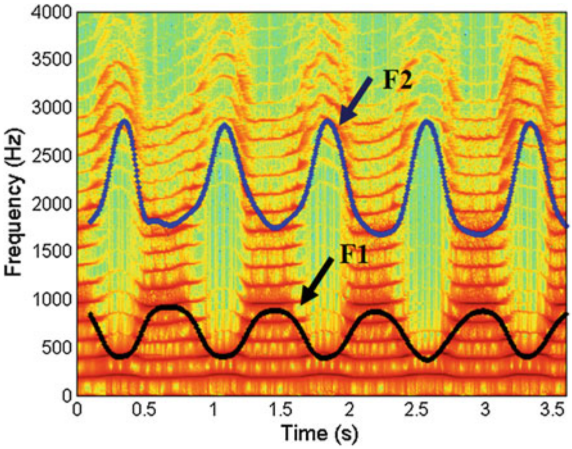
4.3 Speed Profile

The speed profile $\vec{v}(t)$ from the calculated Δx and Δy was estimated with:

$$|\vec{v}(t)| = \sqrt{\Delta x^2 + \Delta y^2} \tag{8}$$

The resulting speed profile can be seen in Fig. 4.

Fig. 2 Formant estimation for diphthong /ai/



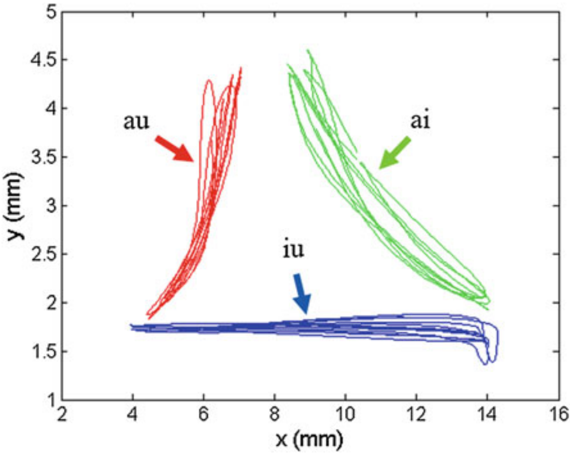


Fig. 3 Formant to position transformation

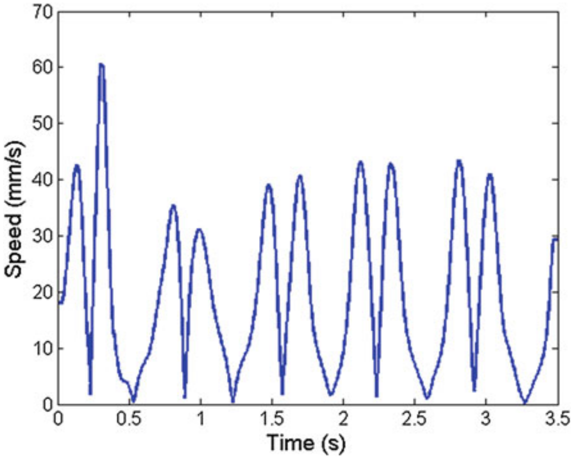


Fig. 4 Speed profile for /ai/ phonation

Finally, the quality of the reconstruction from formants to Sigma-Lognormals was evaluated. For such, The Sigma-Lognormal model was used in order to parameterize the resulting speed profile. Then, the SNR (Eq. 7) between the reconstructed signal and the original one was calculated. The results of the experiment are presented in the next section.

5 Experiments and Results

In the experiment, 4 s long utterances of the three diphthongs (/ai/, /iu/ and /au/) were recorded at 22,050 Hz and 16 bits resolution. As it was explained in previous

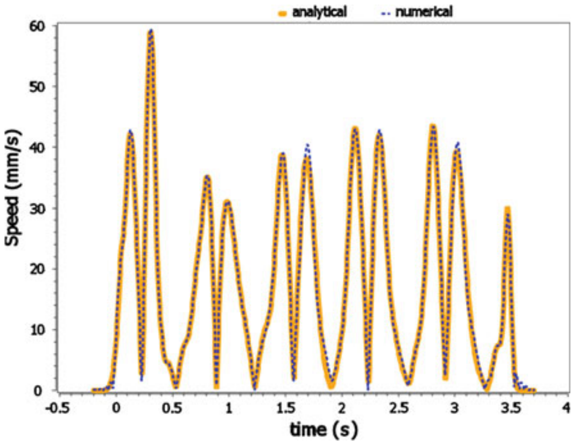


Fig. 5 Corresponding speed profile (*full line*) and its Sigma-Lognormal decomposition (*dot line*)

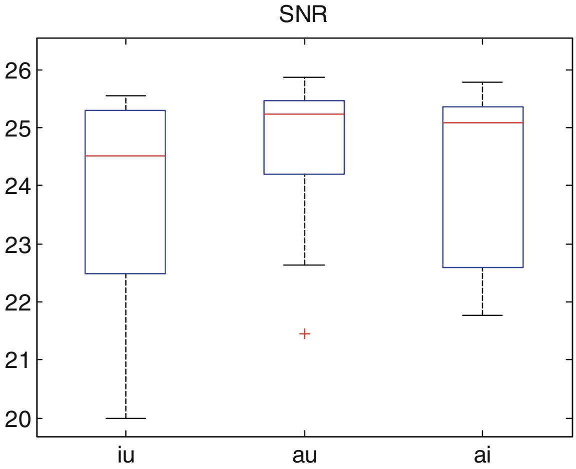


Fig. 6 Signal-to-Noise-Ratio for each diphthong

section, formants positions F1-F2 were extracted each 5 ms and the speed profile was estimated.

Once the speed profile is calculated, the reconstructed speed profile is extracted automatically, using the Sigma-Lognormal model explained briefly in Sect. 3 [11]. A typical original speed profile and its Sigma-Lognomal reconstruction one are shown in Fig. 5. It can be observed how close to the original profile the reconstructed speed profile is. In this particular example the SNR is 25 dB.

The SNR was estimated from Eq. 7. The results are shown in Fig. 6. As it is seen from these results, the SNR is greater than 20 dB, which assesses the excellence of fitting the original speed with the Kinematic Theory as in the case of handwriting [6].

6 Conclusions

In this paper, the possibility of modeling formant dynamics as a function of the articulation organ position representation space has been introduced and evaluated. In order to validate the work hypothesis, the original speed profile was estimated from the distance representation of the formants. This speed profile has been reconstructed using the Kinematic Theory and its associated Sigma Lognormal model. The results obtained from the experiments, grant an estimation fit showing a signal-noise ratio of more than 20 dB, which could be considered as an excellent validation figure of this hypothesis.

These first results open new ways to model speech dynamics and the possibility of applying these advances to handwriting, and to the study of neurodegenerative speech production.

Acknowledgments This study was funded by the Spanish government's MCINN TEC2012-38630-C04 research project and the fellowship program of Universidad de Las Palmas de Gran Canaria.

References

1. Peterson, G.E., Barney, H.L.: Control methods used in a study of the vowels. J. Acoust. Soc. Am. **24–2**, 175–184 (1952)
2. Gómez-Vilda, P., Londral, A.R.M., Rodellar-Biarge, V., Ferrández-Vicente, J.M., de Carvalho, M.: Monitoring amyotrophic lateral sclerosis by biomechanical modeling of speech production. Neurocomputing **151**, 130–138 (2014)
3. Plamondon, R.: A kinematic theory of rapid human movements: part I: movement representation and generation. Biol. Cybern. **72**, 295–307 (1995)
4. Plamondon, R., Djioua, M., Mathieu, P.: A time-dependence between upper arm muscles activity during rapid movements: observation of the proportional effects predicted by the kinematic theory. Hum. Mov. Sci. **32**, 1026–1039 (2013)
5. Gomez-Barrero M., Galbally J., Plamondon R., Fierrez J., Ortega Garcia J.: Variations of handwritten signatures with time: a Sigma-Lognormal analysis. In: Proceedings 6th International conference on Biometrics, pp. 3.16.1–3.16.6. Madrid, Spain, 4–7 June 2013
6. Plamondon, R., O'Reilly, C., Rémi, C., Duval, R.C.: The lognormal handwriter: learning, performing and declining. Front. Psychol. Cogn. Sci. (2013). doi:10.3389/fpsyg.2013.00945, Special Issue in Cognitive Science, Writing words: From brain to hand(s), Topic Editor(s): Sonia Kandel, Marieke Longcamp, 1–14
7. Plamondon, R., O'Reilly, C., Ouellet-Plamondon, R.: Strokes against strokes: strokes for strides. Pattern Recognit. **47**, 929–944 (2014)
8. O'Reilly, C., Plamondon, R.: Design of a neuromuscular disorders diagnostic system using human movement analysis. In: 11th International Conference on Information Sciences, Signal Processing and their Applications, Montreal, Canada, 3–5 July 2012
9. Alku, P.: Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. **11**(2), 109–118 (1992)
10. Itakura, F.: Line spectrum representation of linear predictor coefficients of speech signals. J. Acoust. Soc. Am. **57**(S1), S35–S35 (1975)
11. O'Reilly, C., Plamondon, R.: Development of a Sigma-Lognormal representation for on-line signatures. Pattern Recognit. **42**, 3324–3327 (2009)
12. Dromey, C., Jang, G.O., Hollis, K.: Assessing correlations between lingual movements and formants. Speech Commun. **55**, 315–328 (2013)