ACOUSTIC-ARTICULATORY EVALUATION OF THE UPPER VOWEL-FORMANT REGION AND ITS PRESUMED SPEAKER-SPECIFIC POTENCY

Frantz CLERMONT†* and Parham MOKHTARI‡

† College of Information Sciences, University of Tsukuba,
1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, JAPAN (frantz@milab.is.tsukuba.ac.jp)

‡ Laboratoire Lorrain de Recherche en Informatique et ses Applications,
B.P. 239, 54506 Vandœuvre-les-Nancy Cedex, FRANCE (parham@loria.fr)

ABSTRACT

We present some evidence indicating that phonetic distinctiveness and speaker individuality, are indeed manifested in vowels' vocal-tract shapes estimated from the lower and the upper formant-frequencies, respectively. The methodology developed to demonstrate this dichotomy, first implicates Schroeder's [8] acoustic-articulatory model which can be coerced to yield, on a per-vowel and a per-speaker basis, area-function approximations to vocal-tract shapes of differing formant components. Using ten steady-state vowels recorded in /hVd/-context, five times at random, by four adult-male speakers of Australian English, the variability of resulting shapes aligned at mid-length was then measured on an intra- and an inter-speaker basis. Gross shapes estimated from the lower formants, were indeed found to cause the largest spread amongst the vowels of individual speakers. By contrast, the more detailed shapes obtained by recruiting certain higher formants of the front and the back vowels, accounted for the largest spread amongst the speakers. Collectively, these results contribute a quasi-articulatory substantiation of a long-standing view on the speaker-specific potency of the upper formant region of spoken vowels, together with some useful implications for automatic speech and speaker recognition.

1. INTRODUCTION

Our primary concern in the study reported in this paper, was to obtain acoustic-articulatory evidence substantiating the view held for many decades, that speaker individuality should be more strongly manifested in the upper formant region of spoken vowels. As phonetic distinctiveness amongst vowels uttered by individual speakers is well understood to be mostly encoded in the two lowest formants (F_1 and F_2), it is plausible that the higher formants should then carry more speaker-specific information. This conjecture was indeed offered in 1936 by Lewis [2] in a remarkable study of sung vowels, which preceded the invention of the spectrograph by a decade. In 1959, however, Peterson [7] put forward the more engaging, articulatory hypothesis that the low formants depend more on gross vocal-tract shapes, while the higher formants depend on more exact cavity sizes and constrictions. Peterson's hypothesis then implies that it is the detailed behaviour of the vocal-tract in spoken vowels, which would embody more speaker individuality and which would itself be attributable to the presence of the higher formants. While this phonetic-speaker contrast in articulatory shapes could almost be inferred from Mermelstein's [4] modelling perspective on the vocal-tract geometry, some quantitative and more explicit evidence arose from Liljencrants' [3] correlation analysis of "Fourier descriptors" of two male speakers' vowel-tongue profiles. A shape-based, inter-speaker comparison indeed yielded strong correlations for the DC term and the spatial fundamental frequency, but weak correlations for the second harmonic. The former result led Liljencrants to advance that "gross shapes are similar" for the two speakers articulating different vowels, while he attributed the latter result to the plausible cause of "a more subject dependent fine structure".

Liljencrants' study provides data which we consider to be seminal since they point to the suspected role of the higher formants of spoken vowels, as well as foreshadow the possibility of obtaining articulatory and thus more direct insights into the related question of manifestations of speaker individuality. To this end, we adopted an approach implicating Schroeder's [8] acoustic-articulatory model, which yields area-function approximations to vocal-tract shapes directly from the formants, and therefore lends itself to the estimation of shapes of differing formant components. This fortunate flexibility was thus exploited in an experiment aimed at evaluating the relative importance of the lower and upper formant regions, by quantifying vowel-shape variability on an intra- and an inter-speaker basis.

*On sabbatical leave from U.N.S.W. (U. College), Australia.

2. VOWEL DATASET AND FORMANT ESTIMATION

Since spoken vowels were the sounds of interest in this study and their steady-states are more readily apprehended in acoustic-articulatory terms, we used the time-honoured /hVd/-context in which coarticulatory effects are considered to be minimal, and restricted ourselves to stationary parts of the vocalic nuclei. We also sought to secure a moderate degree of intra- and inter-speaker variability by obtaining not only a good coverage of the vowel space, but also multiple tokens of each vowel on an intra- and an inter-speaker basis. Following these prescriptions, a set of 10 monosyllables with expected front- (HEED, HID, HEAD, HAD, WHO'D) and back- (HARD, HOD, HOARD, HOOD, HUD) vowel nuclei were recorded by 4 adult male and native speakers of Australian English, 5 times at random and on a single occasion. Recordings were conducted in a sound-proof room, and the analogue data were digitised with 12-bit quantisation and at 10 kHz sampling-frequency [1].

The /hVd/-syllables thus collected were analysed using 14th-order, Linear-Prediction (LP) analysis of Hamming-windowed frames of 25.6 msec duration with a frame advance of 5 msec, from which LP-cepstra and LP-poles were obtained. A semi-automatic segmentation algorithm [5] was then employed to generate an inter-frame (cepstral) variance function, around the minimum of which we retained 7 consecutive, steady-state frames for each vowel nucleus. At these frames, the first 4 formants (F_1 to F_4) were estimated using the LP-poles, and an unsupervised tracking method [1] which combines dynamic programming with analysis-by-synthesis. Each of these frequencies was finally averaged over the 7 steady-state frames of every nucleus.

3. VOCAL-TRACT (VT) SHAPE ESTIMATION

In order to quantify the suspected dichotomy between high and low formants in quasi-articulatory terms, we require an articulatory parameterisation which relates as closely as possible to the formants themselves. This requirement does implicate Schroeder's model, which maps formants into corresponding area-functions (or shape approximations) of a lossless vocal-tract. The model is governed by Equation 1, which provides a parametric expression for a logarithmic area-function $lnA(x)$ at x units of length from the glottis:

$$lnA(x) = lnA_0 + \sum_{n=1}^{M} a_{2n-1} \cdot cos[\frac{(2n-1)\pi}{L}x] \ , \quad (1)$$

where

$$a_{2n-1} = -2\frac{F_n - F_{n0}}{F_{n0}} \quad (2)$$

describes the relation between each formant-frequency F_n and a unique, odd-indexed coefficient of the Fourier cosine series used to represent $lnA(x)$; $F_{n0} = \frac{(2n-1)c}{4L}$ defines the nth resonance frequency of a uniform tract of length L; and c is the speed of sound in air. The odd-indexed cosine terms of the series imply that only antisymmetric components of $lnA(x)$ are retained and, while this constraint may not yield realistic shapes for certain vowels, it does alleviate the uniqueness problem [8, 4].

As the model assumes lossless conditions, the area-scaling parameter A_0 is acoustically inconsequential and thus set to unity in Equation 1, which effectively renders normalised, logarithmic area functions. As for the parameter L, we adopted Paige and Zue's [6] formulation:

$$L = \frac{c\sum_{n=1}^{M}[F_n/(2n-1)]}{4\sum_{n=1}^{M}[F_n/(2n-1)]^2} \ , \quad (3)$$

which hinges on the now well-known criterion of minimum VT-shape eccentricity from a uniform tube. In this vein, Paige and Zue's study as well as Wakita's [9] suggest further that the greater the number of formants used, the more realistic is the length estimate likely to be. All of our four, measured formants were therefore used to estimate our vowel-shapes' VT-lengths.

On a per-speaker, a per-token and a per-vowel basis, our procedure for VT-shape estimation then consisted of first estimating L using Equation 3 and $M = 4$, followed by the calculation of $\{a_1, a_3, a_5, a_7\}$ via Equation 2. For a given L, Equation 1 affords the flexibility of varying the spatial resolution of a VT-shape, by simply retaining different subsets of the original $\{a_{2n-1}\}$-set.

Using a differential x value of 0.1 cm, all our VT-shapes were thus estimated from 6 different combinations of the $\{a_{2n-1}\}$, corresponding to either a gradual recruitment of the higher formants $\{F_1,\ F_1F_2,\ F_1F_2F_3$ and $F_1F_2F_3F_4\}$, or to the pairwise combinations $\{F_1F_2,\ F_2F_3$ and $F_3F_4\}$. Illustrations of the former set are shown in Figure 1, where the 4-speakers' VT-shapes of one token of the (7-frame averaged) steady-state vowel in HEED were first centre-aligned, then superimposed onto one another. Shape alignment at mid-length is first motivated by the very antisymmetry property of Schroeder's model. But it is also an acknowledgement of the "tendency for the position of the lips and of the larynx to vary from one articulatory configuration to another", as aptly argued by Mokhtari [5].

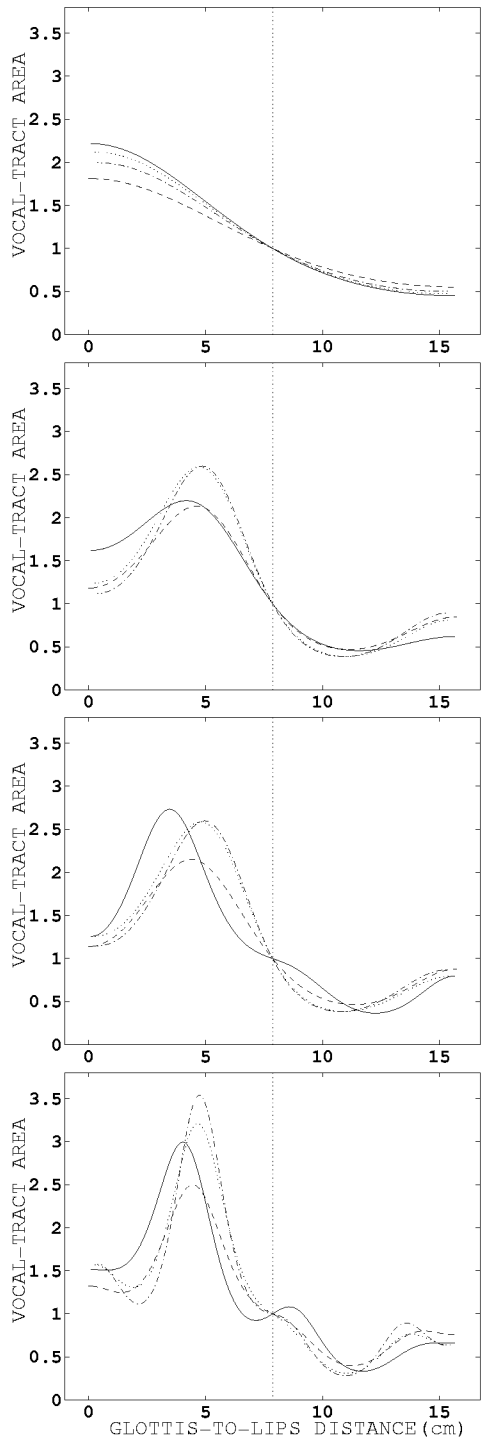

Figure 1: Multi-speaker, Vocal-Tract (VT) shapes estimated from $\mathbf{F_1}$ (top graph), $\mathbf{F_{12}}$ (second from top graph), $\mathbf{F_{123}}$ (third from top graph) and $\mathbf{F_{1234}}$ (bottom graph) of one token of the (7-frame averaged) steady-state of the vowel in **HEED**. Four (4) speakers' VT-shapes were first aligned at their respective mid-length, and then superimposed on each graph. Notation: $\mathbf{F}_{12} \equiv \mathbf{F}_1\mathbf{F}_2$, and so on.

4. VT-SHAPE MANIFESTATIONS OF PHONETIC-SPEAKER DICHOTOMY

The graphs offered in Figure 1 first illustrate the flexibility of the acoustic-articulatory parameterisation adopted in this study. However, they also provide some preliminary evidence pointing to the apparent fact that, as the higher formants are recruited, the shapes' gross outlines remain relatively invariant, whilst details of their features near constricted and expanded regions are progressively acquired and setting our 4 speakers apart from one another. It is this type of phenomenon which we had indeed been hoping to observe, and hence attempted to investigate quantitatively.

To this end, we devised a two-part experiment in order to evaluate the relative importance of the lower and higher formants, in terms of shape variability expressed as root-mean-square distances amongst all vowels of individual speakers and amongst individual vowels of all speakers. Note that distances between centre-aligned and mutually-overlapping shapes [5] were computed using logarithmic areas, rather than their exponentials which admittedly bear more resemblance to directly-measured shapes and are visually more informative as we chose to show in Figure 1. However, a logarithmic scale ensures that differences in expanded regions are effectively de-weighted relative to those in constricted regions. Consequently, our quasi-articulatory distance is rendered more sympathetic to the fact that the formants themselves are more sensitive to variations in constricted regions. Furthermore, it is an inherent property of Schroeder's model, that each formant is mapped to a Fourier cosine component of the logarithmic area-function.

In part 1 of the experiment outlined above, we computed for each token and each speaker at a time, all unique pairs of shape distances amongst the 10 vowels; and averaged those distances over all such pairs, then over each speaker's 5 tokens and finally over all speakers (see top of Figure 2). In part 2 we computed, for each token and each vowel at a time, all unique pairs of shape distances amongst the 4 speakers; and averaged those distances over all such pairs, then over all tokens and over back and front vowels separately, and finally over all vowels (see bottom of Figure 2). In addition, all these computations were performed for each of the 6 formant combinations specified earlier. Our two-part experiment thus yielded a dual set of shape distances, which provided the basis for the following interpretations.

Our first distance profile given on top of Figure 2, indicates that the spread amongst vowels is greater for F_1F_2- than for F_2F_3- and F_3F_4-based shapes, in that order. This contrast represents quantitative evidence that phonetic distinctiveness is indeed embodied more strongly in gross shapes arising from their lower frequency-components. It is not entirely surprising, but all the same pleasing to also observe that relatively little phonetic information is gained in VT-shapes obtained by recruitment of the higher formants F_3 and F_4.

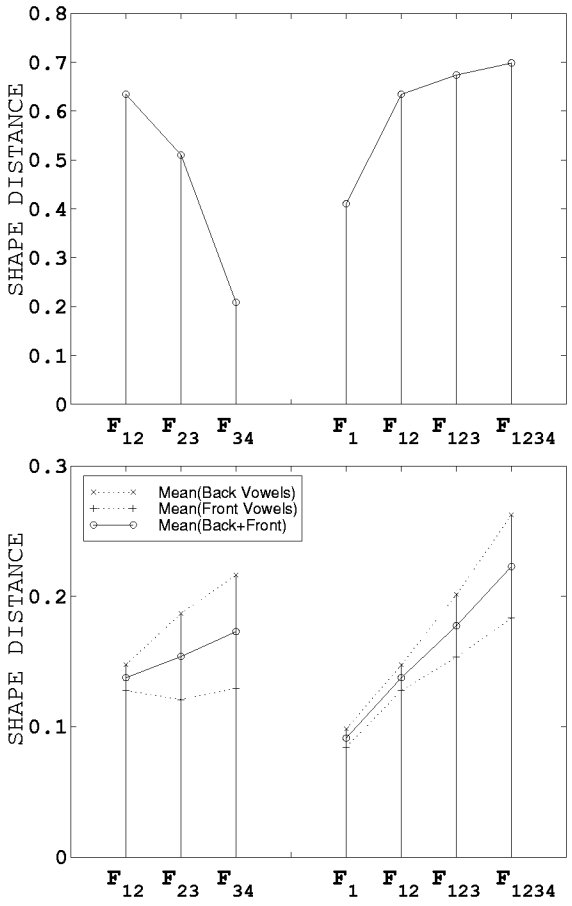

Figure 2: Profiles of distances amongst VT-shapes based on 6 different formant-components. Top graph: Experiment/part 1 (inter-vowel). Bottom graph: Experiment/part 2 (inter-speaker: back, front & all vowels). Notation: $\mathbf{F_{12}} \equiv \mathbf{F_1F_2}$, and so on.

On the other hand, our second distance profile given at the bottom of Figure 2 shows that, for back vowels, the spread amongst speakers is greater for F_3F_4- and F_2F_3- than for F_1F_2-based shapes, in that order. As for front vowels whose phonetic quality is determined primarily by their F_2 which themselves approach nearby higher formants presumed to be more speaker-specific, it is observed perhaps not surprisingly that F_1F_2-based shapes already embody as much speaker spread as F_3F_4-based shapes do. On the whole, however, there is a good indication that speaker individuality strengthens as VT-shapes are estimated from the higher formants.

5. CONCLUDING PERSPECTIVE

In sum, this study has provided some acoustic-articulatory evidence in support of the long-standing claim, that the upper formant region of steady-state vowels contains relatively more speaker-specific information than the lower F_1F_2-region, which itself is well-known to be predominantly a carrier of phonetic cues. Our results also indicate that speaker individuality can be expected to vary with place of articulation, with a strong F_2- and F_4-dependency for front vowels and a relatively stronger F_3- and F_4-dependency for back vowels. Notwithstanding these dependencies which have useful implications for automatic speech and speaker recognition, the higher frequency-components of vowel shapes were indeed found to cause more spread amongst speakers and, in this sense, can be held responsible for speaker-specific behaviours of the vocal tract. However, a more complete elucidation will require either direct measurements which are difficult to acquire extensively, or a more realistic shape estimator which can account for losses found in the human vocal tract.

6. REFERENCES

[1] Clermont, F. "Formant-Contour Models of Diphthongs: A Study in Acoustic-Phonetics and Computer Modelling of Speech", *Doctoral Thesis*, The Australian National University, Australia, 1991.

[2] Lewis, D. "Vocal Resonance", *J. Acoust. Soc. Am.* 8: 91-99, 1936.

[3] Liljencrants, J. "Fourier Series Description of the Tongue Profile", *Speech Transmission Laboratory, Quaterly Progress and Status Report* 4: 9-18, 1971.

[4] Mermelstein, P. "Determination of the Vocal-Tract Shape from Measured Formant Frequencies", *J. Acoust. Soc. Am.* 41: 1283-1294, 1967.

[5] Mokhtari, P. "An Acoustic-Phonetic and Articulatory Study of Speech-Speaker Dichotomy", *Doctoral Thesis*, The University of New South Wales, Australia, 1998.

[6] Paige, A. and Zue V.W. "Calculation of Vocal-Tract Length", *IEEE Trans. on Audio and Electroacoustics* 18: 268-270, 1970.

[7] Peterson, G.E. "The Acoustics of Speech - Part II: Acoustical Properties of Speech Waves", in Travis, L.E. (ed.), *Handbook of Speech Pathology* (Peter Owen, London): 137-173, 1959.

[8] Schroeder, M.R. "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements", *J. Acoust. Soc. Am.* 41: 1002-1010, 1967.

[9] Wakita, H. "Normalization of Vowels by Vocal-Tract Length and its Application to Vowel Identification", *IEEE Trans. on Acoustics, Speech and Signal Processing* 25: 183-192, 1977.