A METHODOLOGY FOR INVESTIGATING VOWEL-SPEAKER INTERACTIONS IN THE ACOUSTIC-PHONETIC DOMAIN

Parham Mokhtari and Frantz Clermont

School of Computer Science
University College, University of New South Wales
Australian Defence Force Academy

ABSTRACT - A long-standing problem in speech research is concerned with the separation of the phonetic and speaker-specific attributes of the acoustics of spoken language. In a previous attempt (Mokhtari & Clermont, 1994) to re-examine this problem in the context of machine classification of spoken vowels, we first confirmed the relative importance of the low spectral regions for maximum phonetic distinction. However, we also provided compelling evidence of the relatively large, speaker-related potency of the high spectral regions, where inter-speaker vowel distinction was found to be adversely affected. This contrast in classification accuracy observed across the available spectral range led us then to advance the notion of dichotomy which is unfolded by way of the methodology described in this paper. The consequences of the dichotomy are thus studied more closely, with a view to gaining a better understanding of the acoustic-phonetic basis for the detrimental effects of vowel-speaker interactions observed in the high spectral regions. The proposed methodology is also put forward as having the potential of paving the way for more robust speech or speaker recognition systems.

INTRODUCTION

An important body of knowledge concerning vowel sounds is rooted in the spectral domain, where consequences of phonetic and speaker variability have been studied primarily in terms of the first three resonance or formant frequencies (e.g. F_1, F_2, F_3) of the vocal tract. The seminal study of Peterson & Barney (1952), for example, has shown that steady-state vowels tend to be well-separated in the F_1-F_2 plane on an intra-speaker basis, but that very little additional, phonetic information is gained for a given speaker by also considering F_3. However, the finding that the higher resonance regions (e.g. F_3 and F_4) of spoken vowels are likely to contain relatively larger proportions of speaker variability appears to date back to Lewis & Tuthill's remarkable study of 1940. More recently, French vowel data were used in a speaker recognition paradigm by Mella (1994), who showed that F_3 contains much more speaker-discriminating properties than either F_1 or F_2 alone. Such findings have also been confirmed by perceptual tests carried out by Kuwabara & Takagi (1991), revealing that relative changes in F_3 adversely affect judgments of voice individuality to a greater degree than changes in F_1 or F_2. Working with more complete spectral information distributed over thirty-five frequency bands (270-10,000 Hz), Li & Hughes (1974) were able to increase inter-speaker distances by de-emphasising frequencies lower than 2200 Hz.

If the emerging understanding is correct that certain spectral regions of vowel sounds contain predominantly either phonetic or speaker-specific information, then it would seem possible to observe this phenomenon in its duality, if machine classification of these sounds were attempted on an intra- and an inter-speaker basis, respectively. Such experiments would also be expected to be more revealing if vowel-speaker interactions were investigated using a more complete representation of the spectral continuum such as the LP-cepstrum, which is not only easily extracted from the acoustic speech signal, but has also been shown to be very effective for both speaker (Furui, 1981) and speech (Paliwal & Rao, 1982) classification tasks. These motivations have collectively led to our previous endeavour (Mokhtari & Clermont, 1994) and to this sequel study, which provide strong quantitative evidence to support the view that the high spectral regions of vowel sounds contain more speaker-specific information. In particular, we advance the notion of spectral dichotomy to characterise the observed vowel-speaker interactions, together with a methodology for unfolding the dichotomy.

VOWEL-SPEAKER INTERACTIONS ACROSS THE ENTIRE SPECTRAL RANGE

We first investigated the problem of vowel-speaker interactions raised above, by studying accuracy profiles obtained from speaker-dependent and speaker-independent, vowel classification experiments. In the former, the emphasis is placed on the phonetic dimension of the problem, while the effects of vowel-speaker interactions are expected to be the strongest in the latter experiments.

Speech materials

The dataset (Clermont, 1991) used to conduct the experiments outlined above, comprises nine non-nasalised vowels in /CVd/ context, where C=/ h, b, d, g, p, t, k / and V=/ i, ɪ, ɛ, æ, a, ɒ, ɜː, ʌ, ɜ /. Five random repetitions of each /CVd/ monosyllable were recorded, in one session, by four adult male, native speakers of Australian English (hereafter referred to as speakers A, B, C, and D). The waveforms were sampled at f_s = 10 kHz and quantised to 12 bits. The acoustic parameters used for training and classification consist of 14th-order linear prediction (LP) cepstra of three consecutive frames chosen near the most stationary part of each vocalic nucleus. The first three formant frequencies (F_1, F_2, F_3) were also estimated at these frames, using a formant-tracking method (Clermont, 1991) based on spectral matching and dynamic programming.

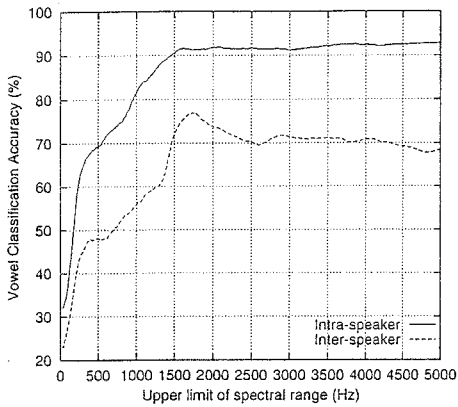

Figure 1. Intra- & inter-speaker vowel classification accuracy as a function of upper limit θ_2 of spectral range $[0,\theta_2]$.

Parametric cepstral distance

Distances based on the cepstrum have todate been formulated in such a way as to yield similarity measures, which are integrated over the entire spectral range defined between zero Hertz and half the sampling frequency, and therefore lack the ability to resolve fine spectral interactions between phonetic and speaker components. We overcame this limitation by using our new formulation (Clermont & Mokhtari, 1994) of the quefrency-weighted cepstral distance (Yegnanarayana & Reddy, 1979), which allows calculation of this distance to be confined within a selected region $[\theta_1,\theta_2]$ Hz of the full spectral range $[0,f_s/2]$ Hz, as shown by the following expression:

$$d^2 = \frac{1}{(\theta_2-\theta_1)}\int_{\theta_1}^{\theta_2}\left[(-\frac{d\phi(e^{j\theta})}{d\theta})-(-\frac{d\phi'(e^{j\theta})}{d\theta})\right]^2 d\theta = (\mathbf{c}-\mathbf{c}')^T\mathbf{W}(\mathbf{c}-\mathbf{c}') \quad ,$$

where $\phi(e^{j\theta})$ and $\phi'(e^{j\theta})$ are the LP-phase spectra of a test frame and of a vowel template, respectively, $\mathbf{c}$ and $\mathbf{c}'$ are the corresponding LP cepstral vectors, and $\mathbf{W}$ is a matrix the elements of which depend only on θ_1 and θ_2. For all experiments, we adopted the methodology of fixing the lower spectral limit to a constant θ_1=0 Hz and incrementing the upper limit θ_2 to the full range, by 20-Hz steps.

Behaviour of classification accuracy across the entire spectral range

The classifier chosen for our experiments is based on the well-known Nearest-Neighbour (k-NN, k=1) method of pattern comparison, in conjunction with the leave-one-out approach to data partitioning, also known as the U-method (Toussaint, 1974). Speaker-dependent results were obtained by training the classifier on four of the repetitions in turn, and then using the fifth repetition as test data. By contrast, speaker-independent results were obtained by training the classifier on three of the speakers at a time, and using all five repetitions of the remaining speaker as test data. The solid curve of classification accuracy, plotted in Figure 1 as a function of upper spectral limit θ_2, represents the mean of four separate intra-speaker experiments, while the dashed curve represents the mean of four separate inter-speaker experiments.

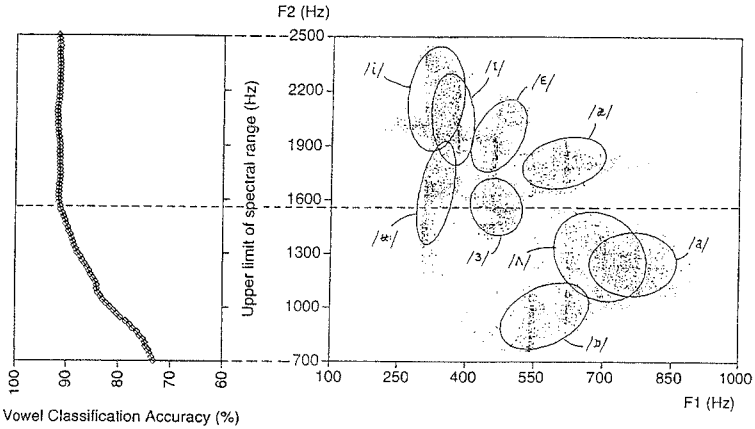

Figure 2. F_1-F_2 vowel space of all 4 male speakers (Australian English). Adjacent to the ordinate is plotted the portion of the intra-speaker accuracy curve (Figure 1) which spans the F_2 range.

The contrast between the behaviour of intra- and inter-speaker vowel classification accuracy as a function of an increasing upper spectral limit θ_2, as portrayed in Figure 1, does suggest a dichotomous relation between the consequences of vowel and speaker influences across the spectral continuum. An intriguing question then is whether these accuracy curves can be interpreted as embodying regions of primary phonetic and speaker influence corresponding to particular formant ranges.

SPECTRAL REGIONS OF PRIMARY PHONETIC INFLUENCE

It is quite clear in the literature that spectral regions which include the first two formants contribute relatively more to vowel discrimination on an intra-speaker basis. It follows that, from a vowel classification point of view, the level of accuracy expected in a speaker-dependent task would be largely independent of the so-called high spectral regions which include and extend beyond the third formant. The nearly asymptotic intra-speaker accuracy curve (solid line) in Figure 1 indeed confirms this implication.

However, a more informative perspective is offered in Figure 2 where, adjacent to the F_1-F_2 plane, is plotted the portion of the intra-speaker accuracy curve which encompasses the F_2 range of the four speakers' formant distribution. Classification accuracy rises to 75% as θ_2 is increased across the F_1 range, to the lowest F_2 (775 Hz) of the formant distribution, and continues to rise to 91% as the spectral range is further extended to the mid-F_2 (1560 Hz) of the speakers' vowel formant space. By contrast, the inclusion of spectral information contained in the regions beyond the mid-F_2 of the four speakers' vowel space, results in a nearly-asymptotic increase of accuracy to 92% at the highest F_2 (2460 Hz), and to 93% at full range (5000 Hz). This contrast does confirm the relatively more important role of the spectral regions encompassing the first two formants, for the purposes of vowel discrimination on a speaker-dependent basis.

SPECTRAL REGIONS OF PRIMARY SPEAKER INFLUENCE

Dare we now hope to observe a similar, nearly asymptotic behaviour in vowel classification accuracy on an inter-speaker basis? Our experimental result shown in Figure 1 (dashed curve) already refutes this contention, and clearly indicates that the consequences of vowel-speaker interactions can be to the detriment of vowel classification accuracy in the high spectral regions.

By analogy with Figure 2, a more enlightening perspective is gained by plotting the relevant portions of the inter-speaker accuracy curve adjacent to the four speakers' F_2-F_3 plane, as shown in Figure 3. Classification accuracy rises to a maximum of 77% as the spectral range is increased up to 1780 Hz, and subsequently drops markedly to a full-range (0-5 kHz) value of 68%. This dichotomous behaviour of classification accuracy across the spectral continuum, together with the nearly asymptotic performance exhibited earlier by the intra-speaker curve, not only adds quantitative support to the

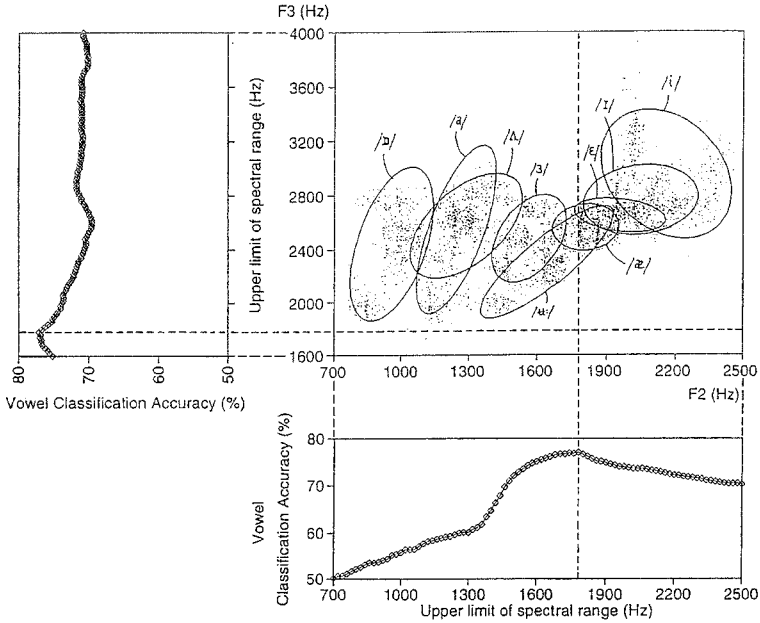

Figure 3. F_2-F_3 vowel space of all 4 male speakers (Australian English). Adjacent to the abscissa and the ordinate are plotted the portions of the inter-speaker accuracy curve (Figure 1) which span the F_2 and the F_3 ranges, respectively.

literature (Ainsworth & Foster, 1985; Furui & Akagi, 1985; Kitamura & Akagi, 1994) which suggests a relatively larger amount of speaker variability in the so-called high spectral regions, but also points to a methodology for examining in greater detail the vowel-speaker interactions which cause this phenomenon.

ACOUSTIC-PHONETIC EXPLANATION OF VOWEL-SPEAKER INTERACTIONS

Thus far, we have observed that vowel-speaker interactions are more strongly manifest in the spectral regions spanning the high-F_2 and the F_3 of the speakers' vowel formant distribution. Indeed, Figure 3 clearly indicates that inter-speaker classification accuracy begins to drop as the spectral range is extended to include the F_2 of the speakers' front vowels and the F_3 of their back vowels, the overall effect of which was identified earlier as the phenomenon of dichotomy. This apparent consistency between certain formant ranges and prominent features of the classification accuracy curves encourages a more detailed phonetic analysis of vowel misclassifications, with a view towards identifying the vocalic subspaces which have caused the dichotomy.

The method adopted to carry the phonetic analysis consists of decomposing the mean, inter-speaker accuracy curve, first in terms of each speaker's individual accuracy curve (obtained by the U-method), and then in terms of the constituent accuracy curves for the individual vowels of each speaker. This decomposition yields the per-speaker contribution of each vowel to the drop in classification accuracy across the high spectral regions, and thus permits an acoustic-phonetic interpretation of the misclassifications listed in Table 1. The entry in the i^{th} row and j^{th} column lists: (1) the vowel(s) confused with the i^{th} vowel of the j^{th} speaker, (2) the formant(s) responsible for the confusion, and (3) the degree of contribution to the dichotomy as measured by the percentage drop in accuracy (to the nearest 5%) in each individual curve.

Misclassifications are observed to arise in the high spectral regions, from confusions: (i) amongst the F_3 of front vowels, (ii) between the F_3 of back vowels and the F_2 of front vowels, (iii) amongst the F_3 of back vowels, and (iv) amongst the F_2 of front vowels, in this increasing order of detrimental effects.

Vowel	Speaker			
	A	B	C	D
/ i /	/ ʉː / (F_2&F_3, 30%)	/ ɪ / (F_2, F_3, 10%)	/ ɪ / (F_2, 35%)	**/ ɪ / (F_2, 90%)**
/ ɪ /	/ i / (F_2, 15%; F_3, 5%)	/ ɛ / (F_2,20%; F_3,10%)	/ ɛ / (F_2,20%; F_3,30%)	/ ɛ / (F_2, 10%)
/ ɛ /	**/ i / (F_2, 85%)**		/ ʉː / (F_2, 30%)	
/ æ /				
/ a /	/ ʌ / (F_3, 50%)		/ ʌ / (F_3, 30%)	/ ɒ / (F_3, 15%)
/ ɒ /		/ a / (F_3, 20%)		/ æ / (F_2&F_3, 20%)
/ ʉː /	/ ɜ / (F_2, F_3, 10%)	/ ɛ, æ / (F_3, 25%)		/ i, ɪ, ɜ / (F_2&F_3, 15%) / i, ɪ / (F_3, 10%)
/ ʌ /				
/ ɜ /		/ ʉː / (F_2, 15%) / æ, ɛ, ʉː / (F_3, 20%)		/ i, ʌ / (F_2&F_3, 10%)

Table 1. Acoustic-phonetic decomposition of the dichotomy in speaker-independent vowel classification behaviour (Figure 1), in terms of the vowel misclassifications that contribute to the drop in accuracy across the high spectral regions. 'F_2&F_3' indicates misclassifications caused by overlapping F_2 and F_3 ranges.

In particular, classification accuracy of the vowel / a / of speakers A and C drops by 50% and 30%, respectively, due to confusions with the neighbouring vowel / ʌ /, as the spectral range is extended to include F_3 of these vowels. The most significant contributions to the dichotomy occur themselves as the spectral range is increased to span the F_2 of all speakers' front vowels. Indeed, classification accuracy of the high front vowel / i / of speaker D drops by 90% due to confusions with the neighbouring vowel / ɪ /, and classification accuracy of the mid-front vowel / ɛ / of speaker A drops by 85% due to confusions with / i /.

The vowel confusions shown in Table 1 and discussed above, are a direct consequence of vowel-speaker interactions in the high spectral regions, and as such, they embody the specific types of speaker differences that are manifest in our spoken vowel data. In this context, it is pertinent to note that Sambur (1975) found the F_2 of front vowels and the F_3 of the back vowel / u / to be the most speaker-discriminating parameters for spoken vowels of American English. More recently, Mella's (1994) study of French spoken vowels has also shown that F_3 of / u / and F_2 of front vowels are the best formant parameters for speaker identification. Whilst these results were obtained under the paradigm of speaker identification, they support our own conclusions regarding the relative speaker-specificity of different regions of the acoustic-phonetic vowel space.

CONCLUDING DISCUSSION

The analysis of vowel-speaker interactions presented above is based only on four speakers who may be considered to form a heterogeneous set, owing to the non-asymptotic accuracy behaviour shown in Figure 1, and to the significant vowel overlap evident in the F_2-F_3 plane shown in Figure 3. There is no guarantee, however, that an equally striking contrast in the behaviour of inter-speaker classification accuracy across the spectral continuum, would be readily obtained for a speaker set which is larger and thus more likely to embody a lesser degree of heterogeneity. Nevertheless, the inter-speaker accuracy curve so obtained, could further be decomposed in order to identify the spectrally more dissimilar subset of speakers. This is exemplified here in Figure 4 by retaining our speakers' individual accuracy curves, and then ranking the former by the degree of dichotomy manifest in the latter.

One can observe, for example, that speaker B's accuracy curve (solid line) is distinctly less dichotomous than the others', and may therefore be considered to be spectrally the most homogeneous speaker in the group, from a vowel classification point of view. This contrast already suggests that the dichotomous behaviour identified earlier in Figure 1 could become blurred as a result of a relatively larger degree of homogeneity amongst the speakers considered. However, this approach to retrieving the dichotomy holds the potential of being able to either perform judicious speaker selection in training a vowel recognition system, or predict the effects of vowel-speaker interactions on the performance of an automatic speaker recognition system.

In sum, we have presented a methodology for investigating vowel-speaker interactions across the spectral continuum, based on intra- and inter-speaker vowel classification experiments performed as a function of an increasing upper spectral limit. Our speaker-dependent experiments have yielded results which confirm the relative importance of the spectral regions encompassing the first two formants, for the purposes of vowel discrimination. By contrast, the speaker-independent results have shown that the influence of speaker variability is most strongly manifest in the spectral regions encompassing the F_2 of the speakers' front vowels and the F_3 of their back vowels. Finally, a more detailed acoustic-phonetic decomposition of the inter-speaker results has shown that the F_2 of mid-to-high front vowels are the most susceptible to the speaker variability-induced vowel confusions that give rise to the observed drop in accuracy in the high spectral regions.

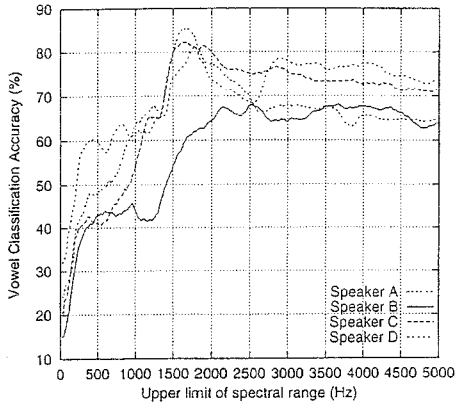

Figure 4. Decomposition of the inter-speaker vowel classification accuracy curve of Figure 1, in terms of the 4 speakers' individual contributions.

REFERENCES

Ainsworth, W.A. & Foster, H.M. (1985) "The Use of Dynamic Frequency Warping in a Speaker-Independent Vowel Classifier", in De Mori, R. & Suen, C.Y. (Eds.), *Proc. NATO Advanced Study Inst. on New Systems and Architectures for Automatic Speech Recognition and Synthesis*, 389-403.

Clermont, F. (1991) "Formant-contour models of diphthongs: A study in acoustic phonetics and computer modelling of speech", Doctoral Thesis, The Australian National University, Australia.

Clermont, F. & Mokhtari, P. (1994) "Frequency-band specification in cepstral distance computation", Proc. 5th Australian Int. Conf. on Speech Science and Technology, Perth, 354-359.

Furui, S. (1981) "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-29 (2), 254-272.

Furui, S. & Akagi, M. (1985) "Perception of voice individuality and physical correlates", J. Acoust. Soc. Japan (English reprint), H-85-18.

Kitamura, T. & Akagi, M. (1994) "Speaker individualities in speech spectral envelopes", Proc. 3rd Int. Conf. on Spoken Language Processing, Yokohama, Japan, 1183-1186.

Kuwabara, H. & Takagi, T. (1991) "Acoustical parameters of voice individuality and voice-quality control by analysis-synthesis method", Speech Communication 10 (5-6), 491-495.

Lewis, D. & Tuthill, C. (1940) "Resonant Frequencies and Damping Constants of Resonators Involved in the Production of Sustained Vowels "O" and "Ah" ", J. Acoust. Soc. Am. 11, 451-456.

Li, K.-P. & Hughes, G.W. (1974) "Talker differences as they appear in correlation matrices of continuous speech spectra", J. Acoust. Soc. Am. 55 (4), 833-837.

Mella, O. (1994) "Extraction of formants of oral vowels and critical analysis for speaker characterization", ESCA Workshop on Automatic Speaker Recog., Identif. & Verif., Suisse, 193-196.

Mokhtari, P. & Clermont, F. (1994) "Contributions of selected spectral regions to vowel classification accuracy", Proc. 3rd Int. Conf. on Spoken Language Processing, Yokohama, Japan, 1923-1926.

Paliwal, K.K. & Rao, P.V.S. (1982) "Evaluation of Various Linear Prediction Parametric Representations in Vowel Recognition", Signal Processing 4, 323-327.

Peterson, G.E. & Barney, H.L. (1952) "Control Methods Used in a Study of the Vowels", J. Acoust. Soc. Am. 24 (2), 175-184.

Sambur, R. (1975) "Selection of Acoustic Features for Speaker Identification", IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-23 (2), 176-182.

Toussaint, G.T. (1974) "Bibliography on Estimation of Misclassification", IEEE Transactions on Information Theory, IT-20 (4), 472-479.

Yegnanarayana, B. & Reddy, D.R. (1979) "A Distance Measure Based on the Derivative of Linear Prediction Phase Spectrum", Proc. Int. Conf. on Acoustics, Speech and Signal Processing, 744-747.