MODELLING INDIVIDUAL SPEAKER CHARACTERISTICS BY DESCRIBING A SPEAKER'S VOWEL DISTRIBUTION IN ARTICULATORY, CEPSTRAL AND FORMANT SPACE

Simon Hawkins, Iain MacLeod, and Bruce Millar

Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Australian National University

ABSTRACT - Of the various methods of encoding vowels, the cepstral representation has consistently produced the best performance in automatic vowel classification studies. We explain this robust phenomenon in terms of our finding that the shape of a speaker's vowel surface is much more similar across speakers in Cepstral Space than it is in either Articulatory or Formant Space. The cepstral representation thus allows an automatic vowel classifier trained on the vowels of one group of speakers to generalise to the vowels of another group of speakers.

INTRODUCTION

As discussed by Waibel and Lee (1990), the first question that arises during the design of a speech recognition system is how to represent or encode the speech signal itself before recognition is attempted. In principle, one could simply use the digitised waveform as the input signal. However, at sampling rates of 10,000 samples per second or more, the amount of processing required would be prohibitive. The waveform also contains information that is redundant for speech processing, such as phase information, which can be detrimental to recognition performance (Waibel and Lee, 1990).

Various encoding schemes have therefore been developed that attempt to provide a compact representation of speech while preserving and enhancing the perceptually relevant cues in speech. Davis and Mermelstein (1980) compared a number of encoding techniques on the basis of their speech recognition performance using a template matching technique. They found that variants of cepstral analysis were the most effective techniques for representing vowels.

More recently, Zahorian and Jagharghi (1993) used a Bayesian classifier and a fully interconnected MLP neural network to classify monophthongal vowels spoken in the context of isolated CVC words under a variety of conditions. The steady state portions of the vowels were represented in terms of either the first three formant frequencies or the coefficients in a cosine expansion of the nonlinearly scaled log magnitude spectrum (the Discrete Cosine Transformation Coefficients or DCTCs). Except for small differences, the authors found that these coefficients were the same as the set of cepstral coefficients. The representation of vowels in terms of ten or more DCTCs was found to result in more accurate automatic vowel classification than the representation of vowels in terms of the first three formant frequencies. However, there was no subset of three DCTCs that provided as much discriminatory power as did the three formant frequencies.

The aim of this paper is to explain why features which encode the smoothed spectra of vowels, such as the set of low-order LPC cepstral coefficients, have been found empirically to provide the best vowel discrimination in automatic vowel classification experiments. We propose that the best encoding strategy will be one that generates a vowel distribution that is highly similar or invariant in shape across speakers. This will allow a classification system trained using the vowels from one group of speakers to generalise to the vowels from a new group of speakers. To test this hypothesis, we compared the structure of the distribution formed by a speaker's vowel nuclei in three representational domains: the formant domain, the cepstral domain and the articulatory domain. The structure of a speaker's vowel distribution will be determined by the articulatory and perceptual constraints that operate during vowel production and can be captured in terms of the (1) the intrinsic dimensionality of the distribution as well as (2) the shape of the surface formed by the distribution.

As well as calculating the shape of the vowel surface, we are also interested in the phonetic quality of the vowel spectra that are separated by the axis of symmetry of the best fitting vowel surface. Broad (Broad, 1981; Broad and Wakita, 1977) observed that the distribution of American English monophthongs produced by an individual speaker in Formant Space formed a piecewise-planar surface. The speaker's front vowel targets were observed to lie on one plane and the back vowel targets on another plane. Broad and Clermont (1989) found that the Australian English monophthongs could be mapped from 12D cepstral space to 3D formant space using a single linear mapping. This finding implies that a speaker's monophthongs will form a piecewise-planar surface in 12D cepstral space just as they have been observed to do in 3D formant space.

In the present study, we use speech frames taken from the entire vocalic nucleii of male speaker's' monophthongs and diphthongs. We are therefore sampling the entire range of vocalic sounds that can be produced by the speaker rather than using just the targets of the monophthongal vowels. We will therefore use a continuous quadratic surface to model a speaker's vocalic surface rather than a pair of discontinuous planes. We hypothesise that the quadratic surface that best fits the vowel distribution of an individual speaker in cepstral space will form a parabolic surface. We predict that this parabolic surface will have an axis of symmetry that separates the speaker's front and central vowels from his back vowels. Because the cepstral domain provides the best method of encoding vowel in vowel classification studies, we predict that the shape of the vowel surface will be more consistent across speakers in cepstral space that in other representational domains.

SPEECH DATA

A vowel distribution was derived from the speech of 12 Australian male speakers recorded by Millar et al (1989). A speaker's vowel distribution is defined to be the set of speech frames taken from the vocalic nuclei of the eleven monophthongs and eight diphthongs in Australian English as uttered by the speakers in [h-vowel-d] context. Speech frames were taken from entire vocalic nucleus of each vowel and not from its steady-state region or from hypothesised vowel targets. The boundaries of the vocalic nucleus were obtained by finding the maximum number of speech frames that had a clear-cut formant structure.

Vowel spectra from the speech frames of the 19 vocalic nuclei produced by a speaker were represented as a distribution of points in three different representational domains. In the articulatory domain, vowel spectra were represented in terms of the estimated area function of the supralaryngeal vocal tract derived from the vowel spectrum. The area function comprised 12 abutting sections of equal length but variable radius. This area function was estimated from the set of 12 LPC reflection coefficients using Wakita's (1977) algorithm. The set of speech frames that formed a speaker's vowel distribution may therefore be represented as a distribution of points in the 12 dimensional space formed by the radii of these 12 sections.

In the formant domain, vowels were represented in terms of the lowest three formant resonances of the vocal tract. Estimates of the first three formant frequencies (in Hertz) were obtained using Talkin's (1987) analysis-by-synthesis method as implemented and modified by Clermont (1991). A speaker's vowel distribution may be represented as a distribution of points within the three dimensional space formed by the first three formants.

In cepstral space, vowels were represented by the first 12 low-order LPC cepstral coefficients after the zeroth order cepstral coefficient had been discarded. A speaker's speech frames may be represented as a distribution of points within the 12 dimensional space formed by these first 12 LPC cepstral coefficients.

INTRINSIC DIMENSIONALITY OF A SPEAKER'S VOWEL DISTRIBUTION

We were interested in the **intrinsic dimensionality** and **shape** of an individual speaker's vowel distribution in the cepstral domain, the formant domain, and the articulatory domain. In the Cepstral and Articulatory domains, both of which are 12 dimensional, we applied Principal Components Analysis (PCA) to find the intrinsic dimensionality of a speaker's vowel distribution. Applied to vowel spectra, PCA is a dimensionality-reduction technique which seeks to find new directions through the original space that capture as much of the original variance of the vowel distribution as possible. These new directions are linear combinations of the original dimensions that lie orthogonal to one another .

On average across the 12 speakers, we found that just three Principal Components were required to explain about 80% of the total variance of a speaker's vowel distribution. The inclusion of a fourth Principal Component to represent a speaker's vowel distribution did not greatly increase the percentage of total explained variance. This suggests that the **intrinsic dimensionality** of both Articulatory Space and Cepstral Space is **three**. In other words, the variance of a speaker's vowel distribution in both Articulatory and Cepstral Space can be characterised in terms of three new orthogonal directions within the original 12 dimensional space. Formant Space is a three-dimensional space formed by the first three formants. As long as a speaker's vowel distribution in this space does not form a plane, the intrinsic dimensionality of Formant Space is **three.**

In other words, the intrinsic dimensionality of a speaker's vowel distribution would appear to be three regardless of the domain in which this distribution is represented. This result is in agreement with the earlier work of Pols et al (1969) and makes it possible to replot a speaker's vowel distribution within the three dimensional space formed by the first three Principal Components of Articulatory Space and the first three Principal Components of Cepstral Space with little loss of information.

SHAPE OF A SPEAKER'S VOWEL DISTRIBUTION

To find the shape of the vowel distribution in a 3D subspace formed by the first three Principal Components (which we labelled X, Y, and Z), we used the least squares criterion to fit a general quadratic equation of the form,

$$\hat{Z} = aX^2 + bX + c + dY^2 + eY + fXY \qquad \text{(Eqn. 1)}$$

to the distribution of a speaker's vowels within the 3D subspace formed by the principal components X, Y, and Z. We found that the vowel distribution of most speakers formed a parabolic surface with an axis of symmetry that separated the speaker's front and back vowels. To simplify the equation required to represent this surface, the three dimensional subspace formed by the three Principal Components was rotated about the speaker's vowel distribution. Two rotations were used to make the axis of symmetry of the best fitting quadratic surface lie perpendicular to the First Principal Component (X). First, the plane formed by the first and third Principal Components (X and Z) was rotated about the axis formed by the third principal component Y to maximise the linear correlation,

$$r(Z', \hat{Z}' = aX'^2 + bX' + c) \qquad \text{(Eqn. 2)}$$

The plane formed by the second and third principal components (Y and Z) was then rotated about the axis formed by the first principal component X to further improve the correlation given by Equation (2). A third rotation was then performed to eliminate the cross-product term XY in the best fitting quadratic equation. After reorienting the 3-dimensional Principal Component space surface using these three rotations, the three axes of this rotated subspace were standardised to have a mean of zero and a SD of one. A general quadratic equation of the form specified by equation (1) was refitted to the vowel distribution and the main axis of symmetry was calculated using the equation,

$$X' = -b/2a. \qquad \text{(Eqn. 3)}$$

where X was the axis formed by the First Principal Component after the three rotations had been performed.

We were also interested in the extent to which this axis of symmetry separated the speaker's front and back vowels. Because we are dealing with speech frames from the entire vocalic nucleus rather than vowel targets, we devised a method for identifying whether the spectrum of a vowel has the timbre of a front vowel or the timbre of a back vowel. For each speaker, we drew a line on the speaker's F1-F2 plane that joined the midpoint of the speaker's vowels [æ] and [a] to the midpoint of his vowels [U] and [u]. We used the average F1 and F2 frequencies of these vowels, averaged over the entire vocalic nucleus of the vowel, to locate these vowels in the F1-F2 plane. Vowel segments lying above the line in the speaker's F1-F2 plane were considered to belong to the phonetic class of the front or vowels and vowel segments lying below this line to belong to the phonetic class of the back vowels. We then calculated the percentage of a speaker's front and back vowel segments that lay to each side of the axis of symmetry axis of the best fitting quadratic vowel surface in each of the three domains.

RESULTS

The vowel distribution in each of the three domains was adequately described by a quadratic vowel surface. In the subspace formed by the three principal components of articulatory space, there was an average correlation across the 12 speakers of 0.8287 between the best fitting quadratic surface and the speaker's actual vowel distribution. The average correlation across speakers was 0.756 in the subspace formed by the three principal component of Cepstral Space and 0.601 in formant space.

By rotating the vowel distribution within these 3D spaces, we found that the best fitting vowel surface could be described by a quadratic equation that had just **four** terms,

$$\hat{Z}' = aX'^2 + bX' + c + eY' \text{ in Articulatory and Cepstral Space,}$$

$$\hat{F}_3 = aF_2^2 + bF_2 + c + eF_1 \text{ in Formant Space}$$

Other terms in the best fitting quadratic equation could be discarded because they were associated with coefficients that had negligible magnitude.

Formant Space. The best fitting quadratic vowel surfaces in formant space are pictured for each speaker in Figure 1. Only a subset of speakers (Speakers 01,04,11,12,24,34) had a vowel surface that was parabolic in shape. The vowel surfaces of other speakers were saddle-shaped (Speakers 30 and 36), bowel-shaped (Speaker 13) and convex rather than concave (Speakers 20 and 35). Nevertheless, these vowel surfaces usually had an axis of symmetry that provided reasonable separation between the speaker's front and back vowels. On average across the 12 speakers, the axis of symmetry of the best fitting quadratic surface separated 97.3% of a speaker's front vowels from 75.3% of his back vowels.

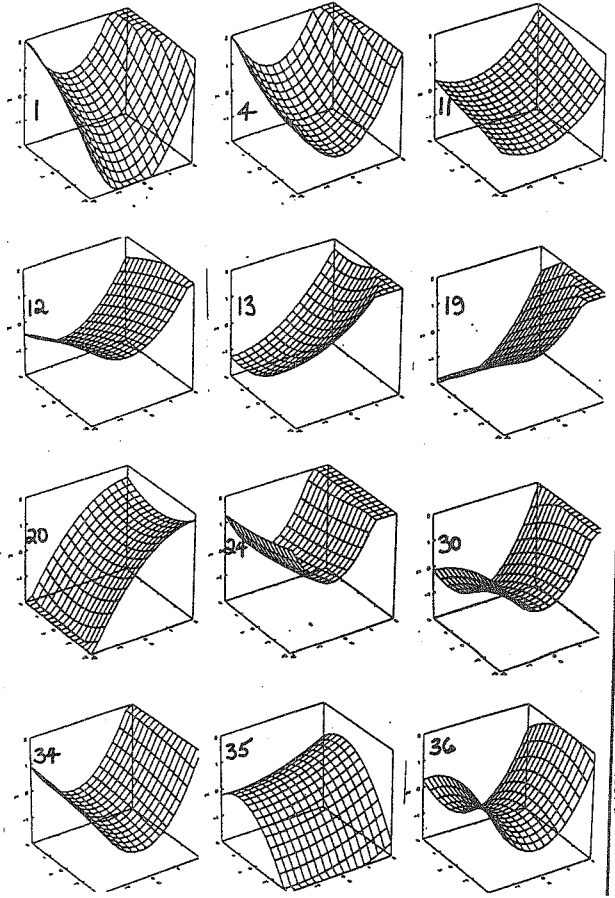

FIG. 1. Vowel surfaces of 12 speakers in Formant space

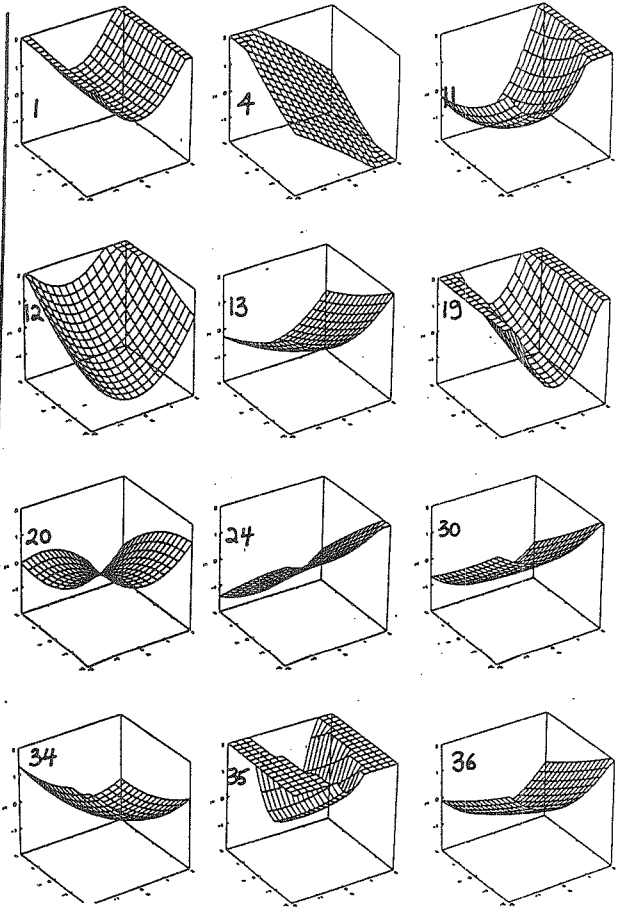

FIG. 2. Vowel surfaces of 12 speakers in Articulatory Space

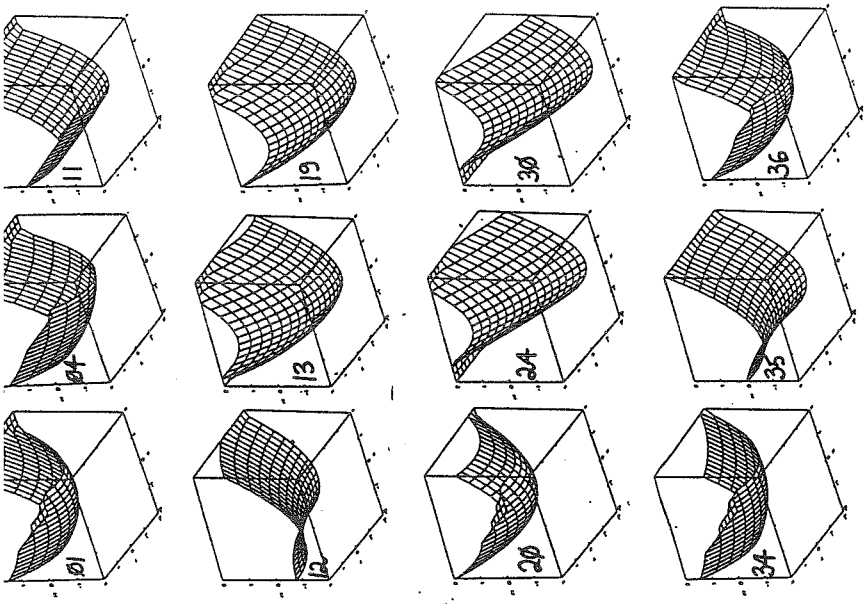

FIG 3. Vowel surface of 12 speakers in Cepstral space.

Articulatory Space. On average across the 12 speakers, the first three Principal Components explained 93.6% of the total variance of a speaker's vowel distribution in 12D Articulatory Space. The best fitting quadratic vowel surfaces in the subspace formed by these three principal components after rotation are pictured for each speaker in Figure 2. The shape of the vowel surface varied considerably across speakers. Although most speakers had parabolic-shaped vowel surfaces (Speakers 01, 12, 13, 19, 24, 30, 34, 35, 36), one had a planar vowel surface (Speaker 04), another had a saddle-shaped surface (Speaker 20), and another had a slightly bowl-shaped surface (Speaker 11). The vowel surfaces of some speakers were highly curved about the axis of symmetry (Speaker 19) whereas the surfaces of others had little curvature (Speakers 35 and 36).

Despite this variability in the shape of the vowel surface, the vowel surface of most speakers had an axis of symmetry that separated the speaker's front vowels from his back vowels. The exceptions were Speaker 04 and Speaker 11. On average across the 12 speakers, the axis of symmetry of the best fitting quadratic vowel surface separated 82.2% of a speaker's front vowels from 76.5% of his back vowels.

Cepstral Space. On average across the 12 speakers, the first three Principal Components explained 77.3% of the total variance of a speaker's vowel distribution in 12D Cepstral Space. The best fitting quadratic vowel surfaces in the subspace formed by the rotated principal components are pictured for each speaker in Figure 3. The vowel surface is highly similar in shape across the 12 speakers being parabolic for all but speaker 12. The vowel surface of this speaker was parabolic but slightly saddle-shaped. The vowel surfaces of all speakers had an axis of symmetry that clearly separated the speaker's front and back vowels. On average across the 12 speakers, the axis of symmetry of the best fitting vowel surface separated 92% of the front vowels from 86% of his back vowels.

CONCLUSION

The surface formed by a speaker's vowel distribution was well described by a quadratic equation with just four terms. This means that just four parameters are required to accurately represent the structure of a speaker's vowel distribution.

The shape of the best fitting quadratic vowel surface was much more similar across speakers in cepstral space than was the case in either formant space or articulatory space. In cepstral space, the vowel surface consistently formed a parabolic surface with an axis of symmetry that separated the speaker's front and back vowels. In the formant and articulatory domains, the axis of symmetry of the best fitting quadratic surface provided good separation between a speaker's front and back vowels but the vowel surface was not consistently parabolic in shape.

This result helps to explain why automatic vowel classifiers perform best when vowels are represented in terms of cepstral coefficients. The consistent shape of the vowel surface in Cepstral Space across speakers means that an automatic vowel classifier trained on vowel distributions of one group of speakers will be able to generalise to the vowel distributions of other speakers.

REFERENCES

Broad, D.J. & Clermont, F. (1989). "Formant estimation by linear estimation of the LPC cepstrum", J. Acoust. Soc. Am., 86, 2013-2017.

Broad, D.J. & Wakita, H. (1977). Piecewise-planar representation of formant vowel frequencies. J. Acoust. Soc. Am., 62, 1467-1473.

Broad, D.J..(1981). Piecewise-planar vowel formant frequencies across speakers, J. Acoust. Soc. Am., 69, 1423-1429.

Clermont, F. (1991). "Formant-contour models of diphthongs: a study in acoustic phonetics and computer modelling in speech", Unpublished PhD thesis, Australian National University.

Davis, S.B. and Mermelstein, P. (1980). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. ASSP-28, 357-366.

Hawkins, S. & Clermont, F. (1990). "Supervised Ceptrum-to-Formant estimation: a new piecewise-linear model", Proc. Int. Conf. on Speech, Science, and Technology, Melbourne, 310-315.

Millar, J.B., O'Kane, M. and Bryant, P. (1989). "Design, collection and description of a database of spoken Australian English", Australian Journal of Linguistics, 9, 165-189.

Pols. L.C.W., Van Der Kamp, L.J. ,& Plomp, R (1969). "Perceptual and physical space of vowel sounds", Journal of Acoustical Society of America, 2, 458- 467.

Talkin, D. (1987). Speech formant trajectory estimation using dynamic programming with modulated transition costs, J. Acoustical Society of America, 82, S-55.

Wakita, H. (1977). "Direct estimation of the vocal tract shape by inverse filtering of acoustic waveforms", IEEE Trans. on Audio and Electroacoustics, Vol AU-21, No.5., 417-427.

Waibel, A. and Lee, K-F. (1990), Readings in Speech Recognition, Morgan-Kaufmann, San Mateo, California.

Zahorian, S.A. & Jagharghi, A.J. (1993). "Spectral-shape features versus formants as acosutic correlates for vowels", J. Acoust. Soc. Am., 94 (4), 1966-1981.