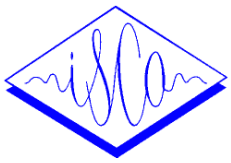

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

CONTRIBUTIONS OF SELECTED SPECTRAL REGIONS TO VOWEL CLASSIFICATION ACCURACY

Parham Mokhtari and *Frantz Clermont*

Department of Computer Science
University College, University of New South Wales
Australian Defence Force Academy
Canberra ACT 2601, AUSTRALIA

ABSTRACT

In this paper we describe the results of several vowel classification experiments, which shed light on the relative importance of certain spectral regions for achieving optimum classification accuracy in either a speaker-dependent or a speaker-independent task. The methodology adopted to conduct this investigation is based on a new formulation [3] of the quefrency-weighted cepstral distance measure, which allows specification of any frequency band within the available spectral range. This more flexible approach to distance computation was used to study the behaviour of classification accuracy as a function of increasing spectral range, and to identify spectral regions which predominantly carry either phonetic or speaker-specific information.

1. INTRODUCTION

Progress towards robust and unconstrained speech recognition has been partly impeded by our incomplete understanding of the interaction between phonetic and speaker-specific properties of the acoustic speech signal. Although this interaction is relatively better understood for the spoken vowels, machine classification of these sounds may still todate be expected to yield worse results in a speaker-independent task than in a speaker-dependent one. This contrast does suggest that the level of classification accuracy will partly depend on the degree of phonetic and speaker separability that is inherent to the given dataset. An interesting question then arises whether certain acoustic properties of speech sounds can readily be related to the phonetic and speaker factors, and be brought to bear on the behaviour of classification accuracy. This paper aims to shed light on this question for spoken vowels in particular.

An important body of knowledge concerning vowel sounds is rooted in the spectral domain, where the consequences of phonetic and speaker variability have been studied primarily in terms of such discrete representations as the resonance or formant frequencies (e.g. F1, F2, F3) of the vocal tract. The seminal studies of Peterson & Barney [11] [12] and of Pols et al. [13], for example, have shown that steady-state vowels tend to be well-separated in the F1-F2 plane on an intra-speaker or a speaker-normalised basis. It also emerges from these studies that, while certain inter-speaker differences can be highlighted in the F1-F2 vowel plane, very little additional, phonetic information is gained for a given speaker by also considering F3. The relative invariance of this formant across vowels was also noted by Stevens [14], who further suggested that the relatively fixed higher formants observed for some speakers may well be related to fixed anatomical structures.

However, the implicit finding that the higher resonance regions (e.g. F3 and F4) of spoken vowels are likely to contain relatively larger proportions of speaker variability dates back to Lewis & Tuthill's [7] remarkable study of 1940. Using very laborious techniques for estimating resonance frequencies of sung vowels, these authors were able to conclude that the first two formants are the strongest determinants of vowel character, while the third and higher resonances "probably contribute only to less basic tonal qualities" [7, p.456]. More recently, French vowel data were used in a speaker recognition paradigm by Mella [9], who showed that F3 contains much more speaker-discriminating properties than either F1 or F2 alone. Such findings have been confirmed by perceptual tests carried out by Kuwabara & Takagi [6], revealing that relative changes in F3 adversely affect voice individuality judgments to a greater degree than changes in F1 or F2. Working with more complete spectral information distributed over 35 frequency bands (270-10,000 Hz), Li & Hughes [8] were able to increase inter-speaker distances by de-emphasising frequencies lower than 2200 Hz. In a similar vein, Hayakawa & Itakura's [5] recent study of time-averaged spectra of Japanese words, yielded substantially high F-ratios (inter- to intra-speaker variances) in spectral regions extending to 16 kHz.

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 appropriate to study implications of this phenomenon for vowel classification based on a more complete representation of the spectral continuum such as the LP-cepstrum. This parameter is not only easily extracted from the acoustic speech signal, but it has also been shown to perform admirably in speech [10] as well as speaker [4] classification tasks. However, distances based on the cepstrum have thus far operated over the entire available spectral range, and therefore lack the ability to resolve fine spectral interaction between phonetic and speaker components. We overcame this limitation in our speaker-dependent and speaker-independent, vowel classification experiments by using a cepstral distance formulation [3], which allows specification of any frequency band within the available spectral range. As a result, we were able to observe, on an intra- and an inter-speaker basis, the behaviour of vowel classification accuracy across a spectral continuum, and to measure vowel and speaker variance within spectral regions encompassing familiar formant ranges.

In section 2 we describe vowel and speaker materials used for this study, and, in section 3, we present and discuss the results of our vowel classification experiments.

2. SPEECH MATERIAL & EXPERIMENTAL METHODS

We investigated the question of vowel-speaker interaction raised above, together with its consequences on vowel classification accuracy across the available spectral range, in the context of two datasets of spoken English vowels.

The dataset [2] used for our speaker-dependent experiments emphasises the phonetic dimension with several repetitions of vowels recorded by a small speaker population. It comprises nine non-nasalised vowels in /CVd/ context, where C=/h,b,d,g,p,t,k/ and V is as in "heed", "hid", "head", "had", "hard", "hod", "who'd", "hudd", and "heard". Five random repetitions of each /CVd/ monosyllable were recorded, in one session, by four adult male, native speakers of Australian English. The waveforms were quantised to 12 bits and sampled at f_s = 10 kHz. The acoustic parameters used for training and classification consisted of 14th-order LP cepstra of three consecutive frames chosen near the most stationary part of each vocalic nucleus. The first three formant frequencies were also estimated at these frames, using a formant-tracking technique [2] based on spectral matching and dynamic programming.

Speaker-independent experiments were also conducted using the 4-speaker Australian English data described above, which yielded preliminary insights into the behaviour of classification accuracy on an inter-speaker basis. We then sought to generalise our speaker-independent results by performing classification experiments based on a dataset of vowels produced by a much larger population of speakers. For this purpose, a subset of Peterson & Barney's [11] data were used, which include the first three formant frequencies of two repetitions of ten American English vowels recorded by 33 male speakers, in the context of the /hVd/ monosyllabic words "heed", "hid", "head", "had", "hod", "hawed", "hood", "who'd", "hud", and "heard". LP-cepstra of order 14 were recursively computed, assuming constant bandwidths of 50, 65 and 120 Hz, respectively, from the first three formants measured by Peterson & Barney.

The vowel classifier itself employed the well-known Nearest-Neighbour (k-NN, k=1) method of pattern classification, in conjunction with the leave-one-out approach to data partitioning, also known as the U-method [15]. Pattern matching was performed using our new formulation [3] of the quefrency-weighted cepstral distance [17], 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. For all vowel classification 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 up to the full range, by 20-Hz steps.

3. RESULTS & DISCUSSION

In sections 3.1 and 3.2, we describe the behaviour of vowel classification accuracy obtained as a function of upper spectral limit, with the aim of unveiling interactions between phonetic and speaker components in the spectral domain. In sections 3.2.1 and 3.2.2, we offer an explanation for the observed vowel-speaker dichotomy in terms of the relative amount of speaker variance dominating the high spectral regions.

3.1 Asymptotic Behaviour in High Spectral Regions

It is quite clear from 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.

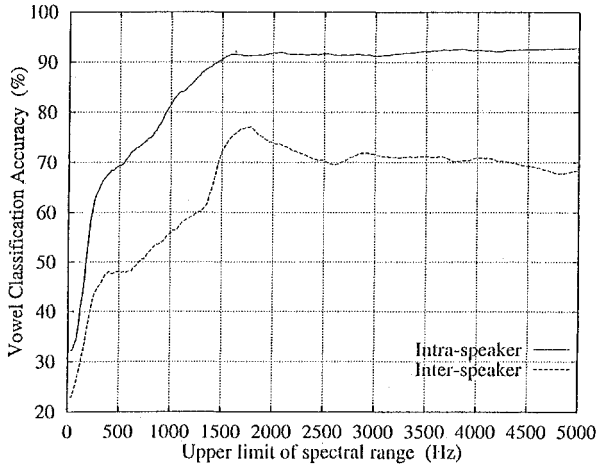

Figure 1. Intra- and inter-speaker vowel classification accuracy as a function of upper limit θ_2 of spectral range $[0,\theta_2]$. Dataset: 5 repetitions of 9 vowels spoken by 4 adult, male speakers of Australian English.

Our experimental result examining this implication is the curve of classification accuracy plotted in Figure 1 (solid line) as a function of upper spectral limit θ_2. This curve represents the mean of four separate speaker-dependent classification experiments based on the Australian English vowel data. Classification accuracy rises to 75% as θ_2 is increased from 20 Hz to the lowest F2 (775 Hz) of the 4 speakers' vowel formant space. A further rise to 91% is observed when the spectral range is extended to the lowest F3 (1746 Hz) of that space. By contrast, the inclusion of spectral information contained in the F3 and higher regions results in a nearly-asymptotic increase of accuracy to 93% at full range. This apparent invariance of classification accuracy beyond a certain spectral limit could almost be predicted from Ainsworth & Foster's [1] speaker-dependent, vowel classification experiments based on 128-point LP-spectra, which resulted in a drop of only 3% after the upper spectral limit was reduced from 5 kHz to 3.2 kHz.

3.2 Non-Asymptotic Behaviour in High Spectral Regions

Thus far, we have observed, on an intra-speaker basis, a nearly asymptotic behaviour of vowel classification accuracy in the high spectral regions. In view of the strong evidence from the literature that suggests a greater proportion of speaker influences in the high rather than in the low spectral regions, an intriguing question arises whether a similar asymptotic behaviour may be expected on an inter-speaker basis. To answer this question, we first conducted a speaker-independent experiment based on the same Australian English vowel data.

3.2.1 Clear Dichotomy. In contrast to the intra-speaker results obtained earlier, speaker-independent vowel classification accuracy as a function of increasing spectral range yielded a clearly non-asymptotic behaviour. The dashed

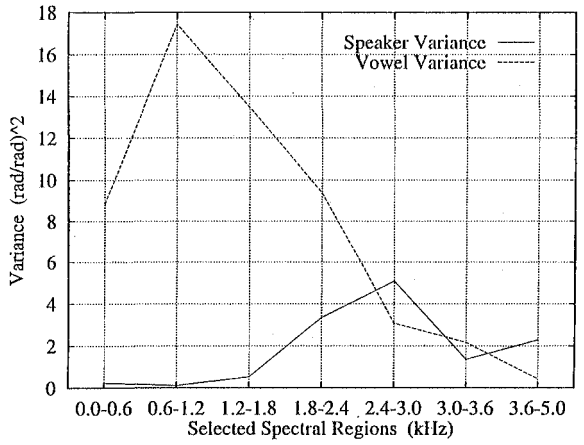

Figure 2. Profile of speaker and vowel variance in seven contiguous spectral regions. Dataset: 5 repetitions of 9 vowels spoken by 4 adult, male speakers of Australian English.

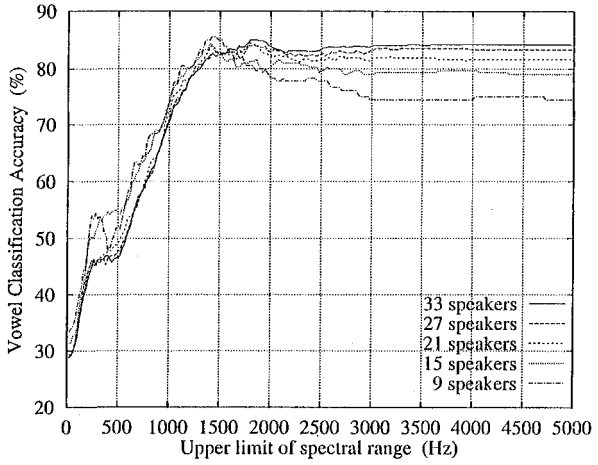

Figure 3. Dichotomy in vowel classification behaviour as a function of the proportion of 'goat' to 'sheep' speakers. Dataset: 2 repetitions of 10 vowels spoken by 33 adult, male speakers of American English; Peterson & Barney [11].

curve in Figure 1 shows that 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 asymptotic performance exhibited earlier by the intra-speaker curve, suggests a relatively larger amount of speaker variability in the so-called high spectral regions, which include and extend beyond the third formant. This conclusion may almost be inferred from Ainsworth & Foster's [1] inter-speaker vowel classification experiments, which showed that accuracy generally increased by reducing the upper spectral limit from 5 kHz to 3.2 kHz.

If the dichotomy observed above is indeed a result of inter-speaker differences causing confusion in vowel classification, then one would expect a large concentration of speaker variance in those spectral regions where vowel classification accuracy has been observed to decrease. The methodology adopted to decompose the vowel and speaker contributions to the total variance is similar to that which Pols et al. [13] and van Nierop et al. [16] used in the frequency analysis of Dutch vowels. However, our variance computation is based on a cepstral distance formulation [3] which allows specification of arbitrary frequency bands.

Figure 2 shows the amount of speaker (solid line) and vowel (dashed line) variance calculated over seven contiguous spectral regions of our Australian English data. These regions were chosen to span, respectively, low-F1, high-F1 to low-F2, mid-F2, high-F2 to low-F3, mid-F3, high-F3, and F3+ ranges of the four speakers' formant space. One can observe that vowel variance is greatest across the entire F1-F2 range (0-2.4 kHz), then decreases in all higher spectral regions. In contrast, the largest amount of speaker variance appears to be contained in the high-F2 (1.8-2.4 kHz) to mid-F3 (2.4-3.0 kHz) regions.

Our profile of vowel and speaker variance indicates that the dichotomy observed earlier is caused by a concentration of speaker variance in the spectral regions, which include and extend beyond the third formant. Indeed, the turning point of the inter-speaker accuracy curve shown in Figure 1 is found to occur at the very spectral region, in Figure 2, where speaker variance begins to rise and vowel variance begins to drop.

3.2.2 Blurred Dichotomy. The clearly dichotomous, accuracy curve obtained earlier across the spectral continuum was generated using a population of speakers, who differ quite significantly in the spectral representation of their vowels. Indeed, the analysis of spectral variance has revealed significant differences amongst our 4 Australian English speakers, in the high spectral regions. Should we now hope to observe a similar dichotomous behaviour in inter-speaker classification of vowels spoken by a more densely populated, and perhaps more homogeneous speaker group?

To answer this question, we attempted to duplicate our results using the 33 males of the Peterson & Barney data. The solid curve in Figure 3, representing speaker-independent vowel classification accuracy as a function of upper spectral limit, exhibits a maximum of 85.0% by including spectral information up to 1840 Hz, then drops to 82.9% for the spectral range extending to 2160 Hz. Although this small dip is reminiscent of our earlier inter-speaker classification accuracy curve, the drop of only 0.8% from the peak (1840 Hz) to full range (5 kHz) suggests that a potentially dichotomous behaviour has been blurred as a result of averaging across 33 speakers.

If the spectral dichotomy observed in Figure 1 is indeed an intrinsic phenomenon, then one could speculate that the blurring effect illustrated in Figure 3 has occurred as a result of an imbalance in the speaker population, and that any significant amount of spectral variability is due to only a small subset of the 33 speakers, the 'goats' as might be referred to. The clearly dichotomous, accuracy curve obtained for our 4 very dissimilar Australian English speakers may therefore be used as typical evidence for identifying the so-called 'goat' speakers. In other words, those individual classification accuracy curves, obtained by the U-method, which show strong evidence of a spectral dichotomy, would characterise the 'goats' of the 33 speakers, while more nearly asymptotic curves would

characterise the more spectrally homogeneous 'sheep'. Applying this criterion to the 33 individual accuracy curves, we were able to rank-order the male speakers in terms of their degree of 'sheepiness' or 'goatiness'.

In Figure 3 we have superimposed the curves of inter-speaker vowel classification accuracy obtained after removing, six at a time, the 'sheepiest' of the speakers, according to our rank-ordered list. The gradual trend towards a non-asymptotic shape first confirms our earlier findings, and, more importantly, indicates that a speaker population dominated by 'goats' will tend to yield a more pronounced spectral dichotomy, by virtue of the greater proportion of inter-speaker variability that exists in the so-called high spectral regions.

4. CONCLUSIONS

We have investigated the behaviour of vowel classification accuracy as a function of increasing spectral range, with the aim of identifying the respective spectral regions in which phonetic and speaker-specific influences are the strongest. 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 speaker variability, if at all present, will be manifest most strongly in the F3 and higher spectral regions. Perhaps more importantly, our results demonstrate, as foreshadowed by Ainsworth & Foster [1], that vowel classification accuracies obtained over the entire spectral range $[0,f_s/2]$ may not be representative of the highest accuracies otherwise attainable in the presence of speaker variability. In addition, we have provided a methodology that may prove to be useful in the selection of speakers used to train a vowel or speaker recognition system.

REFERENCES

[1] W. A. Ainsworth & H. M. Foster, "The Use of Dynamic Frequency Warping in a Speaker-Independent Vowel Classifier", in R. De Mori & C.Y. Suen (Eds.), "Proceedings of the NATO Advanced Study Institute on New Systems and Architectures for Automatic Speech Recognition and Synthesis" (Springer Verlag, Heidelberg), pp. 389-403, 1985.

[2] F. Clermont, "Formant-contour models of diphthongs: A study in acoustic phonetics and computer modelling of speech", Doctoral Thesis, The Australian National University (Research School of Physical Sciences and Engineering), Canberra, Australia, 1991.

[3] F. Clermont & P. Mokhtari, "Automatic frequency-band specification in cepstral distance computation", Technical Report CS11/94, Dept. of Computer Science, University College, University of New South Wales, ADFA, 1994.

[4] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", IEEE Trans. ASSP-29 No.2, pp. 254-272, April 1981.

[5] S. Hayakawa & F. Itakura, "Text-Dependent Speaker Recognition using the Information in the Higher Frequency Band", Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing, I-137 - I-140, April 1994.

[6] H. Kuwabara & T .Takagi, "Acoustical parameters of voice individuality and voice-quality control by analysis-synthesis method", Speech Communication, Vol. 10, Nos. 5-6, pp. 491-495, December 1991.

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

[8] K.-P. Li, & G. W. Hughes, "Talker differences as they appear in correlation matrices of continuous speech spectra", J. Acoust. Soc. Am., Vol. 55, No.4, pp. 833-837, April 1974.

[9] O. Mella, "Extraction of formants of oral vowels and critical analysis for speaker characterization", Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, pp. 193-196, April 1994.

[10]K. K. Paliwal, & P. V. S. Rao, "Evaluation of Various Linear Prediction Parametric Representations in Vowel Recognition", Signal Processing, Vol. 4, pp. 323-327, 1982.

[11]G. E. Peterson, & H. L. Barney, "Control Methods Used in a Study of the Vowels", J. Acoust. Soc. Am., Vol. 24, No.2, pp. 175-184, March 1952.

[12]G. E. Peterson, "The Information-Bearing Elements of Speech", J. Acoust. Soc. Am., Vol. 24, No. 6, pp. 629-637, November 1952.

[13]L. C. W. Pols, H. R. C. Tromp & R. Plomp, "Frequency analysis of Dutch vowels from 50 male speakers", J. Acoust. Soc. Am., Vol. 53, No. 4, pp. 1093-1101, 1973.

[14]K. N. Stevens, "Sources of Inter- and Intra-Speaker Variability in the Acoustic Properties of Speech Sounds", Proc. Seventh Int. Congress of Phonetic Sciences, pp. 206-232, August 1971.

[15]G. T. Toussaint, "Bibliography on Estimation of Misclassification", IEEE Trans. IT-20, No.4, pp. 472-479, July 1974.

[16]D. J. P. J. van Nierop, L. C. W. Pols & R. Plomp, "Frequency Analysis of Dutch Vowels from 25 Female Speakers", Acustica, Vol. 29, No.2, pp. 110-118, August 1973.

[17]B. Yegnanarayana & D. R. Reddy, "A Distance Measure Based on the Derivative of Linear Prediction Phase Spectrum", Proc. IEEE Int. Conf. on Acoust., Speech and Signal Processing, pp. 744-747, 1979.