AN INVESTIGATION OF THE SPEAKER FACTOR IN VOWEL NUCLEI

Clive Cooper and Frantz Clermont

Department of Computer Science
University College, UNSW
Australian Defence Force Academy

ABSTRACT - Experiments in computer speaker identification are reported, which shed some light on speaker variability in the first three formants of the vowel nuclei of selected monosyllabic English words. The experimental results show that there is considerable spectro-temporal variation of speaker information throughout the vowel nuclei, but that the information does not appear to be "localised". Evidence is also presented in support of the hypothesis that speakers could be characterised in certain vowel subspaces.

INTRODUCTION

In speaker recognition systems certain features of the speech sound are exploited in order to distinguish one speaker from another. Any attempt to compare the amount of speaker specific information in a feature, and hence the relative merit of one feature over another, must consider a number of issues, some of which are noted below.

First, a measure of the amount of speaker specific information is required. The ultimate purpose of a speaker recognition system is to distinguish one speaker from another, and so a reasonable measure of the speaker specific content of a feature is the fraction of correct classifications using the feature. Since only one type of error is made in speaker identification experiments, whereas two types of error are possible in speaker verification, for this paper the preferred experimental method to measure speaker specific content is speaker identification.

Secondly, the pattern classifier must be optimally matched to the feature in the sense that the optimal classifier yields the maximum recognition rate. There is no known technique that determines such an optimal classifier, thus the comparisons between features must be assessed in relation to the two tuple: (Feature, Classifier).

Finally, speaker specific information can be extracted from phonetic contexts ranging from phoneme-length to sentence-length utterances. While sentence length processing offers the flexibility of extracting acoustic prosodic information (eg pitch contours), a perceived disadvantage by opponents of this approach is that statistical smoothing takes place. Large amounts of speech data are then used to gain as much knowledge of the underlying probability distribution as possible. But, since so many different acoustic events are contained in a sentence long utterance, the effect is to meld across events and hence obscure the variation in the articulatory peculiarities of different speakers. On the other hand, when a specific part of a phoneme representing a single acoustic event is subject to statistical analysis, then the speaker specific information is less likely to be obscured by the influence of dissimilar events. Based on this premise the vowel nuclei are an attractive phonetic event.

The paper begins with a brief review of some techniques that have been used to characterise vowels. There follows an experimental investigation of the variation of speaker information throughout the nuclei of nine vowels uttered by four speakers of Australian English. The investigation motivates the idea of using vowel subspaces to characterise speakers. The hypothesis is tested in the second investigation using the moderately large (33 speakers) vowel formant data of Peterson and Barney in which the data per speaker is, however, very sparse.

FORMANT DESCRIPTION OF VOWELS FOR SPEAKER CHARACTERISATION.

Whatever utterance length is selected, either a sparse or dense spectral analysis can be adopted. In dense spectral analysis all of the spectral information throughout the auditory spectrum might be used. In contrast, a sparse analysis might focus on the spectral peaks associated with the formants. This paper is concerned with the formants F1, F2, and F3 of nine vowels.

It is well known that the vowels contain significant speaker specific information. An interesting question is how best to characterise the vowels in terms of the formants. For example, Sambur (1975) suggests using the values of Fi (i = 1..3) at the turning point in the vocalic nucleus where dF2/dt = 0. The formant values so obtained are considered to characterise the vowel because the turning point is assumed to coincide with the frequencies corresponding to the vowel target position. However, there is not always a well defined turning point for all the vowels or even for different realisations of the same vowel articulated by one speaker. Sambur ranks F3 for the high back vowel /u/ and F2 for the high front vowel /i/ second and third, respectively, in a list of 38 best performing features.

In her work on vowel perception, Huang (1992) models the vowels by both the formant trajectories and the consonantal context. This yields several involved models beyond the simple "one point" characterisation. The results confirm that better vowel characterisation is achieved by "three point" models and, by taking into account the vowel context, a further improvement is achieved.

Goldstein (1976) computes formant trajectories for diphthongs, retroflex sounds and, in particular, the tense vowels /a/, /e/, /i/ and /u/ in the single b_d context for each speech sound. In the subsequent analysis, some 199 specific events are selected for examination, rather than investigating the continuous variation of the speaker information through the speech sound.

Broad and Clermont (1987) provide a parametric representation of the formant trajectories based on a superposition principle using the onset, target and offset periods of the vowel. In some contexts the models yield substantial agreement with the observed data.

Other researchers arbitrarily select a point in the stationary part of the vocalic nucleus.

To shed additional light on this issue, the first experiment described in this paper investigates the temporal variation of speaker specific information throughout the vocalic nucleus of nine vowels. The results of this experiment motivate the transition from the initial small database to a larger database.

VARIATION OF SPEAKER INFORMATION IN THE VOCALIC NUCLEUS

For this investigation the database consists of 63 /CVd/ words uttered by 4 adult male speakers of Australian English, where C is one of the consonants /h, b,d,g,p,t, k/ and V is one of nine vowels as in "heed", "hid", "head", "had", "hard", "hod", "whod", "hudd" and "herd" . Each of the speakers utters each word 5 times giving a total of 1 260 utterances. The vocalic nucleus of each word is represented in the formant domain by 11 equally time-spaced values of F1, F2 and F3. Since the vowel durations differ, 1/11th of the vowel duration is called one unit of normalised time. Feature vectors are formed by taking the formant values in a window consisting of 1 to 11 consecutive time-slots within the vocalic nucleus. (Clermont, 1991;1992)

For a given part of the vocalic nucleus and a given feature vector, a speaker identification experiment is performed by selecting one utterance from one of the speakers as the test utterance. The remaining 4 utterances for the test speaker and the corresponding 4 utterances of the other speakers are used as data representative of the speaker classes. A nearest neighbour algorithm is used to classify the test utterance as one of the 4 speakers. The temporal variation of speaker-specific information in the vocalic nucleus is determined by sliding, through the vocalic nucleus, a window of width varying from one normalised unit of time to the complete utterance duration. This experiment is repeated for every word and for every speaker and all possible window durations.

The feature vectors used are:

$$f_{ki} = [f_i \ \ldots \ f_{i+k-1}]' \quad \text{for window widths of } k = 1 \ldots 11 \text{ and for time slots } i = 1 \ldots (12-k)$$

where f_{ki} = is the value of the formant (F1, F2 and F3 in turn) in the i'th window of width k.

Figures 1-3 show the results for F1, F2 and F3 when the window width is 1 normalised unit. For F1 there are a few isolated islands of identification where the identification rates rise above the plane drawn at 75% identification. For F2 there are notable regions of high speaker-specific information for the front vowel /i/ and the back vowels, particularly /U/. For F3 there are abundant regions of high identification in the back vowels, and for /U/ there are some regions where 100% identification is achieved. The variation of the average identification rates, over the seven words for each vowel, for F3 are shown in Figure 4.

When the regions of 100% correct identification are visually compared with the formant trajectories, no obvious correlation is apparent. The results suggest that speaker specific information is not localised, but rather that it is distributed over a subspace of the vowel space. The average identification rates over the seven contexts of each vowel using pairs of formant values that are the average value over the vocalic nucleus are given in Table 1.

Vowel	IPA	ARPA	F1-F2	F1-F3	F2-F3
1	/i/	IY	81	61	81
2	/I/	IH	72	61	76
3	/ɛ/	EH	80	54	77
4	/ae/	AE	87	62	80
5	/a/	AA	74	83	80
6	/ɒ/	AO	79	86	84
7	/U/	UH	85	97	97
8	/ʌ/	AH	84	91	91
9	/ɜ/	ER	82	87	94

Table 1: Vowel Performance in Speaker Identification Using Pairwise Formants

Clearly, the first vowel /i/ for F1-F2 has a rich element of speaker specific information, as do several of the back vowels for F1-F3.

To further investigate a vowel subspace that might characterise a speaker, the formant data of Peterson and Barney (Peterson & Barney, 1952) is analysed. This is a "good" database to test the conjecture since it has a moderately large number of male speakers (33) and the data is sparse in the sense that there are only two utterances per vowel: one for the test utterance and one for the reference utterance.

SPEAKER CHARACTERISATION USING VOWEL SUBSPACES

First, the complete vowel subspace of the Peterson and Barney data is examined using the following feature vectors:

One Formant: $f_i = [Fk_i]$ for formants k = 1..3 and vowels i = 1..10;

Two Formants: $f_{jk} = [Fj_i : Fk_i]$ for $(j,k) \in \{(1,2),(1,3)\ (2,3)$ and vowels i = 1..10;

Three Formants: $f_3 = [F1_i : F2_i : F3_i]$ for vowels i = 1..10;

Each speaker in turn is selected as the unknown speaker. The appropriate feature vector is formed and its Euclidean distance from the second utterance for each of the speakers is computed. The

identity of the speaker corresponding to the minimum distance is assigned to the unknown speaker. The results are shown in Table 2:

Formant(s)	% Correct
1	66
2	75
3	79
1-2	79
1-3	89
2-3	92
1-2-3	94

Table 2: Speaker Identification Accuracy using whole Vowel space of Peterson and Barney Data

Finally, all possible subspaces are selected and the identification experiments repeated for each of the formant pairs F1 - F2, F1 - F3 and F2 - F3. The results are shown in Table 3:

Number of Vowels	F1-F2	F1-F3	F2-F3
2	50	50	55
3	62	64	74
4	73	76	83
5	81	79	88
6	82	86	91
7	85	85	92
8	86	88	96
9	82	88	93
10	79	89	92

Table 3: Speaker Identification Accuracy using Vowel subspaces of Peterson and Barney Data.

For all pairs of formants the amount of speaker-specific information has a peak for a subspace of eight vowels. (The vowels not included in this subset are /a/ and /U/.) It is not known if the peaks represent local minima. Remarkably, eight vowels for the F2-F3 pair perform better than all 10 vowels and all three formants.

CONCLUSION

For the nine vowels examined in the first part of this paper there is considerable variation in the speaker-specific information throughout the vocalic nuclei. However, no obvious correlation is apparent between the regions of high speaker-specific information and any features of the vowel formant trajectories. In terms of specific vowel performance the results generally agree with the literature. (Sambur, 1975; Goldstein, 1976; Paliwal, 1984) Namely, F3 for the high back vowels, followed by F2 for the front vowels perform best in distinguishing one speaker from another. Also, speaker specific information is not entirely localised in a few vowels but appears to be distributed throughout a subspace of the vowel space as illustrated graphically in Figures 1-3.

When a moderately large database of 33 speakers with only two utterances for each vowel is examined, there is some evidence to support the hypothesis that a speaker may be characterised by a vowel subspace. Recent evidence offered by Furui & Matsui (1994) indeed suggests that better accuracy and more robust speaker identification is achieved by concatenating phonemes that are characteristic of a speaker into a dynamic sentence length utterance. It would therefore be relevant to investigate the possibility of defining a complete and minimal subspace. In this context, "complete" means that the subspace spans the normal articulatory gestures of the speaker, and "minimal" means that omission of one element will result in a non-spanning set. In such a search the admission of other phonemes into the space is not precluded. Indeed, it is well known that other

phonemes are rich in speaker specific information (Eatock & Mason 1994) and consequently a subspace that might characterise a speaker is likely to also contain non-vowel elements.

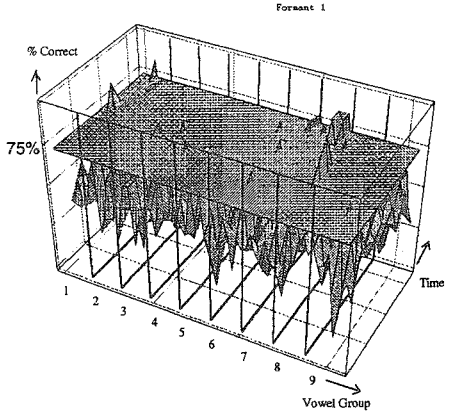

Figure 1:

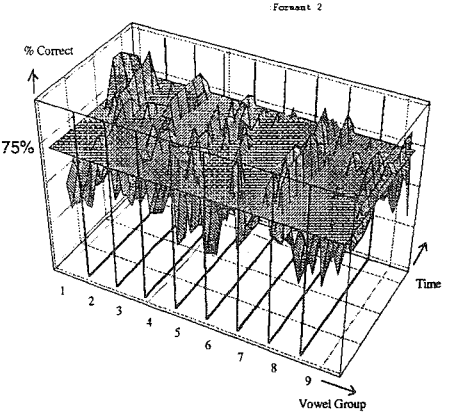

Figure 2

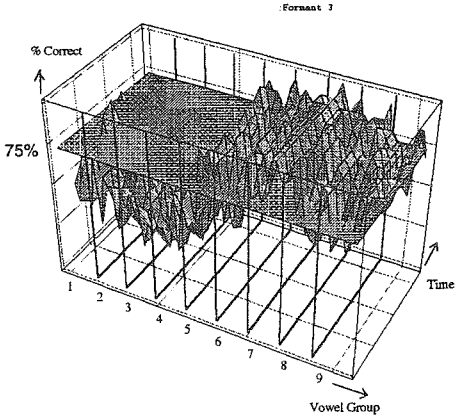

Figure 3

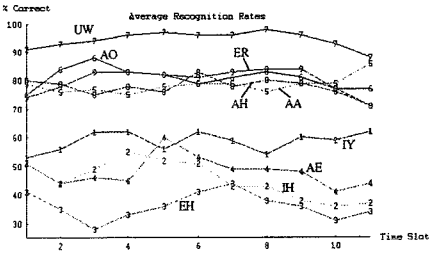

Figure 4

Graphical Representation of Selected Results for All 11 Time Periods.

REFERENCES

Broad D. J., and Clermont F. (1987) *A methodology for modeling vowel formant contours in CVC context.* J. Acoust. Soc. Am. Vol 81, No. 1 pp. 155-165 , January 1987.

Clermont, F. (1991) *Formant-Contour Models of Diphthongs: A study in acoustic phonetics and computer modelling of speech*, PhD Thesis, The Australian National University (Research School of Physical Sciences), Canberra, Australia.

Clermont, F. (1992), *Formant-contour parameterisation of vocalic sounds by temporally-constrained spectral matching*, Proc. IV Australian International Conf. Speech Science and Technology, pp 48-53.

Eatock J. P., and Mason J. S. (1994) *A Quantitative Assessment of the Relative Speaker Discrimination Properties of Phonemes.* IEEE Int. Conf. Acoust. Speech and Sig. Proc., pp I-133-136.

Furui S., and Matsui T. (1994). *Phoneme-Level Voice Individuality Used in Speaker Recognition.* ICSLP 94, Yokohama S25-7.1-4.

Goldstein, U. G. (1976) *Speaker-identifying features based on formant tracks*, J. Acoust. Soc. Am., Vol 59, No. 1, pp 176-182, January 1976.

Huang C. B. (1992) *Modelling Human Vowel Identification Using Aspects of Formant Trajectory and Context.* Speech Perception, Production and Linguistic Structure, Tohkura Y., Vatikiotis-Bateson E., & Sagisaka Y. Eds, (IOS Press: Oxford, England).

Paliwal K. K., *Effectiveness of Different Vowel Sounds in Automatic Speaker Identification.* Journal of Phonetics, 12, pp17-21, 1984.

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

Sambur M. R. (1975) *Selection of Acoustic Features for Speaker Identification*, IEEE Trans. Acoust. Speech and Signal Processing, Vol. ASSP-23, No. 2, pp 176-182, April 1975.