EVALUATING REPRESENTATIONS OF SEGMENT LEVEL DYNAMICS IN ACOUSTIC-PHONETIC MAPPING.

Dave Davies* and J.Bruce Millar
Computer Sciences Laboratory,
Research School of Information Sciences and Engineering,
Australian National University.

ABSTRACT

The time domain of phonetic events is examined with a view to proposing approaches that depart from the regular clock-timed representation of the acoustic analysis of speech and the use of simple time derivatives as a representation of temporal information in the acoustic vector normally fed to automatic speech recognition systems. A novel approach of incorporating temporal information within sequential acoustic vectors is introduced. Basic acoustic parameters derived using a source-synchronous analysis technique are combined with a coded representation of their temporal environment. The concept of similarity length is described and elaborated in various forms that can be applied to the description of sequences of speech up to phonetic segment level. A phoneme-recognition-based evaluation criterion is developed in order to evaluate the performance of such acoustic vectors in efficiently representing acoustic-phonetic mapping. Experiments that apply this analysis and evaluation to the acoustic representation of stop consonant are described. Results are presented in the form of the ranking of individual stop consonants against all other phonemes when this analysis is performed using simple acoustic parameters with and without the addition of temporal information.

1. INTRODUCTION

From experimental studies of speech perception we know that temporal information, in the form of duration or the rate of change of acoustic parameters, can be phonetically distinctive. A central challenge to the principal automatic speech recognition (ASR) technologies of dynamic time warping (DTW), hidden Markov modelling (HMM) and artificial neural networking (ANN) has been to deal effectively with temporal information. This challenge has been apparent both in learning to efficiently detect and use temporal information where it is carries phonetically or lexically distinctive cues and, conversely, to learn to ignore it when it is not useful.

ASR systems typically encode acoustic feature data in an "acoustic vector" which is defined at regular intervals of the order of 10ms. Temporal information is commonly incorporated as the first or second time derivative of acoustic parameters and, generally, also at a higher level in the pattern matching process such as in state-transition probabilities in HMM systems or expanded input vectors in ANN systems. One approach to reducing this temporal modelling burden on the pattern matching system is to construct the acoustic vector from the output of acoustic probes designed to encode temporal contexts at any relevant time scale for any combination of acoustic parameters.

This study has as its primary focus the efficient incorporation of both spectral and temporal information in the acoustic vector. Local estimates of parameter derivatives are prone to noise and do not generally reflect broader segment level dynamics. Broadening the temporal support of acoustic parameters and their derivatives involves estimating the maximum feasible extent of this support. Here we propose a simple method for estimating temporal support and show that it can provide a more valuable component of the acoustic vector than the local acoustic parameter derivatives themselves. The performance of this approach is tested by means of assessing the capacity of the acoustic vector to map distinctively onto the phonemic domain.

2. BACKGROUND

Evidence for the important role played by the temporal patterning in mediating phonemic distinctions is scattered widely throughout the phonetic sciences literature. This evidence includes the influence of the duration of voiced and unvoiced segments, their relationship to overall speaking rate, stress patterns and their position in words, and also to the influence of the dynamic variation of specific frequency-based features which can be good predictors of places of articulation. The challenge of making effective and efficient use of this rich evidence of the relevance of temporal patterning is to a large extent unresolved in mainstream automatic speech recognition. The inadequacy of the duration modelling capacity of standard HMMs has been widely discussed [e.g. 4,5,6].

Several approaches to estimating the stability of acoustic parameters can be found in the literature, notably the temporal decomposition technique of Atal [1] and parametric filtering techniques such as that of Li and Gibson [2]. The approach that we take can be shown to have much in common with that of Li and Gibson to the extent of applying tuned filters to the acoustic signal and looking at the stability of the filter output.

3. ANALYTIC METHODS

This study developed several analytical methods both to measure the temporal extent of acoustic features and to evaluate the impact of incorporation of temporal extent information within an acoustic vector used as a probe of acoustic-phonemic mappings.

3.1 Evaluation Process

Each acoustic vector represents a quantised perspective on the acoustic characteristics of one glottal epoch of a speech utterance. The mapping of this view of the data onto a phonemic view of the data is determined by building an accumulation matrix of co-occurrences of acoustic vector values and the incidence of a specific phonemic label. This matrix can then be used to determine the likelihood of any particular phoneme given a particular acoustic vector. In this study the phonemic map of a particular acoustic vector is expressed as a rank-order of the likelihood of each phoneme given the acoustic vector.

3.2 Derivation and Optimisation of Acoustic Vectors

Using this raw statistical approach, and thereby avoiding the computationally expensive training involved in optimising a full recogniser, we can focus on optimising the derivation and presentation of the acoustic vector and its individual components.

A cost of avoiding training, or smoothing the data, is the requirement for high data densities. This was achieved by coarse quantisation of parameter ranges which were divided, using a parameter value histogram, into ranges of equal population. The impact of quantisation was tested using jack-knife evaluation of stop rankings with all acoustic vectors associated with a stop instance removed from the likelihood matrix as each vector was evaluated. Likelihoods were aggregated and rankings evaluated on a phone instance basis. A quantisation of four bits was used for all parameters reported here.

3.3 Frequency Domain Parameters

Frequency domain analysis was performed on source-synchronous frames generated using the fundamental harmonic extraction technique described in Davies and Millar [3]. This technique applied to the ANDOSL data, which was recorded under anechoic conditions, allowed the tracking of very low level source activity.

A source-synchronous approach was considered to be most appropriate for these tests since it provides a minimum time scale for sampling supra-glottal dynamics. Further, it allowed testing and optimisation of results using sub-ranges of the inter-glottal period. Testing showed best results for samples taken over the first 4.5ms following glottal closure.

Frequency domain transforms were performed using a Goertzel DFT filter with formants taken as the largest peaks found within their normal frequency range. Peak size was taken as the area under the peak between adjacent minima. The irregular shape of broad formant peaks led to the use of a ratio of peak size to height as a measure of bandwidth. The area under the lowest peak, clearly distinguishable from F1 by its relatively stable frequency and increase in relative energy near glottal closure, is presumed to be nasal energy (for illustration see figure 12 in [8]) and is measured as a fraction of total energy (0 to 5kHz). Formant energy ratios are the energy in the upper half of the formants' normal frequency range as a fraction of the total energy in this range.

Jack-knife evaluation was used for optimising many measurement parameters such as the shape of the fundamental harmonic filter, glottal phase and length of the signal sample sent to the DFT filter, frequency domain granularity and formant frequency ranges. In later testing (e.g. Fig. 3), where coarse grain transformed parameters with relatively uniform populations were used, the jack-knife approach was replaced with a simple reference to pre-calculated likelihood tables.

3.4 Temporal Support Parameters

We use a Parametric Similarity Length (PSL) as a computationally efficient estimate of the temporal support of a parameter value and a possible proxy for durational information. The PSL is defined as the temporal interval over which an acoustic parameter maintains its value within a given range. While this range can be optimised for individual acoustic parameters, for the results reported here it was fixed at ±20% which was found to be a suitable compromise for a wide range of parameters. The basic PSL has been enhanced using a bit-coded shape factor that codes, in two bits, the direction of movement of the parameter values at the extremes of the region of similarity - ie each bit represents an end point and its value codes upward or downward movement of the parameter value. We have also evaluated the PSLS estimated for the first and second time derivatives of the base acoustic parameters.

4. EXPERIMENTAL DESIGN

The generic design of all the experiments was first to evaluate parameters and their time derivatives, and the PSLS of parameters and derivatives separately. Then, in a second round evaluation, each base parameter was combined with its temporal measures as a two dimensional acoustic vector. In both cases phonemic prominence was then assessed as evidence for the benefit of each form of temporal information within the acoustic vector.

4.1 Speech Data

The data used consisted of 200 phonetically balanced sentences from four speakers of the ANDOSL speech database. The speakers (S052, S058, S065, S106 comprising 3 females and 1 male) were chosen from the 'Cultivated' accent group of the database to avoid unnecessary bias of the results toward strong Australian accents.

4.2 Selection and Derivation of Acoustic Parameters

The acoustic parameters used are summarised in Table 1 and include the signal energy, glottal period, the frequency and bandwidth of the first three formants and a range of frequency band energies selected from the many alternatives discussed in the ASR literature. Glottal period (Tx) measures the time

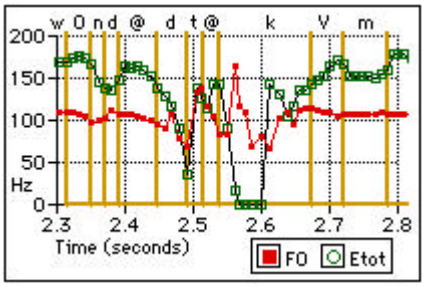
between instants of glottal closure as represented by the fundamental harmonic.

Parameter	Code	Parameter	Code
Total Energy (Log)	Etot	F2 Energy Ratio	ERf2
Fundamental Harmonic Energy*	Efx	0 to 2kHz Energy	E2k
Glottal Period	Tx	600 to 2800Hz Energy	E6to28c
1st Formant Freq*	F1	2 to 3kHz Energy	E2to3k
2nd Formant Freq.	F2	2 to 5kHz Energy	E2to5k
3rd Formant Freq.	F3	Bandwidth F1*	BWf1
F0 to F1 Frequency	F0toF1	Bandwidth F2	BWf2
F1 to F2 Frequency	F1toF2	Bandwidth F3	BWf3
Nasal Energy*	En1	0 to 300Hz Energy	E0to3c
F1 Energy*	Ef1	0 to 400Hz Energy	E0to4c
F2 Energy	Ef2	100-300Hz Energy	E1to4c
F3 Energy	Ef3		
F1 Energy Ratio*	ERf1	* See text for explanation	

Table 1. Acoustic parameters used in this study.

4.3 Acoustic Parameter Evaluations

A total of 24 base acoustic parameters were used over the course of this study. Acoustic vectors comprising their first and second derivatives, PSLS and combinations of these factors were constructed. Testing was performed on 216 acoustic vectors comprising the 24 base acoustic parameters plus the 8 temporal measures of each, listed in the caption of figure 2,



were tested. For each acoustic vector the "rank order" of the phonemes that they predict was recorded.

Figure 1. Acoustic parameters, excitation fundamental frequency and signal energy, around a sequence of stops in the phrase "wanted to come".

In this short paper we have chosen to focus on the evaluation of acoustic vectors with stop-like consonants - the stops plus the affricates. Figure 1 illustrates the time course of two acoustic parameters in the vicinity of voiced and unvoiced stop closures from sentence S002 by speaker S106. In the ANDOSL data the initial boundary of the closure is placed where there is an abrupt drop in energy at all frequencies, or above 300-500Hz where voicing continues into the closure. The initial boundary of an utterance initial stop is placed at the onset of voicing in the case of a voiced stop, or at an arbitrary point 60ms before the beginning of the release phase in voiceless stop. Mirror image rules apply to the placement of final boundary of an un-released utterance-final stop. The final boundary of the closure is placed at the beginning of the release [7].

5. RESULTS

Table 2 summarises the occurrence of the top ten base parameters in the best ranked 100 acoustic vectors aggregated over all stop-like phonemic labels, providing a very general evaluation of the base parameters. Glottal period can be seen to dominate the above average rankings and, importantly, shows strong consistency across speakers whereas other parameters such as F3 bandwidth show strong speaker dependence.

Parameter	Speakers				Total
	S052	S058	S065	S106	
Glottal Period	29	22	26	21	98
F1 bandwidth	10	10	10	17	47
Nasal energy	9	15	12	3	39
Total energy	7	15	7	9	38
F3 bandwidth	3	3	21	3	30
0-300Hz energy	16	4	5	2	27
0-F1 Hz energy	3	4	3	12	22
2-3kHz energy	3	10	0	1	14
2-5kHz energy	3	6	0	3	12
0-400Hz energy	3	4	4	0	11

Table 2. The occurrence of the top ten base parameters in the best ranked 100 acoustic vectors aggregated over all stop-like phonemic labels

Rankings from the preliminary jack-knife tests for the various temporal variants applied to Tx for all phonemes are shown in figure 2. The similarity length and its shape extension generally outperform the parameter derivatives - ie are ranked closer to 1. The base Tx ranks best for several phonemes (eg. /o:/ and /u:/) with 'Tx.Sha' ranking best for the majority of the remaining phonemes. The difference between the derivatives (eg. Tx.Dif.Dif is the second time derivative of Tx) and PSL based parameters is least for the stops, nasals and some fricatives.

Improved performance for the stop-like consonants is illustrated in figure 3 where rankings for two dimensional probes consisting of Tx together with some of its temporal variants are plotted. Best results are obtained with Tx combined with the base PSLS (Tx&Tx.Sha) and the shape of the second derivative (Tx&Tx.Dif.Dif.Sha). The latter parameter appears to be acting more as a pattern detector for periods of extended rapid parameter variation that can be seen in Figure 1 between 2.5 and 2.6 seconds into the utterance.

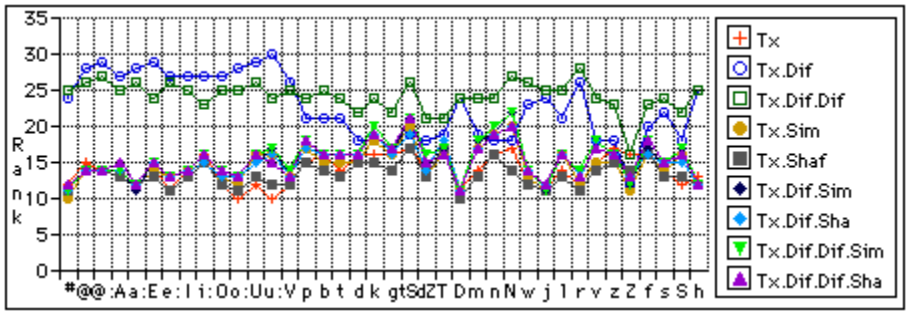


Figure 2. A comparison of all temporal codings applied to the parameter glottal period for the full phoneme set.

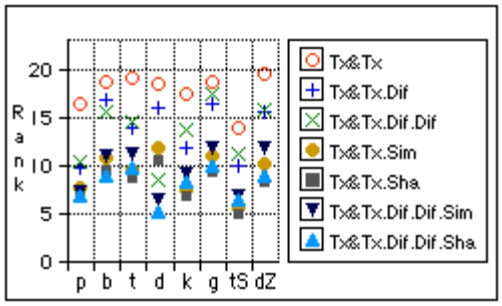


Figure 3. Mean association rankings for 2-D acoustic vectors for the parameter glottal period applied to stop-like phonemic labels.

6. CONCLUSIONS

The results of these experiments point to potential advantages for including some form of durational proxy in the acoustic vector of automatic speech recognition systems. The similarity length tested here consistently outperforms the parameter time derivatives and the addition of terminal shape information provides a further advantage.

The dominance of glottal period in the results for stop-like phonemes may reduce in noisy or restricted bandwidth speech where low energy regions of the fundamental harmonic are not available. Preliminary testing has shown little degradation in performance with moderate thresholding of the fundamental harmonic, however the impact of noise is likely to be much greater than that of a simple logical threshold.

The investigation of this technique is progressing to a detailed evaluation of continuant phonemes and robustness under various signal conditions.

7. ACKNOWLEDGEMENTS

The Australian National Database of Spoken Language was prepared at Sydney University, the National Acoustics Laboratory, Macquarie University and the Australian National University under the auspices of the Australian Speech Science and Technology Association, with major funding from the Australian Research Council.

8. REFERENCES

1. Atal,B.S. (1983) Efficient coding of LPC parameters by temporal decomposition, Proceedings of ICASSP, pp.81-84.
2. Li,T-H., Gibson,J.D. (1996) Speech analysis and segmentation by parametric filtering, IEEE Transactions on Speech and Audio Processing, 4, 203-213.
3. Davies,D.R.L., Millar,J.B. (1996) The evaluation of a computationally efficient method for generating a voiced-source synchronised timing signal, Proc. 6th Australian International Conference on Speech Science and Technology, Adelaide, December, pp.527-532
4. Russell,M.J., Moore,R.K. (1985) Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition, Proceedings of ICASSP, Tampa, March, pp.5-8.
5. Levinson,S.E. (1986) Continuously variable duration hidden Markov models for automatic speech recognition, Computer Speech and Language, 1, 29-45.
6. Pols,L.C.W., Wang,X., ten Bosch,L.F.M. (1996) Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR, Speech Communication, 19, 161-176.
7. Croot,K., Taylor,B. (1995) Criteria for acoustic-phonetic segmentation and word labelling in the Australian National Database of Spoken Language, URL=http://andosl.anu.edu.au/andosl
8. Hess,W.J. (1974) A pitch -synchronous digital feature extraction system for phonemic recognition of speech, Proceedings of IEEE Symposium on Speech Recognition, Pittsburgh, April, pp.112-121.

- http://discus.anu.edu.au/~daved