A REASSESSMENT OF TEMPORAL INFORMATION IN SPEECH PROCESSING

J. Bruce Millar, Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Institute of Advanced Studies, Australian National University

David R. L. Davies Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Institute of Advanced Studies, Australian National University

1. INTRODUCTION

The work described in this paper has been motivated by consideration of both parsimony in the representation of speech acoustics and observations of the degradation of automatic speech recognition (ASR) performance when speaking rate changes. The acoustic-phonetic processing within an ASR system involves the matching of a representation of the acoustic stream with a phoneme symbol sequence that has been promoted by a vocabulary list or a language model. The best match between the continuous acoustic stream and the sequence of phoneme symbols may be based on the probabilistic evidence for the presence, absence, and order of phonemes derived from acoustic models of the individual 40-50 phonemes. The stochastic acoustic models against which the evidence is assessed are based on acoustic feature vectors that conventionally represent the average spectral characteristics of fixed windows, of 20-30ms duration, on the acoustic stream. The models of individual phonemes will typically incorporate fairly crude representations of the sequence of the different acoustic vectors that best approximate the phoneme in a number of contexts.

The acoustic stream is rich with timing information that can be characterised in terms of the temporal extent of locally quasi-static feature values and the trajectory of these values over time. We are concerned to capture more directly some primitives of acoustic feature trajectories that relate to phonetic quality rather than just sequences of regularly sampled spectral values. We wish to build acoustic models that do not simply treat "time" as defined by the "clock on the wall" but rather in a way that is relative to the acoustic-phonetic structure. We therefore aim to capture "temporal extent" and "spectral shape" in an appropriate parameter space, and then to test phonetic discrimination within this space.

An ASR system that receives a stream of regularly sampled acoustic vectors to be matched against a phoneme symbol sequence is faced with simultaneously performing three quite different tasks - quantification of the temporal information, aggregation of adjacent AVs into phoneme scale groupings and performing an existence/order match at the utterance level. This work attempts to evaluate techniques that allow each of these tasks to be performed independently with an emphasis on optimising the information coded in the individual AVs.

After placing our approach within a literature context, we describe a simplified implementation in which individual acoustic features are processed according to the principles outlined. We set the novel spectro-temporal processing in the context of appropriate pre-processing and post-processing options. Specifically, the work discussed in this paper looks at temporal/sequential processing of

speech on three distinct fronts: pre-processing via the source synchronous approach to acoustic signal analysis, some novel spectro-temporal representations of acoustic parameters, and post-processing via the aggregation of sequential acoustic-phonetic likelihood rankings.

1.1.1 Background

The auditory system is well designed for processing temporally organised information while the visual system is well designed for spatial information processing. It is therefore appropriate that the way that speech is organised temporally should influence the way that it is processed by machine. The dimension of "time" is so fundamental to speech processing that it is easy to be overlooked as an object of study.

Considerable attention has been given in the acoustic phonetic literature to the issue of rapid speech processes. This has included studies on the "reduction" of spectral information during shortened vowels [e.g. Lindblom, 1963; Fourakis, 1991; Van Son and Pols, 1992] and the encoding of "phonetic length" which discriminates between categories of vowels which are labelled "long" (or tense) and "short" (or lax) [e.g. Nooteboom and Doodeman, 1980]. It appears that combinations of relative spectral information and relative temporal information are required to provide evidence for a sound phonetic judgement.

Given this acoustic phonetic evidence it should not have come as a surprise that automatic speech recognition (ASR) systems that treat spectral and temporal information separately generate errors when the rate of speech encountered in testing differs from that encountered in training [e.g. Seneff, 1996]. It could be predicted that a given vocalic nucleus, for instance, spoken in different temporal circumstances, could be realised as a number of differently shaped trajectories in an acoustic-feature space where the variation in each parameter of the trajectories forms the "statistical distribution" that represents the phonemic category**.**

The underlying philosophy of the current study is that it would be useful to capture some primitives of syllabic trajectories that relate to phonemic category. As a first approximation to a more adequate syllabic trajectory model we should check a primitive model that characterises "duration" and some simple aspects of "shape" in an appropriate parameter space. Once having defined a simple model then its representation of trajectories may be used to explore phonemic discrimination in this space as a prelude to speech processing in this space.

It is noted that temporal information has been incorporated into the vast majority of speech processing systems by overlapping spectral measurement such that sequential spectral analyses capture changes occuring more rapidly than the underlying spectral measurement theory can support, and by extending these pseudo-instantaneous spectral measurements by differential and double differential versions to the overall spectral feature set.

The issue of the representation of time in speech has been examined from a number of perspectives by others and we briefly review these processes here.

1.1.1 Temporal Decomposition

Temporal decomposition of the speech stream in a way that is sensitive to the data within it rather than by some external reference has arisen in speech synthesis, speech coding, and speech recognition research. The concept of a target spectrum together with the specification of how it is realised in time was the basis of early work on speech synthesis-by-rule (Holmes et al., 1964).

The same principles were applied to speech coding when the redundancies present in externally-clocked LPC parameters were realised (Atal, 1983), and computational efficiencies have continued to be discussed for this method (e.g. Ghaemmaghami et al., 1998). The temporal decomposition approach to speech coding has also been evaluated as a means for creating phonemic labelling for use in speech recognition (Bimbot and Atal, 1991).

The approach to TD used by Atal used Singular Value Decomposition techniques to produce a local set of interpolation functions for describing the speech signal. While capable of producing functions that appear to reflect the distributed and overlapping nature of articulatory gestures the approach is highly computationally intensive. When applied to ASR rather than Atal's signal coding application, the technique poses questions about the stability of the interpolation functions with variations in parameters such as window width and does not necessarily produce functions that have a consistent association with the phonetic context.

Refinements of Atal's original method (eg. Bimbot and Atal 1991) have addressed the above issues with some success and evaluated the technique in a recognition context. Ghaemmaghami et. al. (1998) proposed a hierarchical TD approach in which they first determine Event Approximating Functions as basis functions reducing the complexity of the problem. Van Dijk-Kappers (1988) determined that filter-bank and log-area parameters were suitable for SD analysis while Ahlbom et al. (1987) dropped the SVD step and use clusters of similar AVs as the basis vectors. The complexity of the technique still impedes the ability to generate basis functions that are optimised for phonetic relevance.

1.1.2 The Mutual Information Approach

The temporal distribution of phonetically relevant information has been estimated in perceptual studies and information theoretic analysis (eg. Bilmes, 1998; Yang et.al. 2000). Yang et.al. evaluated the mutual information between signal features in the time-frequency plane and phonetic labels. They also measured the joint mutual information of pairs of acoustic features and the labels.

The significance of this form of analysis to the phoneme recognition problem lies in its direct association of acoustic parameters with the phonemic labels and the generalised nature of the MI measure compared with simpler geometric or correlation measures. As in the current work, the acoustic parameters are treated singly or in pairs, so avoiding the added computation complexity of high dimensionality, allowing multi-dimensional analysis to be performed at a later stage on a locally optimised and simplified AV stream.

Yang et.al. (2000) published results aggregated over a 19 phoneme set. Single phoneme data of this type would provide valuable constraints and reliability estimates for the faster geometrical measures used in the current study.

1.1.3 Time Trajectory models

In a stationary state HMM, signal parameter variation through the duration of a state is seen as noise, increasing model complexity and training data requirements. Trended HMMs model signal variation over the duration of the state. Deng and others (eg. Deng et.al. 1994; Deng, 1997) have used cubic polynomials to model signal variation with some success but increased model complexity.

2. SOURCE SYNCHRONOUS ANALYSIS

The stream of voice source impulses provides the most fundamental temporal segmentation of the speech signal for subsequent, primarily frequency domain, analysis. We have shown that for good quality speech signals, tested over a range of speakers and voice qualities, a simple fundamental harmonic extraction method (Hess, 1983) is likely to provide adequate accuracy for this purpose (Davies & Millar, 1996). The specific method used was a multi-stage time-reversed leaky integrator technique followed by a peak detector (Davies and Millar, 1996).

In the present study all signals were extracted from the broadband and anechoically recorded ANDOSL data corpus (Millar et al., 1994; 1997) and were processed using this method with parameters that were optimised over the speaker set analysed using mean ranking feedback (see section 4). No distinction was made between speech segments that were phonemically labelled as voiced or unvoiced. This had the advantage that phonation was tracked to its extremities. For computational purposes, long periods of unphonated speech were broken into arbitrary frames for subsequent analysis.

The major benefits of the source-synchronous approach lie in two areas. Firstly, the source excitation epoch provides the minimum time scale for catching short-term variation and typically supports more than twice the temporal resolution of the fixed-interval, and hence arbitrary phase, analysis systems. Secondly, it provides for a better representation of low-frequency features such as the lowest nasal and oral resonances. Essentially it produces a signal that is coherent by synchronising the cause and the effect of the excitatory activity in the vocal tract.

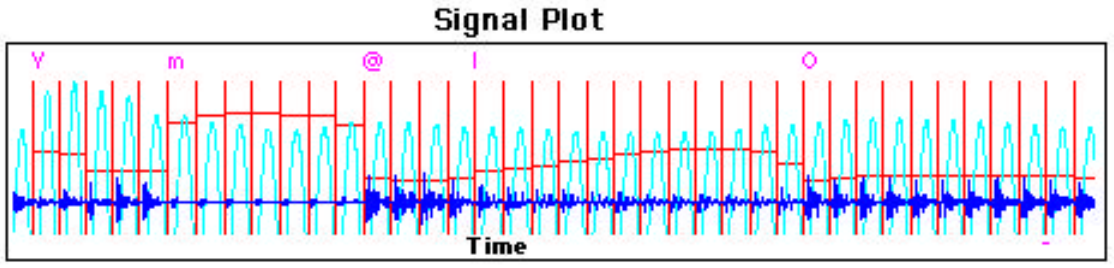

Figure 1. Signal processing example of source synchronous framing and subsequent analysis. The dark waveform is the microphone pressure waveform, the light near-sinusoid is the fundamental of the source synchronous analysis, the vertical lines are the source synchronous frames, and the horizontal bars plotted for each frame are the values of the nasal energy parameter (En1). The utterance is the central part of the words "come along".

3. SPECTRO-TEMPORAL ACOUSTIC VECTORS

The source-synchronously framed signals were then analysed one source epoch at a time. This analysis used the computationally efficient Goertzel DFT filter [Goertzel, 1958] to derive a spectral section representing the first 4.5ms of the source epoch of the speech used in this study. The length of the temporal window was optimised against the aggregated results of all the acoustic-phonetic likelihoods estimated in this study.

A fairly comprehensive set of one-dimensional acoustic parameters were extracted from individual source synchronous frames. The parameter set comprised measures of energy, energy ratios, and band-limited energies, formant frequencies, energies, and bandwidths, some frequency difference values and of course instantaneous excitation frequency in the form of the length of each source-synchronous epoch.

3.1 Excitation frequency measures

The glottal excitation period (Tx) is the reciprocal of the instantaneous excitation frequency (Fx), and was extracted directly from the source synchronous epoch. Tx has so far been the most extensively tested parameter in this work. This was partly due to the initial focus on source synchronous analysis but it also provided a reliable and rapidly evaluated parameter for system testing and early optimisation experiments for the nonlinear parameter transforms and the acoustic-phonetic associations.

3.2 Energy feature measures

Total Frame Energy (Etot) was normalised against the long-term average energy and expressed on a logarithmic scale. All other energy features for the frame were normalised against the total frame energy.

Fundamental Harmonic Energy (Efx) was measured as the energy of the fundamental component signal derived as the output of the fundamental harmonic extraction process used in the source synchronous analysis (Figure 1).

Nasal energy (En1) is the energy of the fixed frequency nasal resonance that consistently appears in source synchronous analysis in periods of oral tract closure or constriction. It was included to enable testing as a cue to consonant closure.

3.3 Formant Energy Ratio measures

The formant energy ratio parameters (ERf1, ERf2, ERf3) quantify the ratio of energy that occurs in the upper half of the formant frequency band compared to the total energy in the band. They combine both formant position and energy information and are expected to be more rubust to noise than F1 and Ef1.

3.4 Bandlimited Energy measures

Certain frequency bands were selected for energy measurement (0-300Hz, 0-400Hz, 100-300Hz, 0-2000Hz, 2000-5000Hz, 600-2800Hz, 2000-3000Hz). The three narrow low-frequency bands were variants used to explore for an optimum low-frequency band, and the other four broad bands were suggested by various published filterbank based front-ends for speech recognition.

3.5 Formant Measures

A frame-by-frame formant analysis was conducted by using formant frequency ranges to select the maximum peak within each range. Published formant frequency ranges were subjected to a sensitivity check for the contemporary data used and were deemed adequate. Formant energy was equated to the area under the peak between adjacent minima, and formant bandwidth was equated to the ratio of area under the peak to height of the peak. This analysis was performed for the first three

formants and generated the nine parameters: 1st Formant frequency, bandwidth, and energy (F1, BWf1, Ef1), 2nd Formant frequency, bandwidth, and energy (F2, BWf2, Ef2), and 3rd Formant frequency, bandwidth and energy (F3, BWf3, Ef3).

3.6 Frequency difference measures

Two frequency difference measures were included to explore phonemic discrimination based on relative rather than absolute formant values. The excitation frequency to first formant frequency distance (FxtoF1) and first formant to second formant (F1toF2) have been found useful in certain phonetic analyses.

3.7 Similarity Length Measures

The conventional way to incorporate the temporal change of acoustic parameters in ASR is to use the first, and maybe second order time derivatives of the parameters. We have previously described (Davies and Millar, 1999) an alternative to parameter derivatives based on a measure of parameter similarity length (PSL). Rather than taking the change in parameter value over a fixed time interval, we have measured the time interval over which a parameter maintains its value within a given range. Thus the basic acoustic vector comprises the two elements: the instantaneous value of the parameter and the similarity length. The related global variable is the range over which the parameter value can change and still be considered "similar". This "similarity tolerance" value was optimised over all the acoustic-phonetic associations and a single compromise value of +-20% used for all acoustic parameters. A more detailed approach would use individually optimised values for each acoustic-phonetic context. The current system allows for composite acoustic vectors comprising several parameter values and their respective similarity lengths.

3.8 Parameter quantisation

Acoustic parameters are initially spread over differing value ranges. Frequencies are measured as a sample number between 1 to 256 corresponding to equally spaced frequency samples taken over a range of 0 to 5000Hz. Energy ratios, initially with floating point values in the range 0 to 1.0, are linearly scaled to an integer range of 0 to 1023.

Parameters are then non-linearly re-scaled to the range of 0 to 15 to create a 4 bit acoustic vector component. Empirical non-linear scales were pre-determined for each parameter such that, over the long term, their values distributed evenly across a set of hexadecile bins representing the full range of the parameter. In the current system a separate non-linear scaling was established for each individual speaker, and, given the extent of the ANDOSL data used, this amounted to deriving equiprobable bins from approximately 10 minutes of spoken phonemically-rich sentence data. The most successful algorithm tested produced equiprobable bins by progressively aggregating pairs of adjacent values with the lowest populations, reducing the parameter ranges incrementally from up to 1024 to 16 discrete values. The same procedure was applied to the quantisation of the PSL.

3.9 Introduction of Shape bits

In addition to the PSL value for each frame, an indication of the way in which similarity of parameter value is terminated at each of its extremes has been added. The PSL together with the similarity tolerance value define a rectangle in parameter-time space. One bit is used to indicate at which corner of the rectangle the parameter enters, and another bit is used to indicate at which corner the parameter leaves. By definition it must leave at one of the corners. These so-called "shape bits" can

indicate in minimal terms something of the immediate context of the observed temporal extent of the parameter as measured at one source synchronous frame.

4. ASSOCIATION MATRIX AND PHONETIC RANKING TABLES

The association matrix is a simple mechanism to enable the evaluation of the performance of such an innovative form of speech representation in the task of phoneme discrimination. It has its roots in the "signature table" technique devised by Samuel (1967) for representing and evaluating moves in a checker-playing computer program, and subsequently applied to the task of speech recognition (Thosar, 1973), and of structured analysis of speech variance (Millar and Wagner, 1983). The matrix comprises a two dimensional array of counters and is established by a single pass of phonemically labelled speech data. A counter in the body of the matrix is incremented when the phonemic label on its first dimension coincides with the binary representation of its acoustic vector on its second dimension. While this form of representation is very simple, it is particularly suited to our spectro-temporal acoustic vectors as the extra complexity required to encode temporal behaviour, as in hidden Markov models, is coded within the acoustic vector itself. It is acknowledged that this form of analysis can have very high memory requirements. Recent increases in the availability of large amounts of memory at low cost have made this approach attractive.

The strength of the acoustic-phonetic association for a particular acoustic vector, X, labeled by phoneme, P, was evaluated by simply taking the rank of the association or the number of other phonemes that have more often been associated with X in the speech data sample. The average value of the mean ranking of each phoneme against phonemically labelled data, as in the ANDOSL corpus, has been used throughout as feedback for the optimisation of otherwise fixed parameters.

The results displayed in this paper are derived from the application of these techniques to material selected from the ANDOSL data corpus. Four speakers were randomly extracted from the section of the corpus representing speakers of cultivated Australian English and in the age range 31 to 45 years.

4.1 Comparison of temporal representations

The PSL approach has been shown to give better acoustic-phonetic associations than parameter derivatives alone. This can be seen in figure 2 where the discriminative power of simply the first formant and its temporal characteristics expressed in both its time derivative and its PSL plus shape bits can be seen. It should be noted that the ranking values on the ordinate are for aggregated values of all levels of F1 and as such give a global picture but are in fact much higher that those achieved for an optimal range of the parameter value. This is clearly shown in Tables 1 and 2 where clustering of low-ranked phoneme associations can be seen.

F1 value	rank 1	rank 2	rank 3	rank 4
0	S	Z	dZ	tS
1	Z	d	g	S
2	m	n	N	j
3	j	n	N	m
4	i@	u:	j	i:
5	i@	u:	i:	j
6	i@	u:	U	I
7	U	o:	oi	u:
8	o:	e:	U	@:
9	@:	e:	b	oi
10	@:	E	e:	oi
11	u@	E	@:	e:
12	@u	O	ei	E
13	O	@u	ei	A
14	A	au	u@	V
15	a:	ai	au	V

Table 1 Phoneme rankings for F1 acoustic vector for a single speaker.

En1 value	rank 1	rank 2	rank3	rank4
0	a:	ai	au	V
1	u@	@:	oi	e:
2	u@	@:	e:	o:
3	o:	U	S	oi
4	o:	U	oi	S
5	U	o:	oi	S
6	u@	o:	S	i@
7	S	U	s	u:
8	i@	u:	I	tS
9	i@	I	tS	U
10	i@	I	u:	i:
11	i@	I	u:	N
12	I	u:	i@	i:
13	N	I	u:	i:
14	m	n	N	j
15	b	g	d	v

Table 2 Phoneme rankings for En1 acoustic vector for a single speaker.

Tables 1 an 2 show phoneme rankings for single parameter acoustic vectors for En1 and F1 generated from 200 sentences for one speaker. While we cannot expect strong clustering of associations for such simple acoustic vectors, the low frequency nasal energy (En1) in particular, displays obvious clustering.

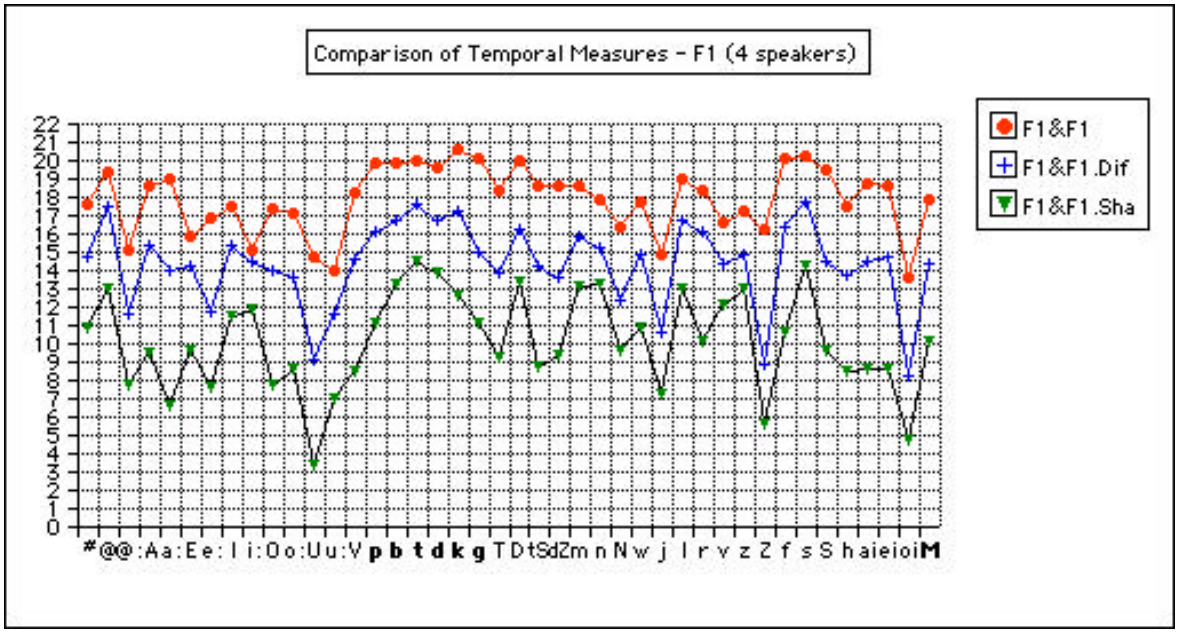

Figure 2 Aggregated phoneme discrimination rankings for all phonemes for the first formant parameter alone (upper trace), together with its first time differential (middle trace) or its similarity length and shape bits (lower trace).

For the nasal energy (table 2) the highest energies are for stops and nasals while open vowels associate with low nasal energy. The midrange (6-13) tend to be associated with high vowels.

In table 1, low F1 is associated with fricatives and stops, next we see nasals, then an approximate trend from close to the more open vowels and glides with increased F1. As expected, schwa and its transitions are located centrally.

4.2 Energy of Nasal Resonance

One aim of this work, partially motivating the source synchronous analysis, was to achieve rapid response times to changing frequency domain information. In Figure 1 the energy of the lowest nasal resonance (horizontal straight lines) can be seen to jump in value by approximately 150% between frames at the junction between the vowel 'V' and final consonant 'm' in the word "come". It has been observed that through sustained near closures this signal can fluctuate between frames in a manner that suggests a rapid mode switching between the excitation of nasal and oral resonances.

4.3 Interspeaker Comparison

Aggregated association strengths such as those in figure 2 were derived from data from four speakers. Figure 3 illustrates the interspeaker variation for F1 results.

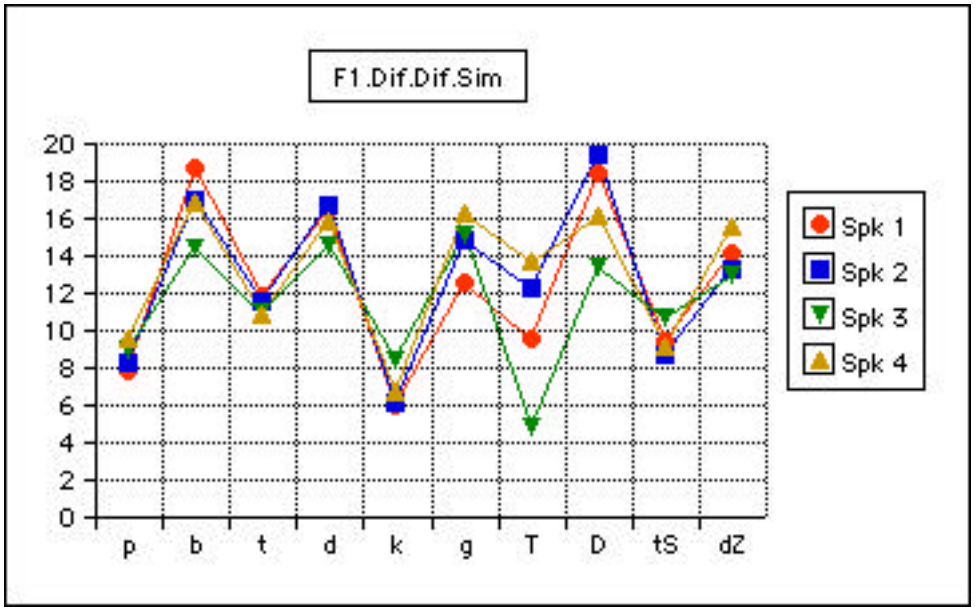

Figure 3 Comparison of stop and affricate rankings across 4 speakers. The parameter used is the similarity length of the second time derivative of the first formant frequency.

5. FUTURE DIRECTIONS

A current direction of this work is to analyse the ranking stream generated by emitting, for each frame in an utterance, the four top ranked phonemes. This stream can be aggregated over a phone length temporal window. The aim is to generate phoneme hypotheses for a word recogniser. In particular we are looking for evidence of increased nasalisation spreading beyond the labelled position of nasals or stops that could serve as a cue to the presence of phonemes that are not explicitly expressed.

6. CONCLUSIONS

This paper has described an innovative approach to the represention of time in speech by integrating it within source-synchronous acoustic vectors. The application of this approach to a set of one and two-dimensional acoustic vectors has illustrated some of the issues arising. A rapid evaluation method using the association of these vectors with phonemic labelling has enabled indications of some of the benefits of this approach. Sample results of this form of processing have been demonstrated.

While such innovation may not lead to immediate global improvements in the performance of speech technology, it is hoped that a consideration of those areas where a some clear advantage can be gained will provide a planning vector towards innovative combinations of ideas that will take us beyond the current speech processing plateau on which we have been sitting for most of the past decade.

The next steps forward will include attempts to integrate some of the ideas in this work into more mainstream speech processing. This will be characterised by a shift from essentially "clock-on-the-wall" temporal processing to a data-driven sequence processing.

7. REFERENCES

Ahlbom,G., Bimbot,F., Chollet,G. (1987) Modeling spectral speech transitions using temporal decomposition techniques, *Proc of ICASSP-87*, pp.13-16.

Atal,B.S. (1983) Efficient coding of LPC parameters by temporal decomposition, In *Proceedings of ICASSP'83*, pp.81-84.

Bilmes,J. (1998) Maximum mutual information based reduction strategies for cross-correlation based joint distribution modelling, *Proc of ICASSP98*, pp.469-472

Bimbot,F., Atal,B.S. (1991) An evaluation of temporal decomposition, In *Proceedings of EuroSpeech'91*, pp.1089-1092.

Davies,D.R.L., Millar.J.B. (1996) The evaluation of a computationally efficient method for generating a voiced-source synchronised timing signal, In *Proc. 6th Australian International Conference on Speech Science and Technology,* Adelaide, December, pp.527-532.

Davies,D., Millar,J.B. (1999) Evaluating Representations of Segment Level Dynamics in Acoustic-Phonetic Mapping, In *Proceedings of 14th International Congress of Phonetic Sciences (ICPhs'99),* 1-7 August, San Francisco, Volume 2, pp.1105-1108.

Deng,L. (1994) Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states, *IEEE Transactions on Speech and Audio Processing*, Vol.2, pp.507-520.

Deng,L. (1997) Speaker-independent phonetic classification using hidden Markov models with mixtures of trend functions, *IEEE Transactions on Speech and Audio Processing*, Vol.5, pp.319-324.

Fourakis,M. (1991) Tempo, stress, and vowel reduction in American English, *JASA* Vol.90, pp.1816-1827.

Ghaemmaghami,S., Deriche,M., Sridharan,S. (1998) Hierarchical temporal decomposition: a novel approach to efficient compression of spectral characteristics of speech, In *Proceedings of ICSLP'98*, pp.2567-2570.

Goertzel,G. (1958) An algorithm for the evaluation of finite trignometric series*, American Mathematics Monthly*, Vol.65, pp.34-35.

Hess,W. (1983) *Pitch determination of speech signals*, Springer-verlag: Berlin.

Holmes,J.N., Mattingly,I.G., Shearme,J.N. (1964) Speech synthesis by rule, *Language and Speech*, Vol.7, pp.127-143.

Lindblom,B. (1963) Spectrographic study of vowel reduction, *JASA* Vol.35, pp.1773-1781.

Millar,J.B., Wagner,M. (1983) The automatic analysis of acoustic variance in speech, *Language and Speech*, Vol.26, Part 2, pp.145-158.

Millar,J.B.,Vonwiller,J.P.,Harrington,J.M.,Dermody,P.J. (1994) The Australian National Database Of Spoken Language, In *Proceedings of ICASSP'94*, Vol.1, pp.97-100.

Millar,J.B., Harrington,J., Vonwiller,J. (1997) Spoken Language Data Resources for Australian Speech Technology, *Journal of Electrical and Electronics Engineers, Australia*, Vol.17, No.1, pp.13-23.

Nooteboom,S.G., Doodeman,G.J.N. (1980) Production and perception of vowel length in spoken sentences, *JASA* Vol.67, pp.276-287.

Samuel,A.L. (1967) Some studies in machine learning using the game of checkers, *IBM Journal of Research and Development*, Vol.11, pp.601-617.

Seneff,S. (1996) Comments on: "Towards increasing speech recognition error rates" by H.Bourlard, H.Hermansky, and N.Morgan, *Speech Communication*, Vol.18, pp.253-255.

Thosar,R.B. (1973) Recognition of continuous speech: Segmentation and classification using signature table adaptation, *Artificial Intelligence Memo AIM-213*, Computer Science Department, Stanford University, California, USA.

Van Dijk-Kappers (1988) Comparison of parameter sets for temporal decomposition of speech, *IPO-AR*, Vol.23, pp.24-33.

Van Son,R.J.J.H., Pols,L.C.W. (1992) Formant movements of Dutch vowels in a text, read at normal and fast rate, *JASA* Vol.92, pp.121-127.

Yang,H.H., Van Vuuren,S., Sharma,S., Hermansky,H. (2000) Relevance of time-frequency features for phonetic and speaker-channel classification, *Speech Communication* 31:35-50.