A SYNTACTIC PATTERN RECOGNITION METHOD FOR THE AUTOMATIC LOCATION OF POTENTIAL ENHANCEMENT REGIONS IN RUNNING SPEECH

Mark HUCKVALE

Abstract

A method for the automatic location of phonetically important events in continuous speech signals is presented. The events chosen are plosive bursts, fricatives, nasals and vocalic onsets and offsets. These events have been chosen to match the requirements of a system that enhances these regions to make the signal more robust to channel degradation. The technique uses a syntactic pattern recognition approach distinguished from automatic speech recognition or from knowledge-based approaches. Performance on read speech from many talkers was approximately 20% misses and 20% false alarms. This performance seems broadly similar to other published techniques, although direct comparisons on equivalent tasks have yet to be undertaken.

1. Introduction

This paper presents a syntactic pattern recognition approach to the issue of finding phonetically important events in continuous speech signals. The approach is distinguished from speech recognition in that it locates events in the signal rather than transcribes or explains the signal as a sequence of phonetic or phonological segments. However, it is distinguished from feature detection through its use of sequential constraints applied to the detected events.

The events chosen relate to vocalic onsets and offsets, nasals, fricatives and bursts. The context for this work is the possibility for speech enhancement demonstrated by Hazan and Simpson (1996). In their work, particular acoustic-phonetic events in the speech signal are targeted for selective enhancement prior to degradation of the entire signal through some channel. Hazan and Simpson have shown that selective reinforcement of bursts and vocalic onsets can provide significant improvements to the intelligibility of the subsequently degraded speech signal, even for the same overall signal-to-noise ratio.

The scenario in which this method of enhancement operates involves a speaker in a quiet environment with a good microphone communicating over a band-limited and distorting channel to a listener in a potentially noisy environment. The task is to manipulate the clean signal such that intelligibility is maximised for the listener, within the limits of maintaining the natural quality of the speech as far as possible. This latter requirement is due to the well known fact that infinite clipping of the speech signal preserves intelligibility in very severe listening conditions, although at the expense of a severe loss of naturalness.

For this speech enhancement technique to be useful in real applications, a number of hurdles need to be overcome: it needs to work on connected speech from any talker, it needs to work in a range of noise conditions and channel distortions, and the locations of the signal regions to enhance need to be found automatically. This paper is a first attempt at the last issue: the automatic location of the regions of the signal for which selective enhancement might yield improved intelligibility. We call these Potential Enhancement Regions (PERs). In the work so far reported by Hazan and Simpson, these PERs have been identified manually, by inspection of the signal.

In this paper I describe the characteristics of five potential enhancement regions, contrast the proposed detection approach to two alternative methods, and describe the implementation and

evaluation of the method on a medium sized database of read speech from a number of talkers.

2. Potential Enhancement Regions

For the current enhancement work, we define five signal regions that have potential for enhancement: bursts, vocalic onsets, vocalic offsets, fricatives and nasals. A description of each follows:

2.1 Bursts (BUR)

Bursts are short regions of turbulent noise generated when the pressure built up behind oral stops is suddenly released. They are interesting for enhancement in that they are short, may be readily masked, but yet carry important place of articulation information. Enhancement of the salience and spectral properties of bursts could lead to improved consonant identification within the manner class for stops.

2.2 Vocalic Onsets (ONS)

The transition region from obstruent or nasal to vowel contains many important transitory features relevant to the identity of the consonant: particularly the F1 transition for manner of articulation cues and F2 & F3 transitions for place of articulation cues. Since these regions are short and the transitions rapid, an increase in the salience of such regions could contribute to an increase in the identification of the preceding consonant. For current purposes, we do not separately locate approximants and instead merge them with the syllable nucleus. Thus the vocalic onset in the sequence [bla-] occurs between the [b] and the [l].

2.3 Vocalic Offsets (OFS)

The transition regions from vowel to obstruent or nasal also contains important transition cues suitable for enhancement. An increase in the salience of these regions could lead to an increase in the identification of the following consonant. Here too, we consider approximants to be part of the syllable nucleus, so in the sequence [ilz] the offset occurs between the [l] and the [z].

2.4 Fricatives (FRC)

When obstructions in the oral cavity cause turbulence to occur, as in fricatives and affricates, then the spectral characteristics of that turbulence gives important cues to the place of articulation. Since such regions are often quiet with respect to the vowel intensity and also easily masked by background noise, an increase in their salience could improve their identification within the manner class for fricatives.

2.5 Nasals (NAS)

Nasal consonants are often quiet and rather similar in sound even for different places of articulation. They also tend to be confused with other quiet voiced sounds, particularly voiced fricatives. Thus an increase in the salience of such regions could lead to improvements in the identification of both the place and manner for nasals.

3. Methods for Automatic Location

We identify three basic approaches to the problem of finding PERs in running speech. In the first, the knowledge-based approach, special rules are constructed from observations of the speech signal which operate on the acoustic properties of the signal in various frequency bands. Using hand-set thresholds the rules form a context-sensitive logic to perform the detection of significant phonetic events. In the second, the speech recognition approach, a phonetic transcription of the signal is

attempted, using as much phonetic and phonotactic information as possible, and the resulting transcription forms the input to a set of mapping rules which then identifies the events of interest. In the third approach, a broad-class transcription of the signal is attempted, with the unit inventory chosen to facilitate the location of events. It is this third approach that forms the basis for the system implemented and tested in this paper. In the following paragraphs we give further descriptions of the three approaches.

3.1 Knowledge-based methods

A recent example of the *Knowledge-based approach* is the landmark detection system of Liu (1996). This system, designed as one component in a speech recognition system based on acoustic-phonetic events rather than statistical modelling, locates so-called 'landmarks' in connected speech signals. These landmarks are defined as times when the underlying phonological description is most clearly realised in the signal. Regions of the signal around these landmarks, once detected, could then act as input to a classification system to drive lexical access. Liu's work describes only the landmark detection procedure for a subset of the events required of a complete recognition system. Her two landmark classes: abrupt and abrupt-consonantal cover quite closely the PERs burst, onset and offset. On the other hand, the transitions between consonantal types (e.g. nasal-to-fricative) are landmarks but not PERs. Nasal and fricated regions are segmented by landmarks from the surrounding signal in Liu's system, but not separately classified as required for enhancement.

The reported performance of Liu's system on a limited subset of landmarks was generally good on a small database of read speech from the TIMIT database. Error rates (misses & substitution errors) were about 10% with false-alarm rates of about 15%. The majority of landmarks were located within 30ms of hand-edited annotations. In future work we shall need to make direct comparisons between Liu's procedure and the system described in this paper on the same data. In the meantime, Liu has provided a target performance figure for us to aim at.

3.2 Phone recognition

One of the highest performance *phone transcription* systems that has been demonstrated on standard data, is the recurrent neural network phone recognition system of Robinson (1994). In this system, a recurrent neural network taking cepstral coefficients as input is trained to produce a set of 39 phone probabilities at each 10ms frame. A viterbi decoder applied to the output of the system parses the network output. When operating as a speech recognition system, the decoder employs a statistical language model built on a bigram or trigram word model and a lexicon. However the system can also operate as a phone transcription system using a statistical language model based on N-grams of phones only. With a phone language model, the system has been shown to be able to provide a phone accuracy of approximately 80% on TIMIT data. A system such as this could provide a means to identify the location of PERs. The recognised phone string, time aligned with the input signal, could be used to derive the type and location of PERs.

In a phone system, it is likely that PER detection accuracy would be higher than the phone accuracy. This is because many of the phone errors would be within a PER class, for example within the class of voiceless plosives which are all labelled with the PER 'burst'. On the other hand, the influence of the phone language model may cause many false alignments with the signal, with units required in the language model 'forced' to be inserted into the recognised string. This in turn might lead to inappropriate enhancements of the signal. A target performance of 10% errors and 10% false alarms looks achievable on the TIMIT data.

3.3 Special syntactic pattern recognition

In contrast to the previous two approaches, a *broad class transcription* approach could be tuned to be sensitive to the acoustic-phonetic requirements of enhancement, while at the same time exploit sequence constraints of the broad class units. The PERs are seen as either broad-class segments or as located at the transitions between broad class segments. In comparison with Liu's system, a syntactic pattern recognition approach allows us to set constraints on the density and ordering of events. While Liu's system, for example, could generate more vocalic onsets than offsets, a syntactic recognition system ensures that equal numbers of onsets and offsets are detected. In contrast to Robinson's system, a special purpose broad-class system can be tuned to the characteristics of the problem: we have freedom to set the acoustic parameters, the complexity of the acoustic modelling, the inventory of units and the sequence constraints. Ideally we would want to ensure that where the system gave errors, these did not reduce the intelligibility of the resulting enhancement. This might mean we would want to suppress false alarms at the expense of increasing the miss rate.

4. Syntactic Pattern Recognition Method

4.1 Syntactic modelling

In the current system, we define the speech signal as being generated by a sequence of five types of acoustic components: silence (SIL), vocalic (VOC), frication (FRC), burst (BUR) and nasal (NAS). Each of these acoustic components is modelled with a three-state hidden Markov model. These models crudely approximate the spectral development that typically occurs in the different regions. At recognition, the signal is 'explained' as a sequence made up from this inventory of five types. It is possible to apply constraints to the recognised sequence according to (i) the transition probabilities between components, and (ii) absolute constraints on longer sequences arising from English phonotactics. The first set of constraints acknowledges the fact that, for example, utterances start with vocalic regions or fricatives more commonly than they start with bursts or nasals, or that burst to nasal transitions are much less common than burst to vocalic regions. The second set of constraints allow us to specify, for example, that long sequences of bursts and fricatives alone are not possible: they must be divided up with vocalic regions according to the sequential rules of English phonology. Thus the sequences BUR FRC BUR FRC or VOC FRC NAS BUR do not occur according to the rules of English phonotactics, although connected speech effects and the use of syllabic consonants might make them possible.

From the recognised broad-class transcription, the BUR, FRC and NAS PERs are found from the BUR, FRC and NAS regions; and the ONS and OFS PERs are identified as the first and last 30ms of each VOC region.

4.2 Measures of performance

Conventional measures of recognition performance used in speech recognition systems may not be appropriate for the evaluation of PER location systems. In particular, there is no requirement to correctly identify every region of the signal providing that the PERs are found. On the other hand it is important that the PERs are found at the correct time - the accuracy of their temporal alignment might be important for subsequent processing.

To address this issue, an evaluation metric was devised which compares the recognised annotation sequence with a reference annotation sequence in terms of the PER location only.

The process is as follows: for each reference PER, the nearest PER of the correct type is identified in the recognised sequence. This is done using the time of the middle of the reference and recognised PER. Each reference and each recognised PER has only a single nearest neighbour.

From this alignment, if the two PERs start within 25ms of each other **or** they overlap in time by more than 50% then this is called a 'hit' else a 'miss'. The overlap criteria ensures that a recognised PER cannot be scored as a hit more than once, while the 25ms criteria allows for some temporal mis-alignment for very short regions. Subsequently, each recognised PER is taken in turn and the mirror process is applied to identify the number of false alarms.

4.3 Use of SFS and HTK

These experiments were conducted on an IBM-PC compatible computer using the Speech Filing System software (version 3.11) from University College London and the Hidden Markov Modelling Toolkit (version 1.31) from Cambridge University. Both of these software packages may be downloaded free of charge by anonymous FTP to pitch.phon.ucl.ac.uk in the directory /pub/sfs.

5. Experiments

5.1 Database

These experiments were undertaken using extracts from the SCRIBE corpus. This material consists of read passages and sentences from a small number of different speakers of British English recorded in an anechoic room using a high quality microphone. They are therefore rather unrepresentative of the speech material for which the enhancement procedures themselves are aimed. On the other hand its is necessary to establish the best methods for PER location and starting with high quality signals will at least mean that deficiencies in location can be attributed to the location algorithm rather than the recording conditions.

Another advantage of the SCRIBE material is the very detailed set of time-aligned phonetic annotations available. This allows for the automatic evaluation of the performance of the location methods. The annotation scheme used for the SCRIBE data uses over 480 different labels for phonetic events. These can largely be divided up into phonetic classes according to the system of the International Phonetic Association, combined with diacritic codes for voice quality, nasalisation, aspiration, frication and velarisation. The general principle followed was to assign each phonetic class to the most directly appropriate broad class: vowels and approximants to the VOC class, nasal consonants to the NAS class, fricatives to the FRC class, plosives to the BUR class. Corrections were then made in the light of experience with the system, with each vowel label having the diacritic fricated being assigned to the FRC class. Utterance initial and final silences and epenthetic silences (but not stop gaps) were assigned to the SIL class.

The actual files from the SCRIBE database used for training and testing are listed in the Appendix.

5.2 Varying number of acoustic parameters

For this modelling task, where only a broad-class phonetic labelling of the signal is required, it is first necessary to establish the degree of spectral detail required to give good labelling performance. One way of investigating this is to use representations of the short-term amplitude spectrum of the signal, stored to various degrees of detail. Although it is possible to do this with a filterbank analysis, where the number of filters can be changed to increase spectral resolution, a much simpler approach is to perform the discrete cosine transform of the spectrum (or 'cepstrum') and then to choose the first N coefficients, varying the size of N. Small values of N give a small vector representing only the basic spectral shape, while large values of N model all of the fine spectral detail. Typically N is less than 16, since it is found empirically that higher cepstral coefficients do not contain useful information as far as speech recognition is concerned.

In this experiment, so-called "mel-scaled" cepstral coefficients (MFCC) were used. These are calculated by simulating a filterbank with increasing bandwidth filters prior to the discrete cosine transform. Spectral magnitudes are fed through a bank of triangular shaped filter responses with a spacing and a bandwidth based on a mel scaling of frequency - so that their bandwidth and spacing increases logarithmically after about 1000Hz. Cepstral coefficients are then calculated from these frequency warped spectral magnitudes. Coefficients are calculated from a 20ms window on the signal each 10ms. In addition, the total energy of the frame is calculated in bels and appended to the vector input to the recogniser.

Training: 5 3-state HMMs representing bursts (BUR), frication (FRC), nasality (NAS), vocalic regions (VOC) and silence (SIL) were initialised and re-estimated using the Baum-Welch algorithm. Recognition was performed with a null language model using bigram transition probabilities estimated from the training data (and using a floor of p=0.001).

Training and evaluation was performed using the following four sets of acoustic parameters:

- energy + first 3 cepstral coefficients (mf4)
- energy + first 7 cepstral coefficients (mf8)
- energy + first 11 cepstral coefficients (mf12)
- energy + first 15 cepstral coefficients (mf16)

The results are presented in terms of the error rate and the false-alarm rate for each of the PERs as calculated using the metric described in section 4.2.

	BUR (N=2057)		FRC (N=2960)		NAS (N=1282)		ONS (N=4106)		OFS (N=4106)	
	E%	F%	E%	F%	E%	F%	E%	F%	E%	F%
mf4	52	31	42	17	42	48	46	15	54	28
mf8	53	33	44	16	30	35	40	16	48	27
mf12	51	33	44	15	25	27	36	16	42	24
mf16	49	33	42	15	24	26	35	16	40	23

Table 1. Error rates as a function of number of cepstral coefficients

The table shows a clear improvement in three of the five PERs with increasing spectral detail. The performance on bursts is very poor, with not only a high miss rate, but also a high false-alarm rate. Furthermore, an increase in spectral detail does not seem to have a large effect. What seems to be happening is that fricatives and burst are inadequately separated and many confusions occur. An increase in spectral detail alone does not help differentiate these two classes - this makes sense in that it is not spectral shape that distinguishes a burst from a fricative, but its sudden appearance and its short duration.

5.3 Dynamic Features

To improve the differentiation between bursts and fricatives, it is necessary to add in to the input vector acoustic information which is different in the two classes. One obvious candidate is information about the rate of change of the spectral coefficients: we expect that bursts, being

sudden and transitory, will have larger spectral rates of change at onset and onset than fricated regions. Indeed, the use of transitional information was also key to the success of Liu's work reported earlier.

In this experiment, each frame of energy and cepstral coefficients is extended with 'delta' coefficients calculated by regression of each parameter over a 40ms window centred on the frame:

- energy + delta energy, 3 cepstral coefficients + 3 delta cepstral coefficients (mf4d)
- energy + delta energy, 7 cepstral coefficients + 7 delta cepstral coefficients (mf8d)
- energy + delta energy, 11 cepstral coefficients + 11 delta cepstral coefficients (mf12d)
- energy + delta energy, 15 cepstral coefficients + 15 delta cepstral coefficients (mf16d)

The training and testing was undertaken as before; error rates and false alarm rates were as follows:

	BUR (N=2057)		FRC (N=2960)		NAS (N=1282)		ONS (N=4106)		OFS (N=4106)	
	E%	F%	E%	F%	E%	F%	E%	F%	E%	F%
mf4d	33	24	32	15	36	42	37	16	44	24
mf8d	34	23	31	20	22	36	30	17	33	21
mf12d	32	25	31	20	16	30	25	17	28	20
mf16d	30	25	31	18	14	30	25	18	27	20

Table 2. Error rates as a function of number of cepstral and delta-cepstral coefficients

The incorporation of deltas has had dramatic effect: 20% more bursts are being detected, and there is about a 10% increase in detections of the other units. As before, the increase in the number of cepstral coefficients has significant effect on the recognition rate of three of the five PERs. The effect of deltas on the detection of onsets and offsets was not expected, since these are not recognised directly, but only inferred from the starting and ending times of recognised vocalic regions. Apparently the deltas have improved the temporal accuracy of the edges of the vocalic labels.

The introduction of deltas has improved things but absolute performance is still poor. The high false-alarm rate for nasal regions provides a clue: it appears that any quiet voiced region is being labelled as a nasal. Similarly too many vocalic onsets are being preceded by burst regions. These give us an indication that the quality of acoustic modelling is still rather poor.

5.4 Using gaussian mixtures

The poor quality of acoustic modelling arises from the fact that we have such a small set of broad classes, each covering a range of phonetic events: the VOC class in particular has to model all vowels, diphthongs and approximants.

One way to accommodate such variability within a broad class model is to use a more flexible method for representing the probability distribution of each acoustic parameter on each state of the HMM. Since the simplest approach is to expect a normal (gaussian) distribution with a single mean and variance for each parameter for each state, the next simplest is to expect a mixture of normal (gaussian) distributions, each with its own mean and variance.

The best performing models from the experiment described in 5.3 with cepstral and delta-cepstral coefficients were used to create multiple mixture HMMs which were further re-estimated using the Baum-Welch algorithm.

- energy + delta energy, 11 cepstral and 11 delta-cepstral coefficients modelled with 2 mixtures (mf12dm2)
- energy + delta energy, 11 cepstral and 11 delta-cepstral coefficients modelled with 3 mixtures (mf12dm3)
- energy + delta energy, 11 cepstral and 11 delta-cepstral coefficients modelled with 4 mixtures (mf12dm4)
- energy + delta energy, 11 cepstral and 11 delta-cepstral coefficients modelled with 5 mixtures (mf12dm5)

- energy + delta energy, 15 cepstral and 15 delta-cepstral coefficients modelled with 2 mixtures (mf16dm2)
- energy + delta energy, 15 cepstral and 15 delta-cepstral coefficients modelled with 3 mixtures (mf16dm3)
- energy + delta energy, 15 cepstral and 15 delta-cepstral coefficients modelled with 4 mixtures (mf16dm4)
- energy + delta energy, 15 cepstral and 15 delta-cepstral coefficients modelled with 5 mixtures (mf16dm5)

Results were as follows:

	BUR (N=2057)		FRC (N=2960)		NAS (N=1282)		ONS (N=4106)		OFS (N=4106)	
	E%	F%	E%	F%	E%	F%	E%	F%	E%	F%
mf12dm2	27	20	27	12	13	30	24	16	27	19
mf12dm3	27	20	26	11	11	29	24	16	26	19
mf12dm4	25	21	24	12	10	27	24	16	26	18
mf12dm5	25	21	25	12	10	27	23	16	25	19
mf16dm2	26	21	27	12	13	29	24	16	26	18
mf16dm3	26	20	26	11	11	27	24	16	26	18
mf16dm4	24	20	24	11	11	26	24	17	25	18
mf16dm5	23	20	25	11	11	26	23	16	26	19

Table 3. Error rates as a function of number of cepstral and delta-cepstral coefficients using 2 mixture models

The incorporation of mixtures makes significant improvements to the recognition of the bursts and fricatives, with little improvement elsewhere. Simply adding a single mixture to the burst and fricative models increases the number of detected events by 5% and reduces the number of false alarms by a similar amount. There are 1% and 2% changes elsewhere. There is a slight trend for performance to increase with an increasing number of mixtures, with the system mf16dm5

providing a mean error rate (equally weighted over the five classes) of 21.4% and a mean false alarm rate of 18.6%.

5.6 Grammar

Finally, the use of phonotactic constraints was tried, using a model of English syllable structure. The aim was to see if constraints on longer sequences of PERs in combination with the bigram probabilitites would benefit performance. To do this, each utterance was modelled as consisting of syllables separated by optional silence, with each syllable constrained to be:

[FRC] [BUR | NAS] VOC [NAS] [FRC] [BUR] [FRC]

That is: starting with an optional FRC region, followed by either a BUR or a NAS region, a mandatory VOC region, an optional NAS region, an optional FRC region, an optional BUR region, and ending with another optional FRC region.

The following, best scoring systems were tried using the application of the language model:

- energy + delta energy, 11 cepstral and 11 delta-cepstral coefficients modelled with 4 mixtures, syllable language model (ml12dm4)
- energy + delta energy, 11 cepstral and 11 delta-cepstral coefficients modelled with 5 mixtures, syllable language model (ml12dm5)
- energy + delta energy, 15 cepstral and 15 delta-cepstral coefficients modelled with 4 mixtures, syllable language model (ml16dm4)
- energy + delta energy, 15 cepstral and 15 delta-cepstral coefficients modelled with 5 mixtures, syllable language model (ml16dm5)

Results were as follows:

	BUR (N=2057)		FRC (N=2960)		NAS (N=1282)		ONS (N=4106)		OFS (N=4106)	
	E%	F%	E%	F%	E%	F%	E%	F%	E%	F%
ml12dm4	25	18	25	11	11	25	21	17	24	20
ml12dm5	25	19	25	11	11	25	21	17	23	20
ml16dm4	24	19	25	11	12	24	21	18	23	20
ml16dm5	23	19	25	11	12	25	21	18	24	21

Table 4. Error rates as a function of number of cepstral and delta-cepstral coefficients and numbers of mixtures using the syllable language model.

The effect of using a syllable grammar, which provided extra constraints on recognised sequences was small and mixed. The most significant improvement was an increase in the number of detected vocalic onsets by about 2%. Elsewhere, performance changed up and down by about 1%. Overall, the system ml16dm5 has a mean error rate of 20.9% and a mean false alarm rate of 18.5%, fractionally better than system mf16dm5.

6. Conclusions

In the preparatory work for an automatic enhancement system described in this paper, we have shown that a broad-class transcription system can locate a set of five important phonetic events with a mean hit rate of about 80% and a mean false alarm rate of about 20%. These rates seem poorer than those obtained by Liu or than those expected from Robinson's system, although the data sets are not the same. One reason for this might be the scoring criterion which has been rather strict in this work, another might be the differences in the system of annotation.

The performance of the system against an independent set of annotations is actually a difficult test, since inevitably the labels tend to reflect the underlying phonological transcription of the utterance rather than the surface form. For example, if a phonological plosive is realised as a fricative then we want our system to label it with FRC and not BUR. However to use annotations of only the regions satisfactory for enhancement would require a laborious labelling task and leave us open to the criticism that we had selected events to maximise location performance. A future exercise will be to directly compare the three approaches on the same data with the same scoring metric.

Certainly, the look of the PERs located by the system on a piece of test data seems satisfactory (see Fig 1). Until the PER location system is integrated with the complete enhancement system, we will not be able to tell whether it meets the goals of the overall project. Our next most important test will be to use the system to enhance spontaneous speech for a real task in difficult listening conditions.
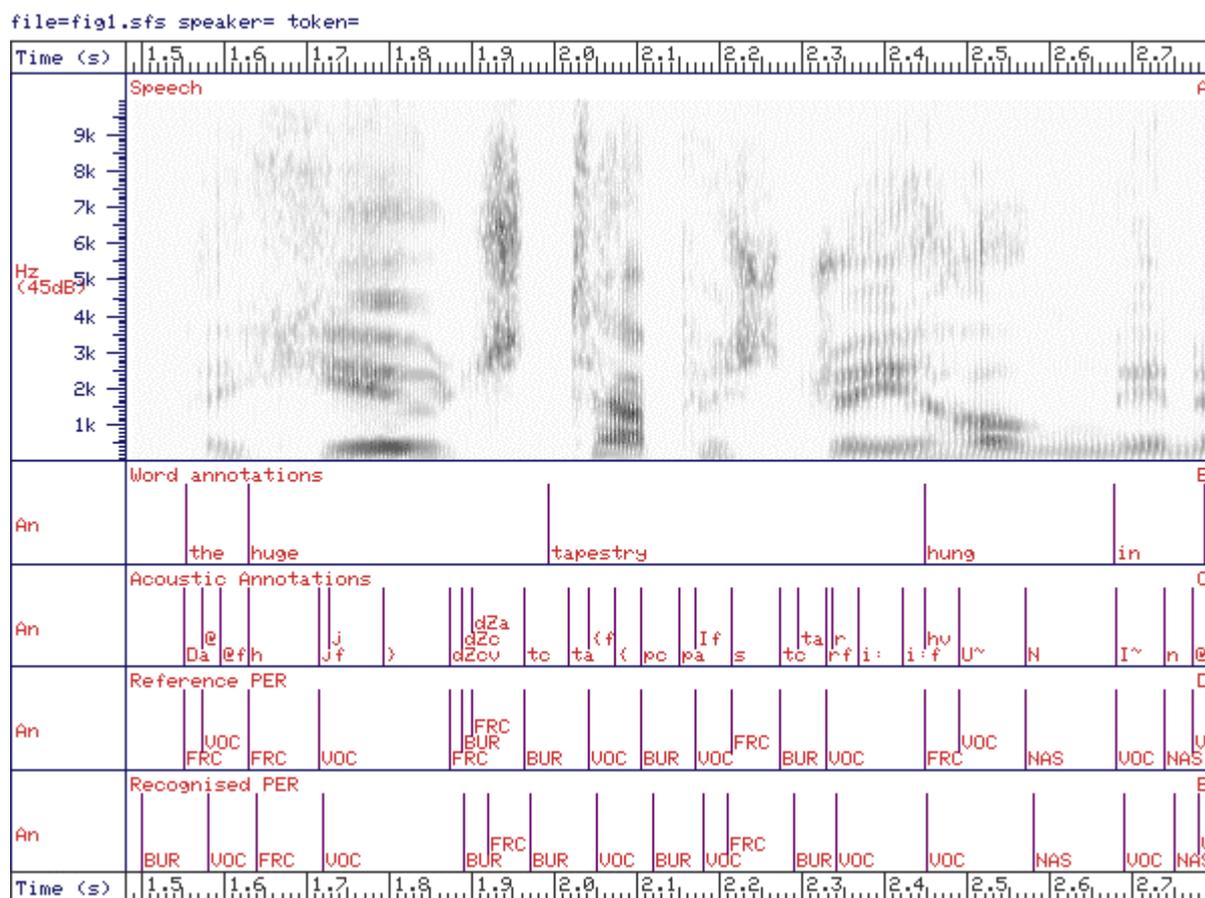

Figure 1. Example recognition of test file WASB0002 using system ml16dm5. A=Speech signal, B=Manually identified words, C=Manually identified acoustic events, D=PER locations found from manual annotations, E=PER locations found automatically by the recognition system in this paper.

References

Hazan, V., & Simpson, A., (1996) Cue-enhancement strategies for natural VCV and sentence materials presented in noise, Speech, Hearing and Language - Work in Progress, Phonetics and Linguistics, University College London, Vol <u>9</u> pp43-55.

Liu, S.A., (1996) Landmark detection for distinctive feature based speech recognition, J.Acoust.Soc.Am. <u>100</u> pp3417-3430.

Robinson, A., (1994) An application of recurrent neural nets to phone probability estimation, IEEE Trans. Neural Networks <u>5</u>.

Appendix

A.1   SCRIBE files used for training and testing

The training files were:

AAPA0001 AAPA0003 AASA0001 AASB0001 ACPA0001 ACPA0003 ACSA0005 ACSB0005 AEPA0001 AEPA0003 AESA0009 AESB0009 AFPA0001 AFPA0003 AFSA0011 AFSB0011 AHPA0001 AHPA0003 AHSA0015 AHSB0015 AMFTA001 AMFTA003 AMFTA005 AMFTA007 AMFTA009 AMFTA011 AMFTA013 AMFTA015 AMFTA017 AMFTA019 AMFTB001 AMFTB003 AMFTB005 AMFTB007 AMFTB009 AMPA0001 AMPA0003 GAPA0001 GAPA0003 LGFTA021 LGFTA023 LGFTA025 LGFTA027 WAPA0001 WAPA0003 WASA0001 WASB0001

The testing files were:

AAPA0002 AAPA0004 AASA0002 AASB0002 ACPA0002 ACPA0004 ACSA0006 ACSB0006 AEPA0002 AEPA0004 AESA0010 AESB0010 AFPA0002 AFPA0004 AFSA0012 AFSB0012 AHPA0002 AHPA0004 AHSA0016 AHSB0016 AMFTA002 AMFTA004 AMFTA006 AMFTA008 AMFTA010 AMFTA012 AMFTA014 AMFTA016 AMFTA018 AMFTA020 AMFTB002 AMFTB004 AMFTB006 AMFTB008 AMFTB010 AMPA0002 AMPA0004 GAPA0002 GAPA0004 LGFTA022 LGFTA024 LGFTA026 LGFTA028 WAPA0002 WAPA0004 WASA0002 WASB0002

The pressure microphone signals (SES and PES) were used, and the set of acoustic annotations (SEA and PEA). Please note that a number of the annotation files have alignment errors in the distributed database. Please contact the author for the corrected annotations.