ACOUSTIC CONFIDENCE MEASURES FOR SEGMENTING BROADCAST NEWS

J. Barker, G. Williams and S. Renals

Department of Computer Science, University of Sheffield
Sheffield S1 4DP, UK
j.barker, g.williams, s.renals@dcs.shef.ac.uk

ABSTRACT

In this paper we define an acoustic confidence measure based on the estimates of local posterior probabilities produced by a HMM/ANN large vocabulary continuous speech recognition system. We use this measure to segment continuous audio into regions where it is and is not appropriate to expend recognition effort. The segmentation is computationally inexpensive and provides reductions in both overall word error rate and decoding time. The technique is evaluated using material from the Broadcast News corpus.

1. INTRODUCTION

Most speech recognition tasks to date have required the recognition of discrete utterances over which both the speaker and channel characteristics remain constant. It is given that the data supplied to the recogniser is speech and so speech detection amounts to little more than trimming off leading and trailing silences. However, practical speech recognition systems cannot expect to be supplied with such pre-segmented data. Faced with an unsegmented stream of audio, from a radio broadcast for example, the first task that must be performed is to decide which regions contain speech and which regions do not.

Given that we accept the limitations of our speech recogniser, a pragmatic goal is not a segmentation into speech and non-speech, but rather into regions that are *recognisable* speech and those which are not. This second class not only contains non-speech audio, such as music, but also speech for which the acoustic conditions are such that the data is not sufficiently well matched by the models to produce a reliable recognition result. A related model based approach to speech detection is described in [1].

If this segmentation can be provided through the use of a purely acoustic confidence measure which is not dependent upon any particular decoding hypotheses (see section 2), it may be computationally inexpensive and computed before recognition is attempted. A segmentation system of this kind concentrates recognition effort exclusively upon regions where it may be usefully applied. The remainder of the paper describes the formation and application of such a confidence measure derived from local posterior probability estimates produced by the ABBOT Hidden Markov Model/Artificial Neural Network (HMM/ANN) large vocabulary continuous speech recognition (LVCSR) system [2].

This work was supported by ESPRIT Long Term Research Project 20077 (SPRACH) and by an EPSRC studentship.

2. ACOUSTIC CONFIDENCE MEASURE

A confidence measure may be defined as a function which quantifies how well a model matches a spoken utterance. Such a measure may be derived from the output of both the acoustic and language models, or from either model separately. An *acoustic* confidence measure is one which is derived exclusively from the acoustic model. The acoustic confidence measure employed here, $S(n_s, n_e)$, is the entropy of the K posterior phone probability estimates q output by a recurrent network averaged over an interval D [5]:

$$S(n_s, n_e) = -\frac{1}{D} \sum_{n=n_s}^{n_e} \sum_{k}^{K} F(q_k|\mathbf{x}^n) \log \left(F(q_k|\mathbf{x}^n) \right) \ . \quad (1)$$

where $\mathbf{x}$ is the acoustic data and the interval $D = n_e - n_s + 1$, with the start and end frames denoted n_s and n_e respectively.

In regions of the signal where the models provide a good match to the data, the distribution of phone posteriors will typically be dominated by a single phone class. Such a distribution has low entropy. However, during regions of non-speech, or poorly modelled speech, several alternative phone models may have roughly equal posterior probabilities, leading to a higher value of S. Ideally, there should be a clear distinction between the regions of well modelled speech where the value of S is low, and regions of poorly modelled speech and non-speech where the value is high. However, there are several factors that weaken the power of the measure.

Firstly, it is possible for certain models to be well matched to the data even during periods of non-speech. This is most obviously true for the silence model, but there are other phone models that might be closely matched to non-speech sounds - e.g background hiss can be mistaken for a sibilant such as *s*. Conversely there are certain *weak* phones that are often ambiguous even in clean, otherwise well-modelled speech. For these experiments we compiled a list of weak phones containing: *ix*, *dx*, *uh*, *axr* and *ax*. By excluding frames which have the highest posterior probability of any of these phone classes, the power of the confidence measure can be increased.

Secondly, due to the piecewise stationary assumption, the per-frame entropy is inherently very noisy. Even in clean speech spikes occur in the entropy profile at regular intervals corresponding to predictably poorly modelled phone transitions. These spikes can easily obscure the underlying trends (see figure 1). However, by applying a median filter with a sufficiently short window (50–80ms) many of these spikes can be removed reducing the value of

S during the speech regions relative to that during the non-speech regions.

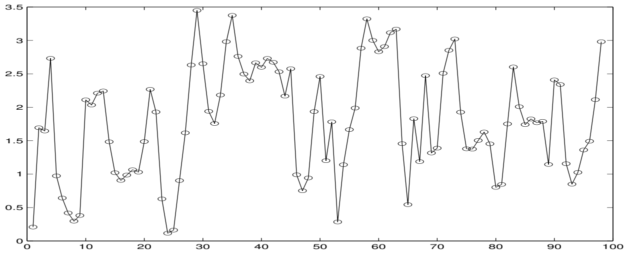

Figure 1: Per-frame entropy for the phrase "America in black and white". Although this is clean studio speech entropy spikes occur at each of the phone transitions.

3. SEGMENTATION

Figure 2 plots values of two versions of the entropy measure over a 10 minute segment of a radio broadcast. The values were calculated by removing silence and the weak phones and averaging over a 40 frame ($\approx$ 600ms) window. It can be seen that even after this averaging there remain rapid fluctuation. These were filtered out prior to segmentation using a further median smoothing stage. This final smoothing, shown in the lower panel, was performed over an approximately 10 second window. Segmentation was performed by locating local maxima or minima in the difference function and declaring these as segmentation points when their absolute value was over an empirical threshold.

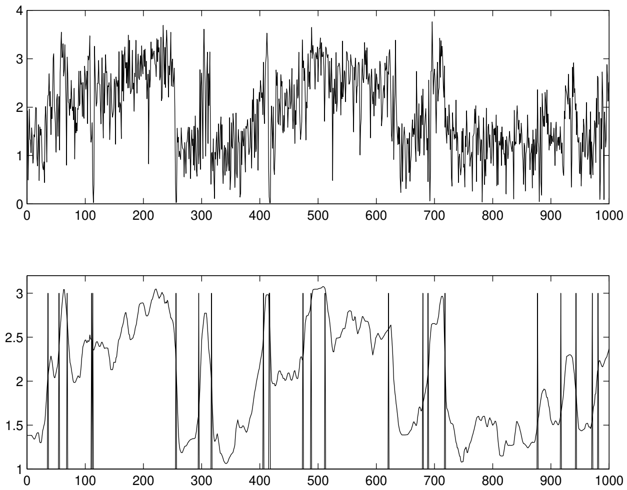

Figure 2: The raw entropy measure (top) and the smoothed and segmented entropy measure (bottom) for a 10 minute extract from a broadcast news program.

Although this procedure is able to provide segmentation points, the heavy smoothing of the entropy function causes these points to be positioned to within a few seconds of their correct location. Therefore, the locations must be 'fine tuned' before they can be used. This was accomplished using the KL2 distance metric, as described by Siegler et al. [3]: Means and variances were calculated for the distribution of the front-end processed acoustics (i.e. prior to any entropy calculations) in two second windows on either side of a putative segmentation point. The segmentation point was then adjusted so as to maximise the distance between the distributions.

4. CLASSIFICATION

Classification of the segments was based on the same acoustic confidence measure employed for the segmentation: The entropy over the phone posteriors was calculated for each frame of the segment and 5 frame ($\approx$ 80ms) median smoothing was applied to reduce the influence of phone transitions. Frames hypothesised as silence or as any of the weak phones were removed and the mean entropy value for the remaining frames was calculated. Low values were taken to indicate well modelled speech worthy of decoding and higher values to indicate poorly modelled speech and non-speech. A threshold was set to decide which segments to excise. A summary of the complete system is given in figure 3.

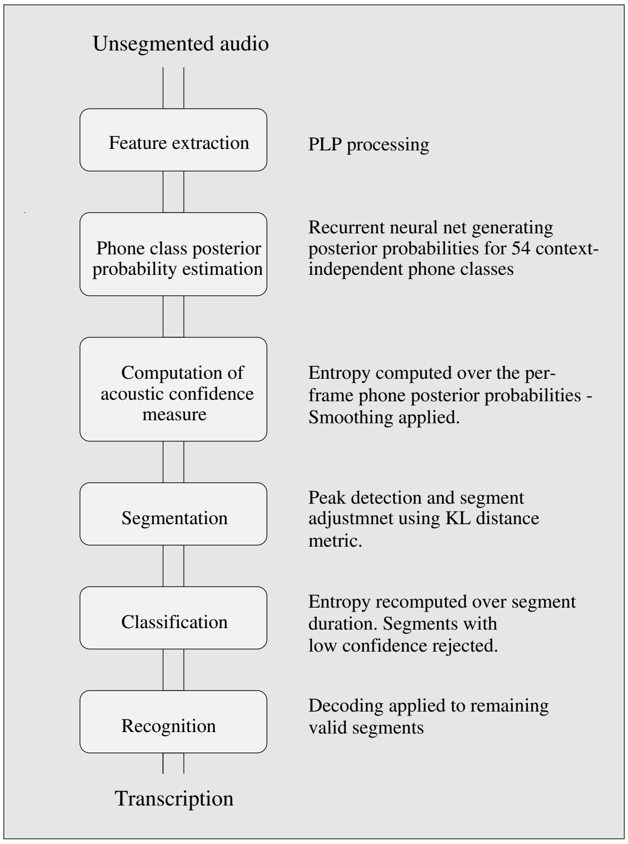

Figure 3: A summary of the segmentation system.

5. EXPERIMENTS

A 30 minute radio show[1] was selected from the the 1996 ARPA Broadcast News (BN) corpus [4] to evaluate the system. Candidate segments were obtained for the entire show, i.e. including commercial breaks not used in the Hub-4 evaluation, and decoded using the ABBOT HMM/ANN LVCSR system. Word error rates

[1] ABC Nightline: Episode 05/23/96.

(WERs) were calculated by aligning the decoded word sequences against a Viterbi alignment of the reference transcription[2]. The classification portion of the system was also evaluated using the 'focus condition' segmentation supplied with the BN corpus for comparison.

The acoustic model used for the experiments was composed of two recurrent networks with 604 context-dependent phone classes (plus silence). One network estimated the phone posterior probability distribution for each frame given a sequence of 12th order perceptual linear prediction (PLP) features. The other network performed the same distribution estimation with features presented in reverse order (since recurrent networks are time-asymmet The two probability estimates were averaged in the log domain. The model was trained on BN data drawn solely from the F0 condition. A 65k word backed-off trigram language model trained on 132 million words was used for the decodings.

6. RESULTS AND DISCUSSION

Table 1 shows the recognition performance for the supplied focus-condition segmentation. The table also shows the number of words in each condition, and the percentage this forms of the total number of words in the show (note that this includes the commercial breaks).

Condition	Words	% Total	WER
F0 - prepared	638	12.3	17.9
F1 - spontaneous	1342	25.9	33.4
F2 - low fidelity	813	15.7	84.4
F3 - music	162	3.1	30.9
F4 - noise	187	3.6	51.4
FX - mixed	358	6.9	81.1
All	3500	67.5	48.5

Table 1: Results using pre-segmented evaluation data.

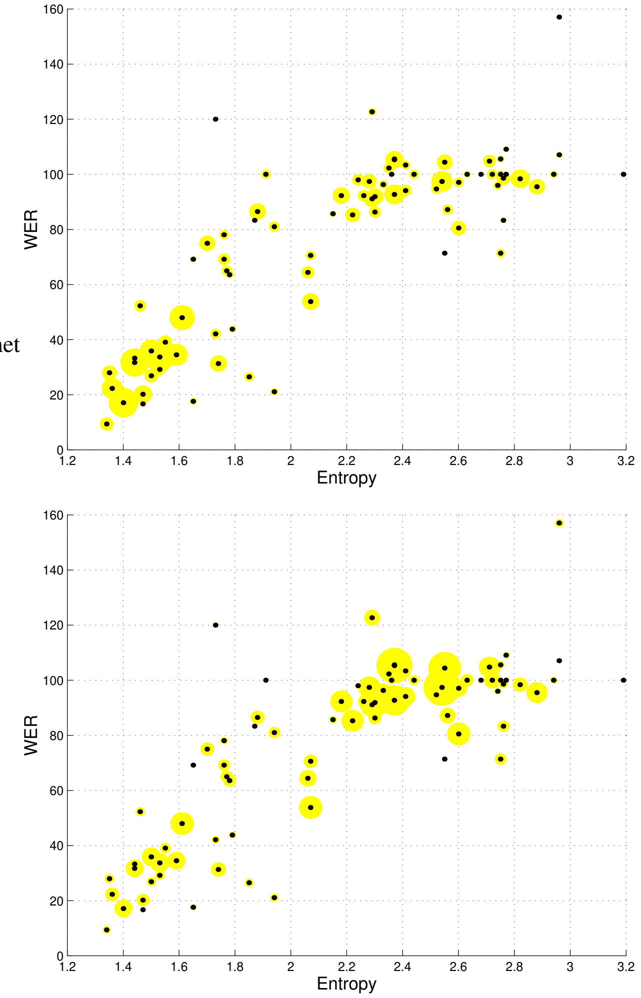

Figure 4: WER for each segment plotted against segment entropy value. Points are weighted by words per segment (top) or decoding time (bottom).

Figure 4 shows the average WER for each of the 81 segments returned by the automatic segmentation procedure. The area of shading around each point in the upper panel is proportional to the number of words in that segment whereas it is proportional to the time taken to decode the segment in the lower panel. It can be seen that there is a high degree of correlation between WER and the confidence value for the segments and also that although many of the 'poor' segments contain few words, they constitute a large proportion of the total decoding time.

The correlations between a segment's confidence estimate and it's WER (and decoding time) are detailed in table 2. Two measures are shown; a simple correlation, and a correlation weighted by the number of words in the segment. This weighting reduces the contribution of very short segments (which can contain as few as four words) for which the WER values are less reliable. Several variations of the confidence measure are shown, illustrating the importance of each stage in the processing of the raw per-frame entropy. The first row, 'raw', refers to the measure derived from a simple averaging of unprocessed frame entropies. The second row, '-transitions', shows how the correlation, and hence the reliability of the confidence measure, is improved by median filtering to remove the effect of phone transitions. The '-silence' row shows the large effect of excluding the silent frames, which may be just as well modelled in non-speech as in speech. The '-weak' row shows the small effect of excluding the set of indistinct phones that generally have intermediate values even in clean speech. The most reliable measure is achieved by combining each of these techniques.

By setting the confidence threshold to an appropriate value it is possible to exclude those segments that are expensive to decode but are nevertheless poorly recognised. In this way decoding time may be reduced by up to 70% without greatly increasing the overall word error rate. This point is illustrated by figure 5 which shows the overall WER as a function of the computational cost as the segment confidence threshold is relaxed and a greater number of segments are decoded. The flattening of the graph clearly indicates the diminishing returns of decoding each successively lower confidence segment[3].

[2] Note that a segment containing few words can receive an artificially reduced WER if a marking algorithm based only upon dynamic programming is used.

[3] Note that the minimum WER of 63% is calculated relative to the full

	WER vs. S		Cost vs. S	
	simple	weighted	simple	weighted
raw	0.684	0.825	0.665	0.845
-transitions	0.695	0.832	0.670	0.850
-silence	0.799	0.915	0.739	0.915
-weak	0.689	0.831	0.643	0.841
all	0.812	**0.923**	0.742	**0.919**

Table 2: The correlation between the entropy measure S and segment WER and computational cost.

By examining the manner in which the average WER *for recovered segments* varies as a greater number of segments are accepted for decoding, we can obtain some measure of the systems segmentation and classification performance. The upper line in figure 6 shows the WERs that are achieved using the acoustic measure and gradually relaxing the segment acceptance threshold. Compare this upper line to the lower line which simulates the WERs that could be achieved if the acoustic confidence measure was a perfect predictor of segment WER. Also plotted are points corresponding to the use of either all the pre-segmented evaluation data and just the F0 condition subset. The point to note here is that the best classification line passes very close to both these 'operating points'. If the system had made an inappropriate segmentation of the data, mixing poorly modelled and well modelled speech within individual segments, reaching the F0 operating point would not be possible.

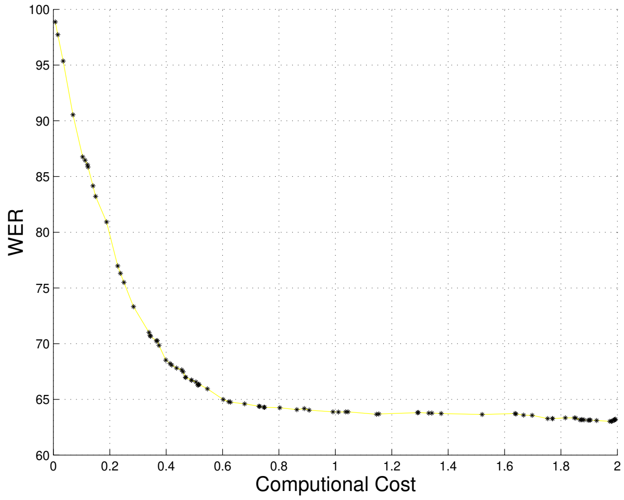

Figure 5: Overall WER as a function of computational cost as the segment confidence threshold is relaxed and an increasing number of segments are decoded.

7. CONCLUSIONS

We have presented a technique that uses a single acoustic confidence measure both to segment continuous audio and also to predict which segments contain speech that may be regarded as recognisable. The technique has two important attributes: First, it is computationally inexpensive allowing for an overall reduction in the computational cost of the recognition task. Second, as the confidence measure is derived directly from the recognition models the segmentation offered is entirely pragmatic, i.e. the data is divided into that which is a good fit to the models and is therefore likely to be recognisable, and that which is not. If different models are used then different segments will be found, but they will be the segments that are most likely to be of practical value. The results presented in this paper are derived from a single half hour radio broadcast. In order to fully assess the technique further evaluation is required over a larger, more diverse set of test data. Additionally, exploiting the durational constraints of speech and non-speech sounds through the use of a simple, two-state HMM may make the confidence measure more robust to non-speech sounds such as music.

5186 words that occur in the half hour broadcast not just the 3500 used for the Hub-4 evaluation.

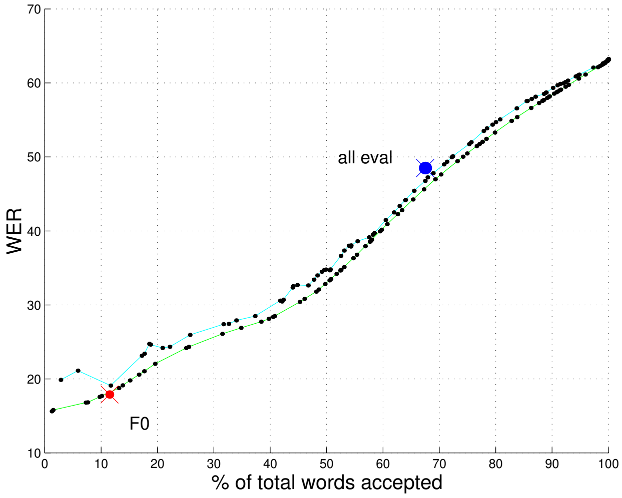

Figure 6: WER of the included segments increases steadily as the confidence threshold is decreased.

8. REFERENCES

[1] B. L. McKinley and G. H. Whipple. Model based speech pause detection. In *Proceedings of ICASSP*, pages 1178–1182, 1997.

[2] A.J. Robinson, M.M. Hochberg, and S.J. Renals. The use of recurrent networks in continuous speech recognition. In C-H. Lee, F.K. Soong, and K.K. Paliwal, editors, *Automatic speech and speaker recognition*, pages 233–258. Kluwer, 1996.

[3] M.A. Siegler, U. Jain, B. Raj, and M. Stern. Automatic segmentation, classification and clustering of broadcast news. In *DARPA Proc. Speech Recognition Workshop*, pages 97–99. Morgan Kaufmann, 1997.

[4] R.M. Stern. Specification of the 1996 HUB 4 broadcast news evaluation. In *DARPA Proc. Speech Recognition Workshop*, pages 7–10. Morgan Kaufmann, 1997.

[5] G. Williams and S Renals. Confidence measures for hybrid HMM/ANN speech recognition. In *Proceedings of EuroSpeech*, pages 19550–1958, Rhodes, 1997.