3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

ROBUST HMM-BASED ENDPOINT DETECTOR

A. Acero, C. Crespo, C. de la Torre, J.C. Torrecilla

Speech Technology Group
Telefonica I+D
Emilio Vargas 6, 28043 Madrid, Spain

ABSTRACT

A new real time HMM-Based endpoint detector is proposed in this paper. Endpoint detection has been shown to be critical in automatic speech recognition systems. The system uses static (energy) and dynamic (delta energy) features of the signal on a frame by frame basis. The endpoint detector is trainable for the working conditions (i.e. telephone lines) and is able to track changes in background noise conditions. Our experiments indicate that high accuracy, low false rejection and low false alarm rates can be obtained with this new endpoint detector.

1. INTRODUCTION

The problem of automatic endpoint detection consists of determining the beginning and end times for an utterance embedded in noise. Typically, isolated-word recognizers perform pattern matching on endpointed speech and it is well-known that the recognizer's error rate is highly dependent on accurate endpointing [1][2]. Explicit endpoint detectors (term proposed by Lamel [1]) work reasonably well with recordings exhibiting an SNR of 30 dB or greater, but fail considerably on noisier speech [3].

A more robust recognizer can be achieved by considering all possible start and end times to select the most likely word candidate. This approach, known as implicit endpoint detection, can be combined with Hidden Markov Modelling techniques [4] to best align, in the maximum likelihood sense, the word preceded and followed by noise with the incoming utterance. Implicit endpoint detection was found to be much more reliable than explicit endpoint detection in the sense of more accurate modelling and more importantly lower error rate. Implicit endpointing requires the word to lie within the boundaries of the signal sent to the recognizer, because insertions will occur if no word is present and deletions will occur if several words are present when only one is expected. While this is not a problem when running from a labelled file, in real-time implementations some kind of explicit endpointing is needed to assure that no word is missed. In addition, given that in practice a speech recognizer may be an expensive system resource, computation may be saved by running the recognizers only on periods of time where we know that speech is present.

Standard endpoint detection algorithms, such as that proposed by Lamel *et al.* [1], running on laboratory databases recorded on clean speech will most likely exhibit very low rejection rate, very low false alarm rate and very high accuracy as defined in Section 2. However, the problem is considerably harder on telephone speech [2] because of the reduced bandwidth and SNR, and variability of telephone lines.

In this paper we present a novel endpoint detector based on Hidden Markov Models and we compare its performance with our implementation of Lamel's [1]. We will describe experiments that show superiority of our algorithm, resulting in higher accuracy, lower false alarm rate, lower rejection rate. The new algorithm can be trained for different environments and it is also computationally inexpensive. In Section 2 we describe the experimental setup while Section 3 describes the algorithm.

2. EXPERIMENTAL SETUP

In order to assess the performance of various endpointers we need to define objective functions:

- *Accuracy*. It can be defined as the average difference and standard deviation in milliseconds between manually endpointed speech and automatically endpointed speech. This can be extracted from histogram of the differences. This measurement is mainly useful for recognizers based on explicit endpointing.
- *Rejection rate*. It can be defined as the percentage of speech pulses that are not detected by the endpointer. In an automatic speech recognizer, one such rejection will lead to a word rejection/deletion.
- *False alarm rate*. It can be defined as the number of false alarms per hour of recordings considered. Any such false alarm will lead to an insertion in the speech recognizer, unless the system has rejection capabilities for clicks, lip smacks and background noise.

A speech database was collected over telephone lines throughout Spain. Speech was sampled at 8 KHz and digitized with a commercial mu-law coder. The database consists of

8992 files amounting to 5 hours of recordings, each one containing one digit embedded in background noise, including breath noises, line clicks and other non-stationary noises. If the endpointer produces more than one speech segment in a file it means there is a false alarm. Likewise, if no segments are detected in a file a false rejection has occurred.

To evaluate the endpointing accuracy, we manually endpointed 347 files of the former database and measured the difference between manual endpoints and automatic endpoints at the beginning and end of the word. Following Wilpon [4], positive difference (for both the beginning point and ending point data) means that more signal was included within the hand endpoints than within the automatically determined endpoints. Trained listeners obtained manual endpoints by visual inspection of waveform and spectrogram, as well as by listening to the waveform segment. Manual endpoints may not be the best for automatic speech recognizers [2], but it is an objective measure that is independent of the recognizer, and therefore easier for comparisons.

Speech is passed through a preemphasis filter (coefficient equals 0.95) to remove DC and 50-Hz hum and to emphasize high frequency components present in fricatives. A 32 ms Hamming window is used every 16 ms to compute short-time log-energies.

Baseline results for our implementation of Lamel's [1] algorithm are shown in Table I and II, where the empirical thresholds were set (K1 = 3.5dB, K2 = 6.5dB K3 = 5dB, K4 = 15.5dB, T1 = 80ms, T2 = T3 = 64ms, NSEP = 96ms) after some trial and error. Figure 1 shows a histogram of differences between manual and automatic endpoints. In the following section we will describe our algorithm and compare it to the previous results.

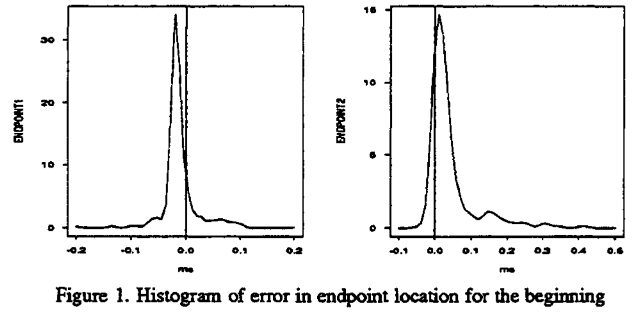

Figure 1. Histogram of error in endpoint location for the beginning and end of utterance for Lamel's algorithm. Time is in milliseconds.

3. ALGORITHM FUNDAMENTALS

The proposed endpoint detector consists of four modules: a feature extractor module, a speech activity module, a noise adaptation module and an endpoint detection module. A block diagram is shown in Figure 2.

Lamel's endpoint detector [1] can also be thought of having the same structure. In this case, the feature extraction module computes short-time energy in dB as the input feature. The noise estimation is accomplished by picking the mode of the 10 dB low energy values (the minimum energy has to be found in the file). The speech activity module performs a plain subtraction producing a noise-normalized log-energy. In this case there is no feedback from the speech activity module. The pulse detection is carried out by using a set of empirical thresholds on the length of the pulse and the values of energy during that pulse. Finally the utterance detection module makes up an utterance out of various pulses if they are close to each other to allow for intra-word silences, such as those present in stops, and inter-word silences, such as those present in real utterances. In the next subsections we will be presenting our algorithm and how it differs from [1].

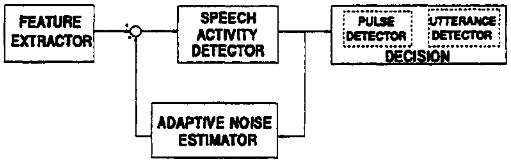

Figure 2. Block diagram of a general endpoint detector.

3.1 Feature Extractor

Lamel [1] proposes as the only feature the use of log-energy values. Although zero-crossing rate has been proposed in the literature [3] to determine endpoint location, it is not so useful for telephone bandwidth recordings [1], and therefore we did not use it here. In this work we propose to augment this feature with the use of delta log-energy (Dynamic information has been shown to provide improvements in speech recognition [6]). The use of delta-energy in Lamel's algorithm is probably complicated given the many *ad hoc* thresholds it contains. In Tables I and II we show that the use of delta log-energy does not improve the accuracy, but it yields lower false rejection rate and lower false alarm rate.

3.2 Speech Activity Detector

Speech activity is modeled here with Hidden Markov Models [5] following the rationale behind implicit endpoint detection. There is one HMM model for background stationary noise and one HMM model for speech and other signals. The feature vector consists of noise-normalized log-energy and delta log-energy. Continuous density HMM models are built with one Gaussian density per state, three states for the noise model and four states for the speech model (the exact number of states was not critical). HMM models for both noise and speech are trained from an isolated word database with digits embedded in noise. A Viterbi search is then run on another database using the network shown in Figure 3a to detect endpoints from a waveform file. The Viterbi search is in this case extremely simple, having only 7 states active each with one Gaussian density with a feature vector of only two components.

Unfortunately the former approach has the same problems of implicit endpoint detection if real-time is needed, i.e. we do not know when the utterance starts and we do not know when it finishes.

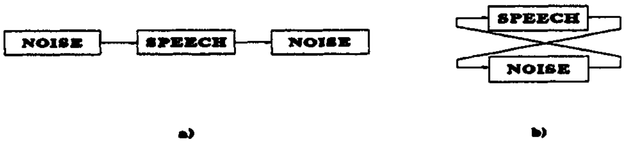

Figure 3.Network of HMMs defining the possible sequences for noise and speech. Network a) is suitable for canned speech, while network b) is required for live input.

To overcome this problem we started to look at the state probabilities while the Viterbi search is performed with time. The network is shown in Figure 3b, where the noise model is followed by the speech model that in turn makes a transition to the noise model. We define the log-likelihood of a frame being speech, log P(O_t/speech), as the average accumulated log-probability for all the N states in the speech model

$$\ln P(O_t/speech)=\frac{1}{N}\sum_{i=0}^{N-1}\ln \alpha_t^s(i) \qquad (1)$$

where $\alpha_t(i)$ represents the accumulated probability in time t for i-state, according to the terminology suggested in [5]. Likewise we can compute the log-likelihood for the M-state noise model as

$$\ln P(O_t/noise)=\frac{1}{M}\sum_{i=0}^{M-1}\ln \alpha_t^n(i) \qquad (2)$$

whose difference is a normalized score that is positive when speech is more likely and negative when background noise is more likely:

$$score[t] = \ln P(O_t/speech) - \ln P(O_t/noise) \qquad (3)$$

Figure 4 shows the log-energy and score as a function of time for a given utterance. Both normalized log-energy and the score can be the output of the speech activity detector.

3.3 Adaptive Noise Estimation

The noise adaptation module constantly updates the log-energy of the background noise, which is subtracted from the log-energy of every frame to make up the noise-normalized log-energy. Noise-level normalization makes the background noise *pdf* sharper, therefore increasing the discrimination between speech and noise.

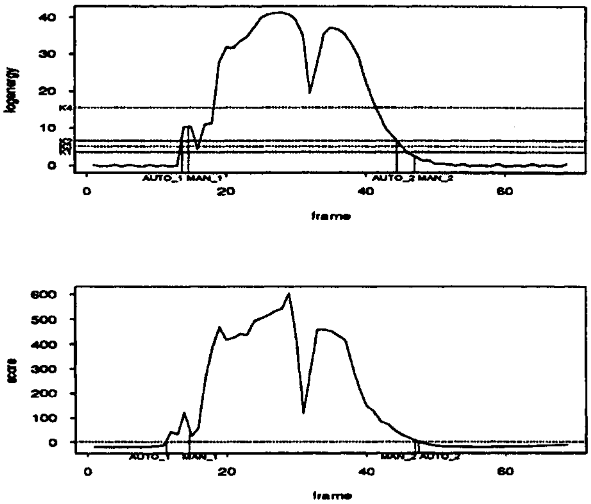

Figure 4. Log-energy and score given by (3) are plotted as a function of time for a given utterance. Both manual and automatic endpoints are shown for Lamel's algorithm (log-energy) and the proposed method (score).

To better track varying background noises, an exponential window

$$w_n[i] = \lambda^{(n-i)} \ , \ 0<\lambda<1 \ , \ -\infty<i\leq n \qquad (4)$$

is used to compute Q[n], a running average of noise frames:

$$Q[n] = \frac{\sum_{i=-\infty}^{n} w_n[i]*E[i]*P[i]}{\sum_{i=-\infty}^{n} w_n[i]*P[i]} \qquad (5)$$

where E[i] is log-energy of frame i and P[i] is the *a posteriori* probability of frame i being noise, obtained from the speech activity module as:

$$P[i] = \frac{P(O_t/noise)}{P(O_t/noise) + P(O_t/speech)} = \frac{1}{1 + \exp(score[t])} \qquad (6)$$

Note that both numerator and denominator in (5) can be implemented recursively:

$$\begin{aligned} Num[n+1] &= \lambda*Num[n] + E[i]*P[i] \\ Den[n+1] &= \lambda*Den[n] + P[i] \end{aligned} \qquad (7)$$

In our case we set the forgetting factor λ to 0.95 that results in a 200 ms time constant.

This method is able to track a decrease in background noise within a time constant. An increase in background noise takes considerably longer, as the energy will be above the noise level. To make possible the adaptation to changing background noises, we had to introduce two *ad hoc* modifications. To account for the modelling inaccuracies of the speech Gaussian, that can give values greater than those obtained by the noise Gaussian for negative noise-normalized log-energies, we forced the score to be a predetermined negative values in those cases. This makes possible to track a decrease in noise level in all cases. Likewise, to be able to track an increase in noise level, we forced the score not to exceed a predetermined value.

3.4 Pulse Detection

The pulse detection mechanism proposed by Lamel [1] consists of imposing a number of thresholds on the speech activity feature (normalized log-energy in this case). We built a similar scheme but only used two thresholds instead of four for values of the speech activity feature. Since the score in (3) is a log-likelihood function, pulse endpoints are determined by zero crossings. A check is made on the pulse so that the speech activity feature exceeds some empirical value (125) . As in the case of [1], we imposed a minimum duration on a pulse to consider it a valid pulse (160ms). These thresholds will hopefully eliminate clicks and other noises that are either too short or too low to be speech. Tables I and II show the performance of this approach.

ENDPOINT DETECTORS			REJECTION RATE (%)	FALSE ALARM RATE (per hour)
LAMEL		(E)	1	56
HMM-BASED	DURATION, AMPLITUDE	(E,ΔE)	0.27	42
	AREA	(E,ΔE)	0.19	35
		(E)	0.24	37

Table I. Rejection and False Alarm Rates for Lamel's algorithm and the proposed method with one (the area under the score function) and two thresholds (maximum speech activity and minimum duration).

ENDPOINT DETECTORS		ACCURACY (ms)			
		INITIAL ENDPT		FINAL ENDPT	
		μ	σ	μ	σ
LAMEL	(E)	-10.2	45.8	33.7	104.6
HMM-BASED (Area Criterion)	(E,ΔE)	-14.8	53.7	6.9	62.1
	(E)	3.7	51.0	14.9	55.4

Table II. Accuracy (mean and standard deviation) measured on the difference between manual and automatic endpoints.

The new criterion proposed is based on the log-likelihood of the a *posteriori* probabilities of a segment of frames being speech or noise, given by

$$Area = \ln \frac{P(speech/O_1 O_2 ... O_T)}{P(noise/O_1 O_2 ... O_T)} = \sum_{t=1}^{T} score[t] \qquad (8)$$

where we have used Bayes rule, and the assumption that the *a priori* probabilities of speech and noise are equal. To reject spikes and other non-stationary noises we requested every pulse to exceed a threshold (we used 1000 for this database). We see in Table I that this method performs better than the previous one, and better than Lamel's.

3.5 Utterance Detection

An utterance consists of a group of pulses separated by silence/background noise. As in [1], an utterance is considered finished when the silence following a pulse exceeds some programmable threshold (this can vary from about 100 ms for isolated word recognizers to almost one second if we are to accept sentences of spontaneous speech with some pauses or hesitations).

This scheme also allows for intra-word silences such as those coming from stops.

4. CONCLUSIONS

We have developed a novel real-time endpoint detector that exhibits higher accuracy than other non-real-time detectors published in the literature [1], has 80% fewer missed speech signals and has a 62% lower false alarm rate.

The system is able to track the evolutions of the background noise and it also offers the advantage that the endpointer is easily trainable from a database. The computational cost of the new algorithm is low.

The use of the delta energy parameter provides dynamic information about the signal, which outperforms systems based only on energy.

REFERENCES

[1] L. Lamel, L. Rabiner, A. Rosenberg and J. Wilpon. "An Improved Endpoint Detector for Isolated Word Recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing. Vol. 19, No 4. Aug. 1981.*

[2] J. G. Wilpon, L. R. Rabiner and T. Martin. "An Improved Word-Detection Algorithm for Telephone-Quality Speech Incorporating Both Syntactic and Semantic Constraints". *AT&T Bell Laboratories Technical Journal. Vol 63, No 3, March 1984.*

[3] L. R. Rabiner and M. R. Sambur. "An Algorithm for Determining the Endpoints of Isolated Utterances". *AT&T Bell Laboratories Technical Journal. Vol 54, No 2, Feb. 1975.*

[4] J. G. Wilpon. "Application of Hidden Markov Models to Automatic Speech Endpoint Detection". *AT&T Bell Laboratories Technical Memorandum. July 15, 1987.*

[5] L. R. Rabiner and B. H. Huang. "An Introduction to Hidden Markov Models". *IEEE ASSP Magazine 3 (1). pp 4-16. Jan 1986.*

[6] F. K. Soong and A. E. Rosenberg. "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition". *Proc. International Conference on Acoustics, Speech and Signal Processing 1986. Tokyo, Japan. Apr. 1986, pp 877-890.*