A ROBUST SPEECH/NON-SPEECH DETECTION ALGORITHM USING TIME AND FREQUENCY-BASED FEATURES

*Brian MAK, Jean-claude JUNQUA, and Ben REAVES**

Speech Technology Laboratory, Division of Panasonic Technologies, Inc.
3888 State Street, Santa Barbara, California 93105

*currently working at Central Research Laboratories, Matsushita, Japan.

ABSTRACT

We address the problem of automatic endpoint detection in normal and adverse conditions. Attention has been given to automatic endpoint detection for both additive noise and noise-induced changes in the talkers' speech production (Lombard reflex). After a comparison of several automatic endpoint detection algorithms in different noisy-Lombard conditions, we propose a new algorithm. This new algorithm identifies islands of reliability (essentially the portion of speech contained between the first and the last vowel) using time and frequency-based features and then applies a noise adaptive procedure to refine the endpoints. It is shown that this new algorithm outperforms the commonly used algorithm developed by Lamel and Rosenberg [1], and several other recently developed methods.

1 INTRODUCTION

A major cause of errors in automatic speech recognition systems is the inaccurate detection of the endpoints of test and reference patterns. It is essential for these algorithms that speech segments be reliably separated from non-speech. Attempts to relax and adjust inaccurate endpoints do not work well in all cases, and robust word boundary detection under adverse conditions remains an unsolved problem. Recently, a real-world evaluation of a discourse system using an isolated-word recognizer showed that more than 50% of the error rate was due to the endpoint detector [2]. According to Savoji, [3] the required characteristics of an ideal endpoint detector are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no a priori knowledge of the noise. Among these characteristics, robustness against adverse conditions has been the most difficult to achieve.

In this paper, we report on the comparative performance of several endpoint detection algorithms and then, based on the results obtained, propose a new algorithm based on time and frequency features and report on its evaluation. Finally, a hybrid implementation of this algorithm on a digital signal processor (DSP) is outlined.

2 A COMPARATIVE STUDY

2.1 Preliminaries

As a first step, we compared the performance of three recently developed endpoint algorithms to an algorithm [1] based on energy levels and timing, which is enhanced by automatic threshold setting [4]. The three algorithms differ in many respects such as the parameters they use (e.g. pitch, energy, zero-crossing rate, duration), their approaches, their complexity, and even the applications for which they are intended. We report their performances when integrated with two commonly used speech recognizers (DTW and discrete density VQ-based HMM) in various types of noisy conditions. Accuracy was judged by agreement with hand-labeled endpoints, and by recognition rates. The next two sections summarize the databases used and discuss the results obtained. More details about this comparative evaluation can be found in [5].

2.2 Databases

The training database for the recognizers was an American English ten-digit vocabulary spoken in a quiet environment by 96 speakers. The test database was the digit vocabulary produced in noisy conditions (2 repetitions) by 30 speakers (which were different from the training database). To simulate speech production in noisy conditions, white-Gaussian noise was played through calibrated headphones at 85 dB SPL. To test different types of noise

0-7803-0532-9/92 $3.00 © 1992 IEEE

disturbances the experiments were run with various additive noises extracted mainly from the RSG-10 noise database [6]: white-Gaussian noise, pink noise, multitalker babble noise, car noise, factory noise, gun noise, and airplane noise. Several levels of signal-to-noise ratio (SNR) have been considered, ranging from clean-Lombard speech (with no additive noise) to 5 dB SNR.

2.3 Results

Results obtained from this comparative evaluation show that 1) a recently developed noise adaptive algorithm (EPD-NAA) using rms energy, zero-crossing rate, and a set of heuristics generally gives the best results at high or medium signal-to-noise ratio (>15 dB). 2) The results of this endpoint algorithm differ on the average by 78 ms from hand-labeling (for clean-Lombard speech). 3) Extensive experiments showed that manual endpoints are not optimal for recognition. By simply varying hand-labeled beginning and ending points by up to 150 ms, in steps of 10 ms, we were able to reduce the HMM error rate by over 70%. 4) The HMM recognizer performs as well with the EPD-NAA algorithm as it does with hand-labeled clean-Lombard speech; for noisy-Lombard speech, depending on the type of noise used and the SNR, there is a degradation from 1% to 43% in recognition accuracy compared to hand-labeling. 5) At very low SNR (5 dB), the algorithm based on Lamel and Rosenberg's method [1] and enhanced by automatic threshold setting (EPD-ATA) generally gives better performance than the other algorithms.

This preliminary experimental evaluation of several endpoint detection algorithms brought to light that the EPD-NAA algorithm performs reasonably well at high or medium SNR. However, at low SNR (<15 dB) *some additional parameters are needed to improve robustness.* One of the methods evaluated was proposed in [7] and used pitch information to detect islands of reliability of the speech signal. However, pitch information is a difficult parameter to extract reliably, especially in adverse conditions. Our evaluation showed that this parameter is very sensitive to certain types of noise, like multitalker babble noise.

3 A NEW ALGORITHM ROBUST AGAINST ADVERSE CONDITIONS

To consistently extract islands of reliability, even in very noisy conditions, we used a parameter (hereafter called the time-frequency (TF) parameter) based on the energy in the frequency-band 250–3500 Hz and the logarithm of the rms energy. Such a feature was used in the identification of broad phonetic classes in the APHODEX system [8]. We selected the energy in the frequency-band 250–3500 Hz, because of its utility for detecting broad boundaries (essentially the portion of speech contained between the first and the last vowel of the speech signal). This energy is first normalized and smoothed by a median average algorithm. Then, the logarithm of the non-bandlimited rms energy is computed, normalized, and smoothed. The final parameter used (TF) is the result obtained after smoothing the sum of the two energy curves. Then, a noise adaptive threshold is computed from the first few frames of the speech signal to determine the beginning of the first vowel and the end of the last vowel (initial broad boundaries). Finally, the EPD-NAA algorithm is applied from the initial boundaries found to a earliest and latest possible boundary limit obtained by subtracting 100 ms from the beginning of the first vowel, and adding 150 ms to the end of the last vowel. If the algorithm, used to detect the islands of reliability, is robust, this new method should reliably yield endpoints which are close to the manual endpoints. In the following sections we will refer to the complete algorithm with the name EPD-TFF.

4 EXPERIMENTAL EVALUATION

We evaluated the EPD-TFF algorithm for clean-Lombard and noisy-Lombard speech. The databases used are the same as in Section 2.2. Performance was assessed by the accuracy of the automatic endpoints compared with the hand-labeled endpoints, and by recognition rates. This time, however, we used only the HMM recognizer. From the previous noisy conditions selected, we extracted the four kinds of noises which best represent various adverse conditions to which the HMM recognizer is sensitive: white, pink, car, and multitalker babble noise. The results obtained are presented in Figures 1 and 2, where automatic endpoints are compared to manual endpoints, and Figure 3, where performance was assessed using the HMM recognizer.

Compared to the other endpoint algorithms, the EPD-TFF algorithm gives the most accurate ending point. Generally, there is less than a 100 ms difference between the computed endpoint and the manually determined ending point (for all the noise and SNR conditions).

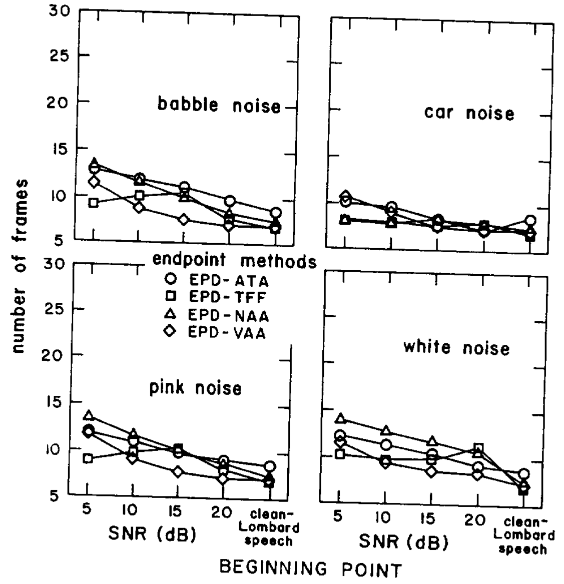

Figure 1 Average difference between the automatically and manually determined beginning point. The results are presented at various SNR for the test words, and different types of noise.

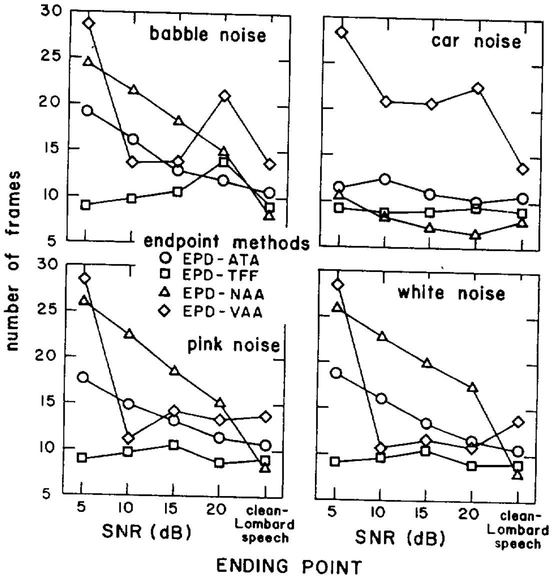

Figure 2 Average difference between the automatically and manually determined ending point. The results are presented at various SNR for the test words, and different types of noise.

For the beginning point, good performance is obtained at low and high SNR by the EPD-TFF algorithm. However, at medium SNR the EPD-VAA algorithm is generally the most accurate.

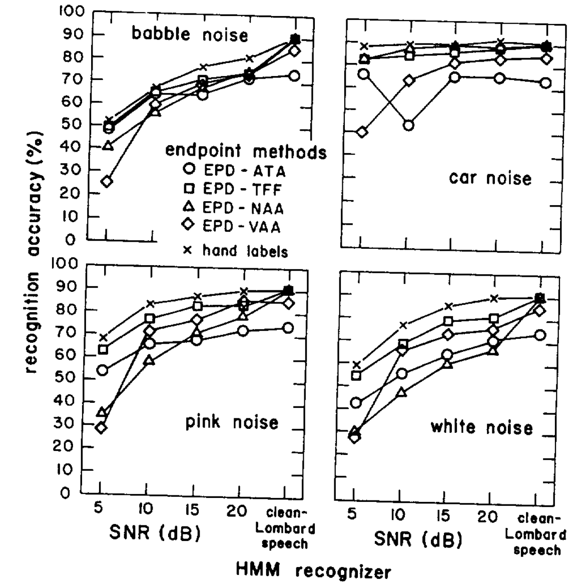

Figure 3 Recognition accuracy obtained with the HMM recognizer for the different endpoint algorithms and manual labeling. The results are presented at various SNR for the test words, and different types of noise.

When the HMM recognizer was used to assess the automatic endpoints, we found that 1) in the case of clean-Lombard speech, the recognition scores obtained with the endpoints produced by the new algorithm (EPD-TFF) are similar to the ones obtained with manual endpoints. 2) In the case of additive noise, EPD-TFF outperforms the other endpoint detection algorithms, especially at low SNR. Only for car noise the EPD–NAA algorithm performs slightly better than the EPD–TFF algorithm but both algorithms give good performance. 3) The degradation due to automatic endpoint detection, when using the EPD-TFF algorithm, is quite consistent across the various noise conditions. This was not the case for the other endpoint detection algorithms.

For the different endpoint detection algorithms we evaluated the error percentage due to the endpoint algorithm (hand-labels were taken as a reference) relative to the total number of errors. Figure 4 presents the results obtained as a function of the SNR.

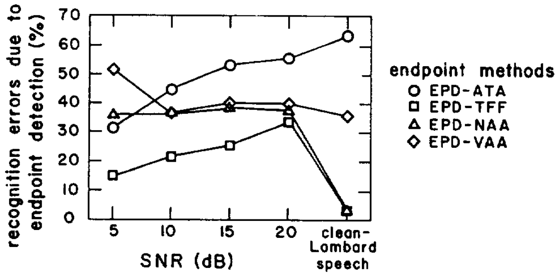

Figure 4 Percentage of recognition errors (averaged across the noise conditions) due to automatic endpoint detection as a function of the SNR

Compared to the EPD-NAA algorithm, the EPD-TFF algorithm provides a major improvement, especially at low SNR. This is essentially due to the additional TF parameter (based on time and frequency features) used to determine islands of reliability. It is interesting to notice that the percentage of recognition errors due to the EPD-TFF algorithm is the greatest at medium SNR (15 to 20 dB).

5 IMPLEMENTATION ON DSP

The EPD-TFF algorithm requires the computation of three parameters (the logarithm of the rms energy, the zero-crossing rate, and the TF parameter) on the entire speech signal before being able to determine the endpoints. However, it is possible to compute these parameters during the acquisition of the speech signal. Using the voice activation capabilities of the EPD-VAA algorithm (for more details see [5]), we first determine, in real-time, a rough estimate of the beginning sample of the speech signal in the continuous stream of input data. Then, the three parameters are continuously computed. Finally, after a rough estimation of the ending boundary of the speech signal, the final endpoint boundaries are determined. As the parameters necessary to find the final boundaries are already computed, the decision procedure is very quick. The role of the EPD-VAA algorithm is to compute some rough boundaries which contain the speech signal before applying the EPD-TFF algorithm. This hybrid algorithm has been implemented on a DSP board which is based on the TMS320C30.

6 CONCLUSIONS

Based on the results of a comparative study of several endpoint detection algorithms, we proposed a new algorithm (EPD-TFF) which uses a parameter derived from time and frequency features. For clean-Lombard speech, EPD-TFF provides as good recognition accuracy as that obtained with manually determined endpoints. In the presence of additive noise, the EPD-TFF algorithm outperforms the other endpoint detection algorithms studied. The use of a reliable parameter, robust against adverse conditions, to determine islands of reliability is found very beneficial, especially at low SNR. At high SNR, such a parameter maintains the high performance already obtained with the EPD-NAA algorithm. Nevertheless, according to our results, there is still room for improvement at medium SNR (15–20 dB). By taking into account in the algorithm possible pauses between words, it should be straightforward to apply the EPD-TFF algorithm to continuous speech. Finally, it is worth recalling that all the experiments reported in this paper used hand-labeled training data. Given the good results provided by the EPD-TFF algorithm at high SNR, an important future direction will be to assess the recognition performance when both training and recognition use automatic generated endpoints.

Bibliography

[1] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon. An Improved Endpoint Detector for Isolated Word Recognition. *IEEE ASSP*, 29, 1981.

[2] J.C. Junqua. Robustness and Cooperative Multimodal Man-Machine Communication Applications. In *Second Venaco Workshop and ESCA ETRW*, September 1991.

[3] M. Savoji. A robust algorithm for accurate endpointing of speech. *Speech Communication*, (8):45–60, 1989.

[4] B. Reaves. Comments on an improved endpoint detector for isolated word recognition. *Correspondence IEEE ASSP*, 39:526–527, February 1991.

[5] J.C. Junqua, and B. Reaves, and B. Mak. A study of Endpoint Detection Algorithms in Adverse Conditions: Incidence on DTW and HMM Recognizers. In *Eurospeech-91*, 1991.

[6] H. Steeneken and F. Geurtsen. Description of the RSG-10 Noise Database. Technical report, TNO Institute for Perception, 1990.

[7] M. Hamada, Y. Takizawa, and T. Norimatsu. A noise robust speech recognition system. In *ICSLP-90*, pages 893–896, 1990.

[8] Carbonell et al. "APHODEX, Design and Implementation of an Acoustic-Phonetic decoding Expert System". In *ICASSP-86*, pages 1201–1204, 1986.