A Robust Algorithm for Word Boundary Detection in the Presence of Noise

Jean-Claude Junqua, *Member, IEEE,* Brian Mak, and Ben Reaves, *Member, IEEE*

***Abstract*—We address the problem of automatic word boundary detection in quiet and in the presence of noise. Attention has been given to automatic word boundary detection for both additive noise and noise-induced changes in the talker's speech production (Lombard reflex). After a comparison of several automatic word boundary detection algorithms in different noisy-Lombard conditions, we propose a new algorithm that is robust in the presence of noise. 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, after a noise classification, applies a noise adaptive procedure to refine the boundaries. It is shown that this new algorithm outperforms the commonly used algorithm developed by Lamel *et al.* and several other recently developed methods. We evaluated the average recognition error rate due to word boundary detection in an HMM-based recognition system across several signal-to-noise ratios and noise conditions. The recognition error rate decreased to about 20% compared to an average of approximately 50% obtained with a modified version of the Lamel *et al.* algorithm.**

I. Introduction

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

Many applications of speech recognition require identifying words or short phrases in either speech or noise. Continuous speech recognition using background modeling and a finite-state grammar directly addresses this problem. The combination of word boundary detection and isolated word recognition addresses only the problem of speech embedded in noise. However, it has had more development, and in noise (with proper boundaries) currently achieves higher recognition rates than does implicit word boundary detection with background modeling.

This paper focuses on word boundary detection and its use in automatic speech recognition. After a brief overview of the literature on word boundary detection algorithms (Section II), the performance of three recently developed word boundary detection algorithms are evaluated and compared (Section III) with a newly enhanced version of a commonly used algorithm [1] based on energy levels and durations. Then, based on the results obtained, we propose a new algorithm (Section IV) based on time and frequency features, report on its evaluation, and show that this new algorithm outperforms the other methods to which it is compared. Section V presents an optimization of this algorithm and outlines a hybrid implementation on a digital signal processor (DSP). Finally, the main conclusions of our work are summarized. Throughout the paper, the term boundary refers to the beginning or ending frame of the speech patterns studied.

II. Overview of Existing Word Boundary Detection Algorithms

Currently, most word boundary detection algorithms use one or more of the following parameters: signal energy, zero-crossings, duration, and linear prediction error energy (e.g., [1], [3]–[6]). Recently, Hamada *et al.* used pitch information to distinguish speech signals from noisy signals [7]. In their algorithm, pitch information is extracted directly from the waveform. Generally, word boundary detectors that use only one parameter are algorithmically more complex in order to achieve good performance [1], [5]. A different approach adopted by Wilpon and Rabiner was to determine a set of speech boundaries based on the output of a Viterbi algorithm [8]. Hansen and Bria proposed a noise adaptive boundary detection [9] derived from [1]. An advantage of such an algorithm is that it deals with the variation of duration and intensity due to the Lombard effect and the additive noise induced by a noisy environment. However, threshold adaptation depends often on the type of additive noise used and it is difficult to rely only on an adaptive procedure to deal with various noise conditions.

Manuscript received March 20, 1992; revised November 12, 1993. The associate editor coordinating the review of this paper and approving it for publication was Dr. David Nahamoo.

J.-C. Junqua is with the Speech Technology Laboratory, Division of Panasonic Technologies, Inc. Santa Barbara, California 93105 USA.

B. Mak is with the Oregon Graduate Institute of Science and Technology.

B. Reaves is with the Central Research Laboratories, Matsushita, Moriguchi, Osaka 570, Japan.

IEEE Log Number 9400753.

III. A Comparative Study

A. Preliminaries

As a first step, we compared the performance of three recently developed word boundary detection algorithms to an algorithm [1] based on energy levels and durations, which is enhanced by automatic threshold setting [10]. We report their performances when integrated with a commonly used speech recognizer: vector quantization-based hidden Markov model (VQ-based HMM). The VQ-based recognizer used first and second order regression features, R_1 (with a 150 ms window) and R_2 (with a 230 ms window) [11], extracted from the index weighted cepstral coefficients derived from the twelfth model order of perceptually based linear prediction analysis (PLP) [12]. The training was done on clean speech produced in a normal environment (without background noise) and the testing on Lombard or noisy-Lombard speech. Accuracy was judged by recognition rates.

B. Databases

The training database for the recognizer 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 (two repetitions) by 30 speakers (who were different from the training speakers). 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 disturbances the experiments were run with various additive noises extracted mainly from the RSG-10 noise database [13]. Several levels of signal-to-noise ratio (SNR) have been considered, ranging from clean-Lombard speech (with no additive noise) to 5 dB SNR.

C. Overview of the Word Boundary Detection Methods Evaluated

1) An Energy-Based Algorithm with Automatic Threshold Adjustment (EPD-ATA): This algorithm from Lamel and Rosenberg [1] with some modifications from [10], is a general form of an intuitive approach based on energy levels and durations of silence and speech. As the noise thresholds are adapted dynamically, this algorithm is similar to the one proposed by Hansen and Bria [9]. A unique feature of this algorithm is that it yields not only the most likely pair of boundaries, but also other possibilities in order of their rank of being correct. Five energy thresholds are adapted automatically according to the voicing peak and the ambient noise estimated from the first few frames. These frames are taken from the beginning of a disk file containing one utterance each.

2) Use of Pitch Information (EPD-PCH): This word boundary detection algorithm [7] relies on pitch extraction and energy variations. It was designed for use in a real-time speech recognizer for controlling home appliances. This algorithm first attempts pitch detection directly from the waveform by a straightforward method of finding a peak whose amplitude is higher than the amplitudes of those surrounding it. Next, the regions where the pitch appears to be somewhat stable are declared to be an island of reliability. Finally, the beginning of the first island of reliability is taken to be an initial guess for the starting boundary; similarly, the end of the last island is the initial guess for the ending point. The initial boundaries are further refined using the energy curve.

3) A Noise Adaptive Algorithm (EPD-NAA): This algorithm, introduced here for comparative evaluation purposes, uses the log of the rms signal energy, the zero-crossing rate, duration information, and a set of heuristics. The thresholds used for the energy and the zero-crossings are adapted automatically from a few frames provided by the signal environment. First, the frame of maximum energy is located from the speech signal. Then, a search for the word ending boundary candidate begins on the basis of the logarithm of the rms energy. This candidate is then refined using zero-crossing rate, energy, and heuristic rules based on the previously calculated thresholds. The same procedure is applied to the word beginning boundary candidate. When the boundary candidates are found, a procedure to refine the boundaries based on the same parameters is applied. This procedure gives as output the initial word candidates and the refined boundaries. This algorithm, which is simple and fast, relies on the determination of the frame of maximum energy to start the search. Generally, this frame is easy to locate even in presence of noise.

4) A Voice Activation Algorithm (EPD-VAA) This algorithm is based on energy and zero-crossing parameters and a set of decision rules and threshold settings. Its implementation, which continuously looks at the input samples and detects the beginning and ending boundaries without an a priori knowledge of the maximum duration of the input signal, makes it suitable for real-time purposes. To detect the beginning boundary, speech is classified broadly in two categories: fricative-like speech or vowel-type speech. To each class is associated a set of conditions based on some functions (determined empirically) of the energy and zero-crossing parameters. The ending point is determined by classifying the energy in three categories: low, medium, high. To avoid confusion between a pause and the end of speech and to deal with transient noises, a history mechanism is implemented. A decision procedure based on the number of frames classified as noise and the noise category allows the ending boundary to be determined. This algorithm is fast and very practical: the speech signal is acquired at the same time as the word boundary detection is done.

D. Results

The evaluation has been done on two repetitions of Lombard speech produced by 30 speakers (600 utterances). To compare the evaluated algorithms to a reference, we ran the particular method described in [1] with thresholds determined empirically rather than automatically. This we call EPD-REF. Empirically set thresholds are likely to give better performance than automatically adjusted thresholds. This algorithm [1] is not well adapted to noisy environments—for this reason one of the proposed methods (EPD-ATA) is a modification of this algorithm to automatically adapt the thresholds to the

TABLE I
RECOGNITION ACCURACY OBTAINED WITH AN HMM VQ-BASED RECOGNIZER; THE BOUNDARIES WERE DETERMINED MANUALLY AND WITH THE DIFFERENT WORD BOUNDARY DETECTION ALGORITHMS EVALUATED

Word Boundary Detection Method	HMM Recognition Accuracy
Manual endpoints	90.5%
EPD-REF	84.3%
EPD-ATA	74.0%
EPD-PCH	61.5%
EPD-NAA	90.2%
EPD-VAA	85.2%

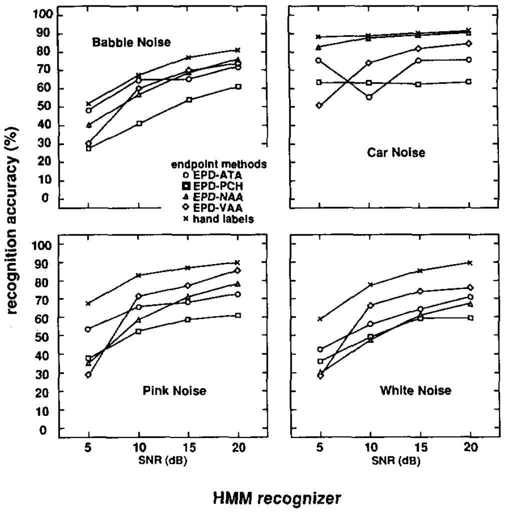


Fig. 1. Recognition accuracy obtained with an HMM VQ-based recognizer for the different word boundary detection algorithms and manual labeling. The results are presented at various SNR for the test words and different types of noise.

background noise level. The reference algorithm and the other word boundary detection algorithms were evaluated for clean-Lombard speech (no additive noise). The results obtained are presented in Table I.

For clean-Lombard speech the EPD-REF and EPD-NAA algorithms give the best results. The noise adaptive algorithm (EPD-NAA) gives recognition accuracy about as good as that obtained with hand labels. To see if boundaries could be manually optimized for a particular type of recognizer, we ran extensive experiments where hand-labeled beginning and ending boundaries varied by fixed amounts up to 150 ms, in steps of 10 ms. These experiments showed that there is a strong interaction between the boundary values and the recognition algorithm used. The optimum boundaries reduced the error rate by over 70% compared to the error rate obtained with hand-labels. When the boundary values vary, the errors made by the recognizers change. However, we could observe that long words (e.g., "zero") are less sensitive to word boundary values and generally better recognized than short words (e.g., "two").

To evaluate these different word boundary detection algorithms on noisy-Lombard speech, additive noise was used to simulate different noise conditions. The results obtained are presented in Fig. 1 for car, white-Gaussian, pink and multitalker babble noise.

Each word boundary detection algorithm is very sensitive to the noise spectrum. The EPD-NAA algorithm generally performs well at high SNR, while EPD-ATA, when compared to the other word boundary detection algorithms, gives its best recognition performance at low SNR values. Babble noise was not properly handled by the EPD-PCH algorithm. This is probably because, in this case, pitch information is difficult to extract.

E. Discussion

For clean Lombard speech, the EPD-NAA algorithm gives recognition scores as good as those obtained with manually determined word boundaries. However, for noisy-Lombard speech there is a degradation in recognition accuracy ranging from 1% (car noise SNR = 20 dB) to 43% (pink noise SNR = 5 dB) compared to manually determined word boundaries. The degradation in recognition accuracy for noisy environments is essentially due to the word boundary detection algorithm.

Looking closely at the boundary locations given by the various methods, it was found that, in the case of clean-Lombard speech, the EPD-ATA and the EPD-PCH algorithms tend to misclassify the beginning and ending portion of the words (the beginning and ending boundaries tend to be inside the words studied), while the EPD-VAA tends to misclassify the ending portion of the words (the ending boundary tends to cut the end of the words). With the EPD-NAA algorithm, the speech signal is generally contained between the boundaries detected.

This comparison shows that the relative performance of all four algorithms evaluated is strongly dependent on the noise condition. The reliability of the parameters used by different algorithms depends on the noise characteristics. Apparently, the energy and zero-crossing parameters are not sufficient to consistently determine reliable boundaries, even when using a complex decision strategy. No improvement was obtained using pitch information because it can be difficult to extract this parameter for certain noise conditions. There is a need to consistently detect what can be called islands of reliability, or in other words to base the word boundary computation on a rough speech detection, robust against noise conditions, followed by a refinement procedure. The EPD-NAA algorithm, which reliably identifies the frame of maximum energy before starting the search, is a very basic positive first step towards solving this problem. However, the results obtained are not yet satisfactory. In the following sections we present: 1) a new algorithm that addresses this need, 2) a comparative evaluation of this new algorithm and the best of the word boundary detection methods presented above, and 3) an optimization of this new method by introducing a noise classification procedure.

IV. A NEW ALGORITHM ROBUST AGAINST NOISE

A. Description of the Algorithm

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 computed on the entire frequency band of the speech

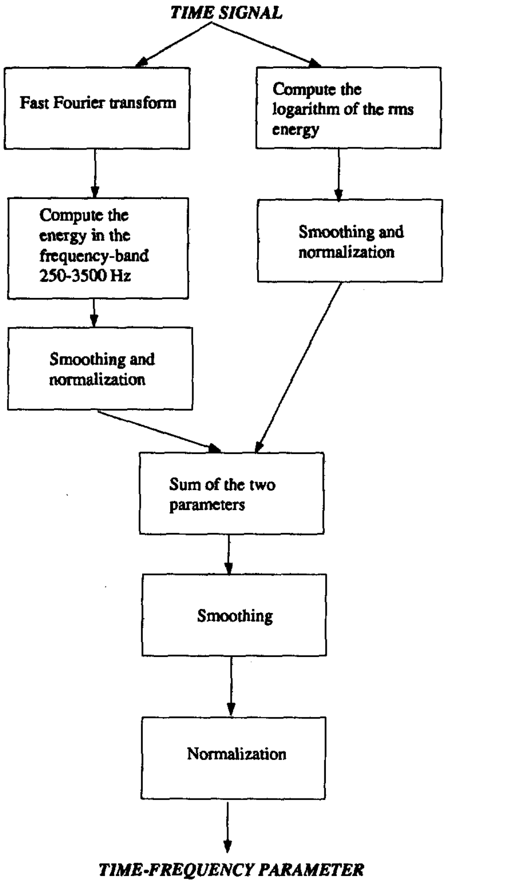

TIME SIGNAL
Fast Fourier transform
Compute the logarithm of the rms energy
Compute the energy in the frequency-band 250-3500 Hz
Smoothing and normalization
Smoothing and normalization
Sum of the two parameters
Smoothing
Normalization
TIME-FREQUENCY PARAMETER

Fig. 2. Block diagram of the computation of the time-frequency parameter used in the EPD-TFF algorithm.

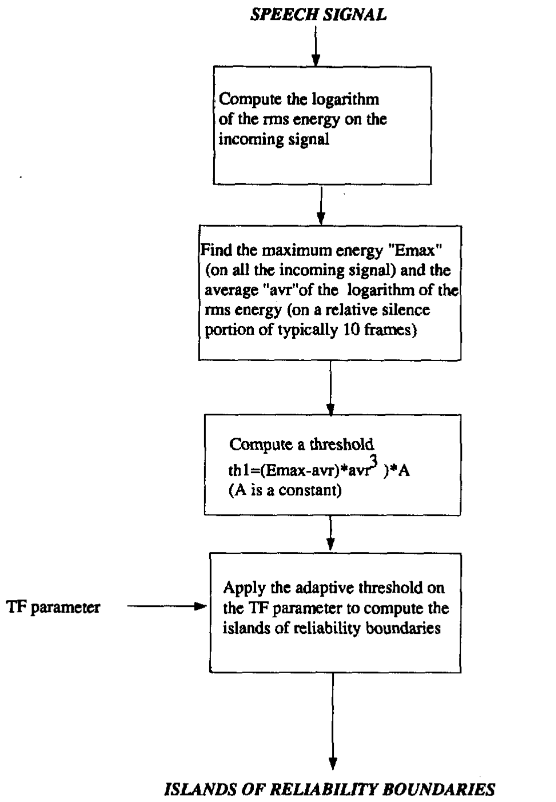

SPEECH SIGNAL
Compute the logarithm of the rms energy on the incoming signal
Find the maximum energy "Emax" (on all the incoming signal) and the average "avr"of the logarithm of the rms energy (on a relative silence portion of typically 10 frames)
Compute a threshold th1=(Emax-avr)*avr3)*A (A is a constant)
TF parameter
Apply the adaptive threshold on the TF parameter to compute the islands of reliability boundaries
ISLANDS OF RELIABILITY BOUNDARIES

Fig. 3. Block diagram of the computation of the islands of reliability in the EPD-TFF algorithm.

signal. Such a feature was successfully used in the identification of broad phonetic classes in the APHODEX system [14]. We selected the energy in the frequency band 250–3500 Hz because of its usefulness for detecting high energy regions (in the incoming signal) that correspond essentially to the vowel portions of the speech signal. This frequency-band helps the algorithm to make the distinction between speech and noise. By determining the portion of speech contained between the first and the last vowel of the speech signal, broad boundaries can be detected. This bandlimited energy is first normalized and smoothed by a median average algorithm. Then, the logarithm of the nonbandlimited 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). Fig. 2 illustrates the computation of the time-frequency parameter and Fig. 3 the detection of the islands of reliability using this parameter and the adaptive threshold. Finally, a refinement procedure is applied from the initial boundaries found to an earliest and latest possible boundary limit obtained by padding a fixed 100 ms to the beginning of the first vowel, and 150 ms to the end of the last vowel. The refinement procedure is presented in Fig. 4. If the algorithm used to detect the islands of reliability is robust, this new method should reliably yield boundaries that are close to the manual boundaries regardless of the type of additive noise. Fig. 5 shows a spectrogram of the word "one" with white-Gaussian noise added to the speech signal to obtain an SNR of 15 dB. On this figure the time-frequency parameter used to detect the islands of reliability has been plotted. As can be seen, the application of an adaptive threshold on the TF parameter gives a rough approximation of the word boundaries. In the following sections, we will refer to the complete algorithm with the name EPD-TFF.

B. Experimental Evaluation

We evaluated the EPD-TFF algorithm for clean-Lombard and noisy-Lombard speech. The databases used are the same as described above. Performance was assessed by recognition rates. The results obtained are presented in Fig. 6.

We found that 1) in the case of clean-Lombard speech, the recognition scores obtained with the boundaries produced by the new algorithm (EPD-TFF) are similar to those obtained with manually determined boundaries; 2) in the case of additive noise, EPD-TFF outperforms the other word boundary detection algorithms, especially at low SNR. Only for car noise is EPD-TFF outperformed slightly by EPD–NAA, but both

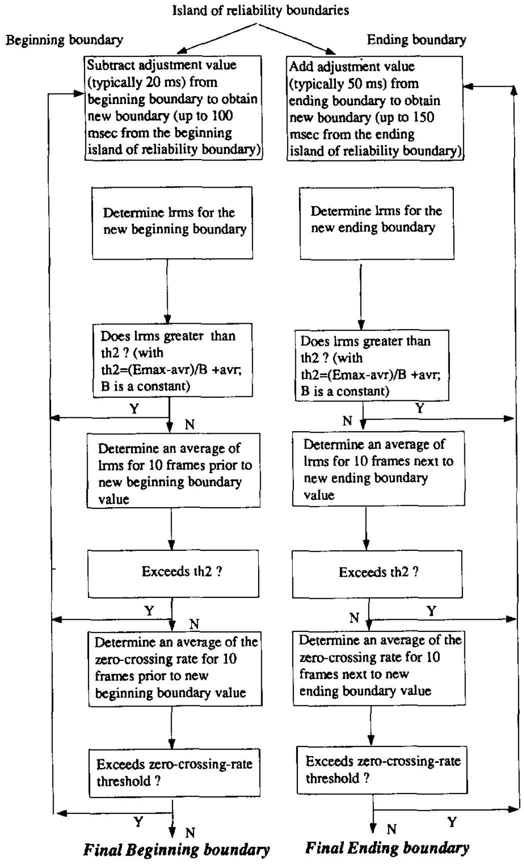

Island of reliability boundaries
Beginning boundary
Ending boundary
Subtract adjustment value (typically 20 ms) from beginning boundary to obtain new boundary (up to 100 msec from the beginning island of reliability boundary)
Add adjustment value (typically 50 ms) from ending boundary to obtain new boundary (up to 150 msec from the ending island of reliability boundary)
Determine lrms for the new beginning boundary
Determine lrms for the new ending boundary
Does lrms greater than th2 ? (with th2=(Emax-avr)/B +avr; B is a constant)
Does lrms greater than th2 ? (with th2=(Emax-avr)/B +avr; B is a constant)
Y
N
N
Y
Determine an average of lrms for 10 frames prior to new beginning boundary value
Determine an average of lrms for 10 frames next to new ending boundary value
Exceeds th2 ?
Exceeds th2 ?
Y
N
N
Y
Determine an average of the zero-crossing rate for 10 frames prior to new beginning boundary value
Determine an average of the zero-crossing rate for 10 frames next to new ending boundary value
Exceeds zero-crossing-rate threshold ?
Exceeds zero-crossing-rate threshold ?
Y
N
N
Y
Final Beginning boundary
Final Ending boundary

Fig. 4. Block diagram of the EPD-TFF refinement procedure.

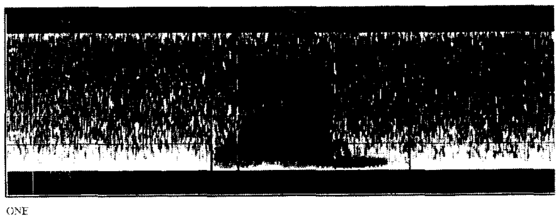


Fig. 5. Spectrogram of the word "one" (produced in noise by a male speaker) with additive white-Gaussian noise (SNR=15 dB). The time-frequency parameter, the islands of reliability boundaries, and the final boundaries are displayed.

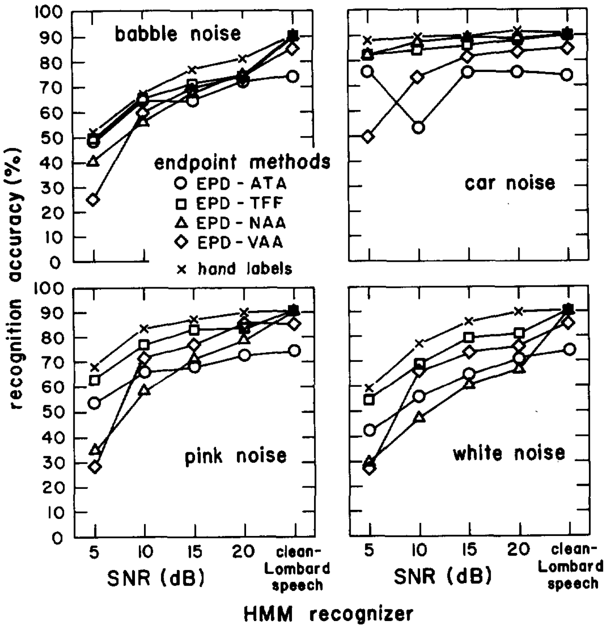


Fig. 6. Recognition accuracy obtained with an HMM recognizer for the different word boundary detection algorithms studied and manual labeling. The results are presented at various SNR for the test words and different types of noise.

algorithms give good performance; and 3) the degradation in recognition accuracy by EPD-TFF due only to automatic word boundary detection is quite consistent across the various noise conditions. This was not the case for the other word boundary detection algorithms.

Compared to the other word boundary detection algorithms, the EPD-TFF algorithm gives the most accurate word ending boundary. Generally, there is less than 100 ms difference between the computed ending boundary and the manually determined ending boundary (for all the noise and SNR conditions). For the beginning boundary, good performance is obtained at low and high SNR by the EPD-TFF algorithm. However, at medium SNR the EPD-VAA algorithm gives generally the most accurate beginning boundary.

In the following we will refer to the recognition errors obtained when using hand-labels by $E_{\rm HL}$and to the recognition errors obtained when using automatic word boundary detection by $E_{\rm AL}$. For each word boundary detection algorithm, we evaluated the percentage of recognition errors attributable to word boundary detection errors relative to the total number of errors (recognition accuracy with hand-labeled boundaries was used as a reference). This percentage was obtained by evaluating the ratio $\frac{E_{\rm AL}-E_{\rm HL}}{E_{\rm AL}}$. Fig. 7 presents the results obtained as a function of the SNR. Compared to the EPD-ATA algorithm, which yields an average across the SNR values of approximately 50% error rate due to word boundary detection, the EPD-TFF algorithm shows 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 ratio of the recognition errors due to the EPD-TFF algorithm to the total number of recognition errors is greatest at medium SNR (15 to 20 dB). In this case, the EPD-TFF algorithm still gives the best absolute recognition performance compared to the other word boundary detection algorithms. However, at medium SNR (it is also true at high SNR), as the noise does not influence too much the recognition accuracy, the word boundary detection precision becomes important and causes a large portion of the errors made by the recognizer. At high SNR, the EPD-TFF algorithm provides very good boundaries due to the precision of the refinement procedure. However, at medium SNR this procedure needs to be improved. At low SNR the refinement procedure is

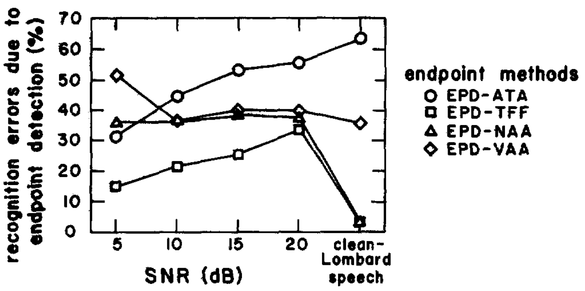


Fig. 7. Percentage of recognition errors (averaged across the noise conditions) due to automatic word boundary detection as a function of the SNR. It is important to note that this figure shows the ratio of the recognition errors due to word boundary detection (taking recognition scores obtained with hand-labels as a reference) to the total number of recognition errors obtained with these algorithms; it is not the absolute recognition performance obtained when using the different word boundary detection algorithms.

less important because of the important influence of the noise on the recognizer performance. In this case, even accurate word boundaries do not prevent the recognizer from making mistakes. To improve the performance of our new algorithm, we propose, in the next section, a noise classification procedure that improves performance at medium SNR without degrading word boundary results at other SNR values.

V. Optimization for Real-Time

By using 10 frames of "relative" silence at the beginning of the recording and computing an average of the logarithm of the rms energy and the zero-crossing rate on these frames, we determine the noise level (high, medium, or low) and the noise category (high or low zero-crossing rate). A set of threshold values, empirically determined, are used to perform this classification. If the noise level is low the same refinement procedure as the one used in the EPD-NAA algorithm is applied. However, if the noise level is classified as medium or high, a linear adjustment procedure depending on the noise rms energy is applied to adjust the islands of reliability boundaries. A single adjustment value or adjustment factor is subtracted from the approximate word beginning boundary and added to the approximated ending boundary.

We evaluated the final EPD-TFF algorithm for the four noise conditions at different SNR. The results are presented in Table II, where recognition rates of the improved and previous EPD-TFF algorithm are shown at medium SNR (15 and 20 dB), and in Fig. 8 where, as in Fig. 7, the percentage of errors due to the improved EPD-TFF word boundary detection method, the previous EPD-TFF algorithm and EPD-NAA (used in this figure as a reference), is presented. As can be seen in Fig. 8, the noise classification procedure results in a large improvement at medium SNR, leading to a substantial decrease, across different noise types, of the recognition error rate due to word boundary detection. At low or high SNR (5 dB and clean-Lombard speech) the noise classification procedure did not change the recognition accuracy.

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 speech boundaries. It is possible to compute these parameters during the acquisition of the speech signal. Using the voice activation capabilities of the EPD-VAA algorithm, 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 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 that contain the speech signal before applying the EPD-TFF algorithm. This hybrid algorithm has been implemented on a DSP board based on the TMS320C30 digital signal processing chip.

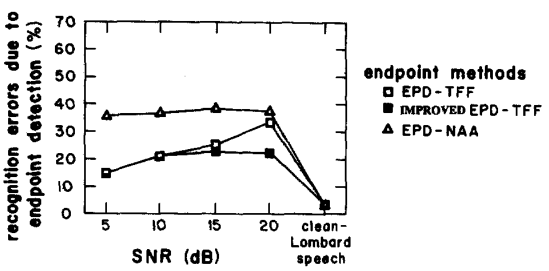


Fig. 8. Comparison of the percentage of recognition errors (averaged across the noise conditions) due to automatic word boundary detection between the improved EPD-TFF algorithm and the same algorithm without the noise classification procedure. The results obtained with the EPD-NAA algorithm are also shown in the figure.

TABLE II
Recognition Accuracy Obtained with the Improved EPD-TFF Algorithm and an HMM Recognizer at 15 dB and 20 dB SNR for Four Different Kinds of Additive Noise; the Results Obtained with the EPD-TFF Algorithm without the Noise Classification Procedure Are Indicated between Parentheses

Type of Noise	Recognition Accuracy of Improved EPD-TFF and EPD-TFF at SNR = 15 dB	Recognition Accuracy of Improved EPD-TFF and EPD-TFF at SNR = 20 dB
Car	87.0% (86.3%)	92.2% (88.7%)
White-Gaussian	80.3% (79.2%)	84.7% (80.7%)
Pink	83.5% (83.0%)	86.7% (83.3%)
Multitalker babble	71.3% (71.3%)	74.2% (74.0%)

VI. Conclusion

Based on the results of a comparative study of several word boundary detection algorithms, we have proposed a new algorithm (EPD-TFF) that 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 boundaries. In the presence of additive noise, the EPD-TFF algorithm outperforms the other word boundary detection algorithms studied. The use of a reliable parameter that is robust against noise conditions is found to be very beneficial, especially at low SNR. At high SNR, such a parameter maintains the high performance already obtained with the EPD-NAA algorithm. A noise classification procedure allows the algorithm to improve the performance at medium SNR. It is shown that the EPD-TFF algorithm reduces greatly the recognition error rate due to word boundary detection when dealing with noisy-Lombard speech. We evaluated the average recognition error rate due to word boundary detection in an HMM-based recognition system across several signal-to-noise ratios and noise conditions (without taking into account clean speech, for which very good results can be obtained). The recognition error rate decreased to about 20%, compared to an average of approximately 50% obtained with EPD-ATA, a modified version of the Lamel *et al.* algorithm. By taking into account in the algorithm possible pauses between words, which are mainly a problem in practical implementations where a speech signal is acquired continuously, it should be straightforward to apply the EPD-TFF algorithm to continuous speech.

Acknowledgment

The authors wish to thank T. Applebaum and B. Hanson from STL for providing the HMM algorithm, M. Hamada and Y. Takizawa from Central Research Laboratories of Matshushita for providing the EPD-PCH algorithm, and K. Tsuga of the Kansai Information Communication Research Laboratory who wrote a preliminary implementation of the EPD-NAA algorithm.

References

[1] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE ASSP Mag.*, vol. 29, pp. 777–785, 1981.

[2] J-C. Junqua, "Robustness and cooperative multimodel man-machine communication applications," in *Proc. Second Venaco Workshop and ESCA ETRW*, Sept. 1991.

[3] M. H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Commun.*, vol. 8, pp. 45–60, 1989.

[4] C. Tsao and R. M. Gray, "An endpoint detector for lpc speech using residual error look-ahead for vector quantization applications," in *Proc. ICASSP-84*, 1984, pp. 18b.7.1–4.

[5] H. Ney, "An optimisation algorithm for determining the endpoints of isolated utterances," in *Proc. ICASSP-81*, 1981, pp. 720–723.

[6] L. R. Robiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, no. 2, pp. 297–315, 1975.

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

[8] J. G. Wilpon and L. R. Rabiner, "Application of hidden Markov models to automatic speech endpoint detection," *Comput. Speech, Language*, vol. 2, pp. 321–341, 1987.

[9] J. Hansen and O. Bria, "Lambard effect compensation for robust automatic speech recognition in noise," in *Proc. ICSLP-90*, 1990, pp. 1125–1128.

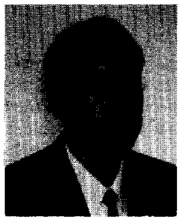
[10] B. Reaves, "Comments on an improved endpoint detector for isolated word recognition," *Corresp. IEEE Acoust., Speech, Signal Processing*, vol. 39, no. pp. 526–527, Feb. 1991.

[11] T. H. Applebaum and B. A. Hanson, "Robust speaker-independent word recognition using spectral smoothing and temporal derivatives," in *Proc. EUSIPCO-90*, 1990, pp. 1183–1186.

[12] H. Hermansky, B. A. Hanson, and H. Wakita, "Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain," *Speech Commun.*, vol. 4, pp. 181–187, 1985.

[13] H. J. M. Steenekan and F. W. M. Geurtsen, "Description of RSG-10 noise database," *Tech. Rep.*, TNO Institute for Perception, 1990.

[14] Carbonel *et al.* "APHODEX, design and implementation of an acoustic-phonetic docoding expert system," in *Proc. ICASSP-86*, 1986, p. 1201–1204.



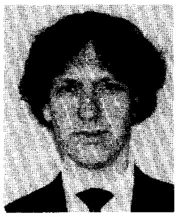
Jean-Claude Junqua (M'90) received the Engineer degree in electronics and automation from ENSEM (France) in 1980, the Master and Doctorate degrees in 1981 and 1989, respectively, and the "Habilitation à diriger des recherches", in 1993, from the University of Nancy I (France), in computer science.

From 1981 to 1986, he was Technical Director of the Computer Science Laboratory CRIN (France). From 1987 to 1988, he was Visiting Researcher at the Speech Technology Laboratory, Santa Barbara, CA, and in 1989, he joined the Laboratory. From April 1992 to August 1993, he was Visiting Researcher at Matsushita, Osaka, Japan. Currently Project Leader at Speech Technology Laboratory, his work has concentrated on improving robustness of isolated-word and connected-word automatic speech recognizers. His current interests cover all aspects of automatic speech recognition, e.g., the development of auditory models, the study of noisy-Lombard speech, the development of speech processing environments, and the design of knowledge-based and dialogue systems. He has published numerous papers in the above areas.

Dr. Junqua coorganized the ESCA workshop "Speech Processing in Adverse Conditions" and is currently on the Editorial Board of the *Speech Communication Journal.*

Brian Kan-Wing Mak received the B.S. degree in electrical engineering from the University of Hong Kong in 1983 and the M.S. degree in computer science from the University of California, Santa Barbara, in 1989. He is working toward the Ph.D. in computer science at the Oregon Graduate Institute of Science and Technology.

In 1990, he joined the Speech Technology Laboratory of Panasonic Technologies, Santa Barbara, CA, and worked on endpoint detection research in noisy environments. His interests include speech processing and recognition, neural networks, and machine learning.



Ben Reaves (S'83–M'83) received the Bachelor's degree from the University of Southern California, Los Angeles, in 1981, and the Master's degree from Stanford University, Stanford, CA, in 1983, both in electrical engineering (statistical signal processing).

From 1983–1985, he was a Member of the Technical Staff at Hughes Aircraft Company, Fullerton, CA, working on digital hardware for a spread spectrum modem. Since 1985, he has been a Research Engineer at the Speech Technology Laboratory, Santa Barbara, CA. Since 1989, he has been working with the Speech Recognition Group (CG3) of Matsushita's Central Research Laboratory, Osaka, Japan. His primary research interests are in improving the usability and reliability of automatic speech recognizers: His work is in the area of real-time automatic detection of speech in noisy environments. He has published several papers in this area.