Robust Word Boundary Detection Using Fuzzy Logic

F. Beritelli

Indexing terms: Endpoint detection, Speech recognition, Pattern recognition, Fuzzy logic

The Letter proposes a pattern recognition approach to robust word boundary detection in adverse acoustic noise conditions. The algorithm uses four simple differential parameters calculated in the time domain and a pattern matching based on a set of six fuzzy rules extracted by a hybrid learning tool. The experimental results demonstrate that the new endpoint detector outperforms traditional methods, above all with high levels of background noise.

Introduction: In the last few years the growth of multimedia applications has increased the demand for new and more efficient speech command and control systems for man-machine interaction. In this context, the large new application scenarios will require systems and methodologies able to guarantee good performance levels, even in adverse acoustic noise conditions, with as low a computational load as possible [1]. As the recognition rate of an isolated word recognizer strongly depends on the accuracy of the End Point Detector (EPD) [1][2], for the effective automatic recognition of speech a robust word boundary detection algorithm is essential [3].

In this Letter we propose a new fuzzy logic-based boundary detection algorithm that meets requirements of both computational simplicity and robustness to background noise. The Fuzzy End Point Detector (FEPD) uses a set of four simple differential parameters and a matching phase based on a set of fuzzy rules. Experimental evaluation shows that the FEPD outperforms traditional word detection methods.

Description of the algorithm: The architecture of the endpoint detector proposed is based on a pattern recognition approach. More specifically, it consists of pre-processing the speech signal, extracting its significant features, a matching phase, a post-processing module and finally a decision block. Pre-processing of the speech signal consists of 140-Hz high-pass filtering in order to eliminate the undesired low-frequency components. In order to guarantee a robust word boundary detection in the presence of high noise levels, rather than using absolute parameters like energy, correlation and zero crossing, in this Letter we propose a different approach based on a set of six differential parameters. The fuzzy system has the task of mapping the pattern of input parameters onto a scalar value, ranging between 0 and 1, which indicates the degree of membership in the voice inactivity/activity classes [6].

The four differential parameters used for speech/noise classification are calculated in a window of 80 speech samples (a 10-ms frame, sampling at 8 kHz) and are: the full-band energy difference $\Delta E_b$, the high-band energy difference $\Delta E_h$, the zero-crossing difference $\Delta ZC$ and the spectral distortion $\Delta S$ [4].

For the matching phase a new methodology is used having the advantage of exploiting all the information in the input pattern by means of a set of six fuzzy rules automatically extracted by a hybrid learning tool [5]. The fuzzy system has the task of mapping the pattern of input parameters onto a scalar value, ranging between 0 and 1, which indicates the degree of membership in the voice inactivity/activity classes [6].

In order to reduce sharp variations in the fuzzy system output, it is post-processed by a 7-th order median filter. Finally, the decision module, by means of a threshold comparison, returns the start and the ending point. Experimentally we chose a threshold value of $T_h=0.9$ in that, observing the output of the post-processing module for several types of background noise and signal-to-noise ratios, it was rarely observed to exceed 0.9 in segments of pure noise, whereas inside word boundaries it is always close to 1. Further, in order to avoid false alarms, due to non-stationary noise, or premature endpoints due to possible intra-word pauses or unvoiced sounds, we inserted a simple control mechanism with two windows $W_1$ and $W_2$ frames in width, similar to the one proposed in [7]. To determine the start point, the window slides along the post-processed signal; if the number of elements that exceed the threshold value $T_h$ is greater than $k^*W_1$, it is a start point; otherwise it is a false alarm. The same procedure is followed to determine the endpoint, which is detected if the signal is below the threshold $k^*W_2$ times. We experimentally fixed $k=0.8$, $W_1=5$, $W_2=40$.

Experimental results: In this section we evaluate the performance of new endpoint detection algorithm analyzing the results of a series of tests carried out in various acoustic noise and speech level conditions. The testing database consists of isolated words sampled at 8 kHz, linearly quantized at 16 bits per sample, normalized at -20 dBmO and preceded and followed by about 0.5 seconds of silence. More specifically, we chose as the test words a ten-digit Italian vocabulary pronounced by 20 male and 20 female speakers different from those used in the training phase. In order to evaluate the FEPD robustness to noisy environments we digitally added three typical background noises (car, babble, traffic) at different signal-to-noise ratio values (30, 15, 5, 0 dB). Figures 1 (a)-(d) show a comparison between the FEPD performance and that of three traditional methods in terms of weighted accuracy (WA), an evaluation parameter, expressed in frames, which gives greater weight to the effect of word clipping than to widening, as suggested in [2]. In the comparison we applied to each method the same window mechanisms for the false alarm and intra-word silence problem. Further, in the graphs the WA values were saturated at a value of 50 frames. As regards the performance of traditional methods, the first two reference endpoints detectors, proposed by Tsao et al. [7] and by Lamel et al. [8], present quite good performance only with SNRs greater than 20 dB, but they have the advantage of a low complexity. The third method, recently proposed by Junqua et al.[3], performs better in noisy environments but is computationally more complex. FEPD performance is better than traditional methods with all SNRs and types of background noise. More specifically, the results demonstrate not only an improvement with high SNRs, but also optimum FEPD behaviour in very noisy conditions. At SNR=0 dB, for example, the WA parameter is halved with respect to traditional word boundaries detectors. In the presence of babble noise we can observe that the WA tends to saturate when the background noise level increases and is reduced with traffic noise at SNR=0 dB. This anomalous behaviour is due to the combined action of two factors: a) a saturation of the fuzzy output with high noise levels, which causes a widening of the endpoints (thus reversing the trend occurring up to SNR=5 dB); b) the criterion used to define the WA parameter, which gives greater weight to clipping than to widening.

In all conditions, we observed a great reduction in the amount of clipping and widening greater than 180 ms (18 frames) in length, i.e. of the errors that most degrade the performance of a speech recognition algorithm.

Finally, one of the problems of an energy-based speech activity detector is the large dynamic range of speech signal levels. Above all in the presence of noise, the performance of a word boundary detector deteriorates as the speech level decreases. This problem is usually dealt with by using an activity gain control (AGC) system. Thanks to the efficient set of differential parameters together with a robust, non-linear methodology for the pattern matching, the performance of the FEPD is quite independent of speech level changes. Figure 2 gives the weighted accuracy with varying SNRs and signal levels ($\pm 10$ dB as compared with the nominal level). Except for the 0 dB/-30 dBmO case, performance is similar to the reference nominal case of -20
dBmO, so the FEPD is robust to level variations and does not need an AGC system.

![Graph](image)

**Fig. 1 (a-d)** Weighted accuracy versus signal-to-noise ratio for different types of background noise.

**Conclusion:** We have presented a computationally simple and robust endpoint detection algorithm. Unlike traditional methods, the FEPD does not require adaptive thresholds or refinement procedures. Compared to the main traditional word boundary detection algorithms, the FEPD algorithm gives the most accurate word boundary detection, above all with high levels of background noise.

Francesco Beritelli (Istituto di Informatica e Telecomunicazioni - University of Catania, V.le A. Doria 6, 95125 Catania, Italy)

**References**


