An Adaptive and Fast Speech Detection Algorithm

Dragos Burileanu[1], Lucian Pascalin[1], Corneliu Burileanu[1], and Mihai Puchiu[2]

[1] "Politehnica" University of Bucharest, Faculty of Electronics and Telecommunications
Blvd. Iuliu Maniu 1-3, Sect. 6, Bucharest 77202, Romania
(bdragos, lucianp, cburileanu)@mESsnet.pub.ro
[2] Graphco Technologies Inc. Romanian Office
Calea Floreasca 91-111, Sect. 1, Bucharest 714011, Romania
mpuchiu@graphcotech.ro

Abstract. The detection of speech from silence (actually background noise) is essential in many speech-processing systems. In real-field applications, the correct determination of voice segments highly improves the overall system accuracy and minimises the total computation time. This paper [1] presents a novel robust and reliable speech detection algorithm to be used in a speaker recognition system. The paper first introduces some basic concepts on speech activity detection and reviews the techniques currently used in speech detection tasks. Then, the proposed speech/non-speech detection algorithm is described and experimental results are discussed. Conclusions about the algorithm performances are finally presented.

1 Introduction

The problem of separating the speech segments of an utterance from the background noise is called *speech detection* (or *endpoint detection*, or even *voice activity detection*). This is very important in many areas of speech signal processing: speech and speaker recognition, speech compression and transmission. A proper location of regions of speech (sometimes together with pause removal), not only reduces the amount of processing, but also increases the accuracy of the speech processing system.

Speech detection is an old and known problem; however, it has not been completely answered until today. The major difficulty of the speech detection task deals with the variability of the speech and background noise patterns. The noise and voice inconsistency often leads to the inaccurate detection of the speech endpoints, by cutting phonemes or passing non-speech events to the speech processing system.

Another important problem is the absence of an accurate method of checking the correctness of a speech detection algorithm, or for comparing two or more

[1] This work was sponsored by Graphco Technologies Inc., Newtown, Pennsylvania, USA

candidate methods. The location of the speech endpoints in the presence of high-level noise becomes a complex task even for experienced phoneticians. The commonly used manual method consists of a rough approximation, followed by a more precise endpoint location with acoustical and visual assistance.

The methods used in speech detection can be classified into two general categories, according to their basic principles:

- *explicit methods*, that detect the endpoints independently of the subsequent processes; the main techniques are *energy-based*, *pattern recognition-based*, or use *Hidden Markov Models* or *neural networks* to explicitly determine the endpoints [1],[3],[5],[6];
- *implicit methods*, that use the application itself to detect the endpoints; this approach assumes that the endpoint detection is an implicit part of the speech processing system and is developed along with the application [7].

All these techniques may work well enough for high signal to noise (SNR) ratios, but most of them are not really adaptive and degrade their performances as the noise signal increases. Furthermore, the techniques based on training have difficulties in finding features that are discriminatory between unvoiced speech and silence and also can fail when the environment noise does not match the training or when human artefacts and background voice corrupt the speech.

Many speech detectors developed for real-field applications (such as those used in noisy office, factory, moving cars or airplane cockpit environments) additionally use *speech enhancement techniques* [4]. Without any doubt, these hybrid methods perform better, but they usually need a priori knowledge of the noise statistics and are rather time consuming, leading to slow system response.

2 Description of the Speech Detection Algorithm

The proposed algorithm is based on short-term parameter estimation. It relies on two classical measures of speech, one with meaning of magnitude and the other with meaning of frequency: E_x - the *pseudoenergy* of the input signal, and ZCR - the *zero-crossing rate*, calculated over all samples of a signal frame [2],[5],[6]; we found that a frame-length of 15 ms is a good choice for accurate detection.

The procedure of speech detection and silence elimination for a recorded utterance consists in four steps:

I. Pre-processing. Because the ZCR is highly susceptible to 50/60 Hz hum, very low frequency noise, DC offset, etc., the waveform is first high-pass filtered; the filter is used only for energy and ZCR computing and does not affect the input signal.

II. Speech boundaries rough estimation. Two energy thresholds, T_1 and T_2 are calculated over all M frames of the input signal, based on three typical values of the speech signal and according to the following equations:

$$En\_max = max\, E_x\,(i) \ , \quad i = 1\,,\ 2\,,\ \ldots\ ,\ M \tag{1}$$

$$En\_min = min\, E_x\,(i)\ ,\quad i = 1\,,\,2\,,\,\ldots\,,\,M \tag{2}$$

$$T_1 = En\_min\left(1 + 2\,log_{10}\,\frac{En\_max}{En\_min}\right) \tag{3}$$

$$SL = \frac{\sum_i E_x\,(i)}{\sum_i 1}\quad ,\, where\, i\, is\, the\, index\, for\, all\, frames\, having\, E_x > T_1 \tag{4}$$

$$T_2 = T_1 + 0.25\,(SL - T_1) \tag{5}$$

In the previous equations, En_max is the peak energy value of the input speech signal, En_min is the minimum energy value (typically the silence or background noise minimum level), and SL is the average level of the signal above T_1.

The speech boundaries are approximately estimated based on the following energy criteria:

1. When energy exceeds T_1 and subsequently exceeds T_2, the crossing point with T_1 level is declared a 'preliminary startpoint' (PS).
2. When energy falls below T_2 and then falls below T_1, the crossing point with T_1 level is declared a 'preliminary endpoint' (PE). The region between the PS and the PE will be further called 'word'.
3. A short isolated 'word' is a 'noise spike'; it is marked as unnecessary and it will be rejected in the silence removal procedure.

III. Speech boundaries fine adjustment. After a first guess of the 'word' boundaries based on energy criteria, possible unvoiced regions near the 'word' (and belonging to the 'word') will be also considered as "speech", according to *ZCR*-based methods. Only after this step, one obtains the real start-endpoints of the declared 'voice segments'.

The *ZCR* parameter has generally a different behaviour in speech regions than in the silence (background noise) ones. However, because of its rapid variations, we found that it is rather difficult to correctly determine this behaviour; therefore, our algorithm uses an additional parameter: a *smoothed ZCR* (*SZCR*), that uses the three left and right neighbour values of the *ZCR* and describes the overall trend of the *ZCR*. It has a clean curve and roughly indicates the endpoints location; ZCR will be used to accurately refine the found points. To distinguish consonants from silence, we also calculate a *medium ZCR* (ZCR_{med}) and two variation areas for *ZCR* (denoted D_1 and ranging between ZCR_{low} and ZCR_{high}) and *SZCR* (denoted D_2 and ranging between $SZCR_{low}$ and $SZCR_{high}$).

After computing these values, an unvoiced speech segment is added to the 'word' using the following criteria:

4. A region near to a 'word', for which *SZCR* leaves D_2 area just closed to the 'word', belongs to the 'word' until the *SZCR* curve re-enters into D_2 area.

5. A region near to a 'word', having $SZCR$ outside the D_2 area and not complying with rule no. 4, belongs to the 'word' until the $SZCR$ curve enters into D_2 area.
6. After the new boundary of the 'word' is established (using rules no. 4 or 5), it is refined back until the ZCR curve is in the same side of the D_1 area like the $SZCR$ curve.
7. A long region that is added to the 'word' using rules no. 4 / 5 and 6 is truncated if it is the left region of the 1st 'word' in the utterance or the right region of the last 'word'. This rule helps to remove the deep breath at the beginning or at the end of an utterance.

IV. Silence elimination. All frames that are not detected as 'voice segments' by steps **II** and **III** are considered as 'silence' and removed from the given utterance.

3 Experimental Studies

To evaluate the accuracy of our algorithm and its robustness to various acquisition conditions, we performed a great number of experiments: optically and acoustically assisted tests on various commercial databases (YOHO, TIMIT, CTIMIT, NTIMIT, HTIMIT, FFMTIMIT) and tests using a specific recognition system.

Qualitative tests were done on hundreds of recorded utterances from the above-mentioned databases. We verified the endpoint detection in distinct conditions: high and low SNR, typical background noises, different microphones and transmission channels; silence removed utterances were optically checked and listened and finally compared with the original utterances. The conclusion was that the algorithm accurately detects the speech boundaries and copes with many different acquisition conditions; the few errors are mainly related to high noise energy or limited bandwidth transmission and telephone line attenuation. Figures 1 to 4 show the overall speech detection procedure applied on a same phrase ("Don't ask me to carry an oily rag like that"), uttered by the same speaker, in four databases: TIMIT (clean speech), CTIMIT (cellular telephone speech in various noise conditions), NTIMIT (standard telephone speech), and HTIMIT (different telephone handsets used); white zones indicate speech regions and grey zones indicate the rejected 'silence' regions.

We also completed a quantitative test: a precise comparison of automatically and manually labelled endpoints on 300 sentences from the YOHO database. The time distances between the 1,740 manually defined endpoints and the endpoints detected by our algorithm were measured. Assuming an acceptable distance of one frame (15 ms) between a real and an estimated endpoint, the resulted algorithm accuracy was better than 98%.

Finally, we performed a group of tests using a speaker verification system and four subsets of the YOHO database (26 speakers, each with 24 enrolment phrases and 40 verification phrases). The tests were accomplished on the original files that contain silence and then on the processed files (with silence removed by

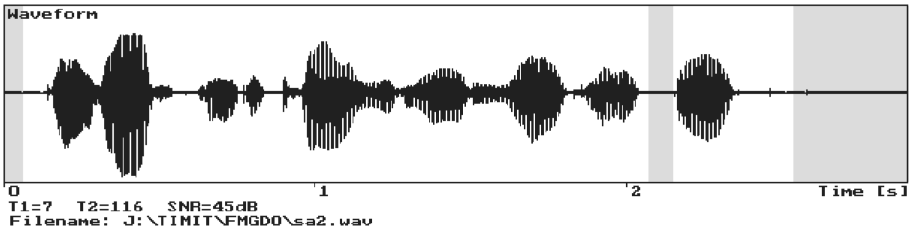

Fig. 1. TIMIT sentence

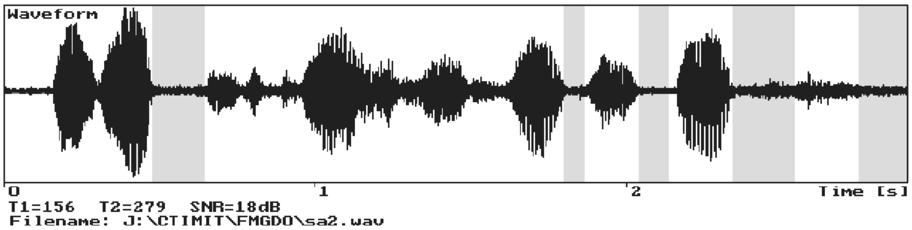

Fig. 2. CTIMIT sentence

our algorithm); a basic parameter representing *the Equal Error Rate* (EER) was calculated. The results indicated that using the proposed algorithm an average EER improvement of 2.67% (absolute value) is obtained.

4 Conclusions

Our main purpose was to develop and test a robust speech detection algorithm for a speaker recognition system having the following distinctive features:

- the possibility of use in any speaker recognition system and therefore independent of the recognition algorithm;
- the adaptability on the background noise level;
- an accuracy "as high as possible" of the detected speech boundaries; this means to include all significant acoustic events within the detected speech segments (and then to eliminate the non-speech regions of an utterance);
- the possibility of detecting and rejecting some typical background noises: isolated short external noises (including human artefacts), or very low frequency noise (for example 50/60 Hz hum);
- a reasonable complexity and consequently operation in real-time on a previously acquired utterance.

Considering these fundamental requirements, we designed and implemented a novel *explicit energy-based speech detection algorithm*. Tested on many of the

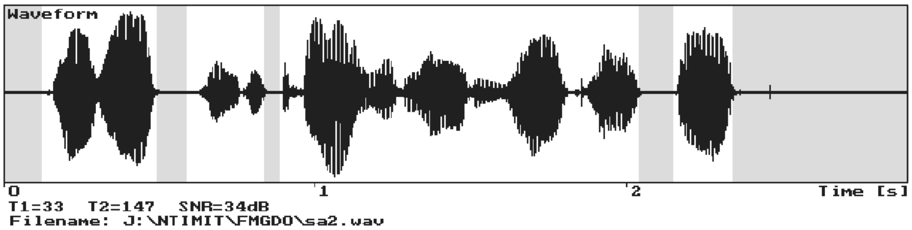

Fig. 3. NTIMIT sentence

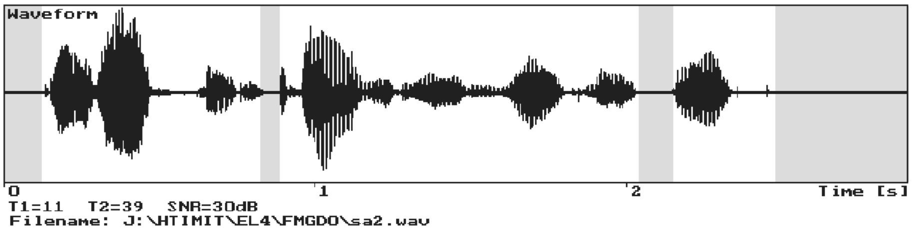

Fig. 4. HTIMIT-EL4 sentence

available databases and also using a particular speaker recognition program, it proved to be an accurate, adaptive and fast speech detection algorithm. As it was mentioned, the algorithm was designed for an off-line processing; nevertheless, it could be easily changed for an on-line version, with only a few minor modifications.

References

1. Cohn, R.P.: Robust Voiced/Unvoiced Speech Classification Using a Neural Net. Proceedings of ICASSP'91, Vol. 1, Paper S7.6 (1991).
2. Draganescu, M., Stefan G., Burileanu C.: Electronica functionala. Ed. Tehnica, Bucharest (1991), 443–480.
3. Herrera, A., Ramos A, Yamasaki K.: Speech Detection in High Noise Condition. Proceedings of ICSPAT'96, Boston, Vol. 2 (1996), 1774-1778.
4. Le Bouquin Jeannes, R., Faucon G.: Voice Activity Detector Based on the Averaged Magnitude Squared Coherence. Proceedings of ICSPAT'95, Boston, Vol. 2 (1995), 1964-1968.
5. Rabiner, L.R., Sambur M.R.: An Algorithm for Determining the Endpoints of Isolated Utterances. The Bell System Technical Journal, Vol. 54 (1975), 297-315.
6. Tolba, H., O'Shaughnessy D.: Voiced-Unvoiced Classification Using the First Mel Frequency Cepstral Coefficient. Proceedings of ICSP'97, Seoul, Vol. 1 (1997), 137-142.
7. Tucker, G.B., Spanias A.S., Loizou P.C.: A HMM-Based Endpoint Detector for Computer Communication Application. Proceedings of ICSPAT'95, Boston, Vol. 2 (1995), 1969-1973.