Robust Phonemes Detection Algorithm in Noisy Environments

Badia MOUNIR¹, Laila Elmazouzi², Ilham Mounir³, Abdelmajid FARCHI⁴

¹,²,³LIMMI Laboratory, Graduate School of Technology, University Cadi Ayyad, Morocco
⁴LIMMI Laboratory, Science and Technical Faculty, University Hassan 1er, Morocco.

Abstract—Speech endpoint detection is the problem of determining the beginning and ending of speech utterance. In this paper, we present new endpoint detection algorithm based two main stages. In the first stage, the speech signal is enhanced using frequency domain multiband spectral subtraction. The second stage is the algorithm based on spectral coefficients extraction. The performance of the proposed algorithm was evaluated for Arabic fricative phonemes in noisy environments with SNR ranging from 0 to 30 dB.

Keywords—Endpoint detection, frequency domain, spectral moment, speech enhancement.

I. INTRODUCTION

One major step which affects directly the performance of automatic speech recognition (ASR) systems is the detection of speech from non-speech. Too many non speech segments wrongly detected as speech and used in the training can corrupt the acoustic models, and hence reduces recognition accuracy. In a signal that has high signal-to-noise ratio (SNR), this can be solved simply by using an energy threshold [1, 2, 3]. However, when the signal is corrupted by noise, it Speech endpoint detection is the problem of determining the beginning and ending of speech utterance. can be very hard to distinguish between speech and non-speech. Depending on the surrounding environment of the recording, non speech can be silence or a variety of other acoustical signals such as cars or conversations. Such noises can be filtered out using speech enhancement techniques [2, 4, 5].

Noise can be stationary, that is, does not change over time, such as the noise coming from PCs. Noise can also be non stationary, such as conversations where the spectral and temporal characteristics of noise are constantly changing as people speaking in the background. Another distinctive feature of the various types of noise is the shape of their spectrum or the distribution of noise energy in the frequency domain. A number of algorithms have been proposed in the literature for speech enhancement [6] with the primary goal of improving speech quality. These algorithms can be divided into three main classes.

The first class contains the spectral subtractive algorithms based on the basic principle that as the noise is additive, one can estimate/update the noise spectrum when speech is not present and subtract it from the noisy signal [7, 8]. The second class concerns statistical model based algorithms: the speech enhancement problem is posed in a statistical estimation framework [9, 10, 11]. The third class is interested in subspace algorithms rooted primarily on linear algebra theory. The decomposition of the vector space of the noisy signal into “signal” and “noise” subspaces can be done using orthogonal matrix factorization techniques from linear algebra [12, 13].

In this article, the proposed endpoint detection algorithm consists of two main stages. In the first stage, the speech signal is enhanced using frequency domain multiband spectral subtraction. The second stage is the algorithm based on spectral coefficients developed in [14]. The performance of the proposed algorithm will be evaluated and compared to Sambur Rabiner algorithm [15] for Arabic phonemes CVC (C is a fricative consonant and V is a vowel) in presence of different type of noises. The rest of article is organized as follows. In section 2, we will present the proposed endpoint detection. In section 3, several experiments are conducted. The results show the efficiency of the proposed algorithm.

II. THE PROPOSED ENDPOINT DETECTION ALGORITHM

The proposed endpoint detection algorithm contains two main stages: speech enhancement and segmentation speech/ non speech.

II.1 Speech Enhancement

In the first stage of the proposed endpoint detection algorithm, we used Multi-Band Spectral Subtraction (MBSS) approach [16] to increase the accuracy of the proposed method. The motivation behind using multiband spectral subtraction stems from the fact that, in general, noise is unlikely to affect the speech signal uniformly over the whole frequency domain. In other words, some frequencies will be affected more adversely than the others, depending on the spectral characteristics of the noise.
In the multiband approach, the speech spectrum is divided into N non-overlapping bands, followed by spectral subtraction performed independently in each band.

The block diagram of the multiband method is shown in Fig. 1. First, the signal is windowed and the magnitude spectrum is estimated using DFT. Subsequently, the noisy speech spectrum is preprocessed to produce a smoothed estimate of the spectrum. Next, the noise and speech spectra are split into N frequency bands and the over-subtraction of each band is calculated. Then, the individual frequency bands of the estimated noise spectrum are subtracted from the corresponding bands of the noisy speech spectrum. Finally, the modified frequency bands are recombined and the enhanced speech signal is obtained by taking IDFT of the enhanced spectrum using the noisy speech phase.

### II.2 Segmentation Speech/Non Speech

In the second stage, we used the algorithm to separate speech segments from background noise (Figure 2). The spectrogram of a continuously recorded utterance is first derived. For each frame, the spectrum is obtained by fast Fourier transform (FFT). The probability density function for the spectrum can thus be estimated by normalization over all frequency components:

$$p(f_k) = \frac{P(f_k)}{\sum_{k=0}^{N/2} P(f_k)} \quad (1)$$

Where $P(f_k)$ is the power spectrum, $f_k = 2f_{Nq}k/N$, $k = 0, 1, \ldots, N/2$, and $N$ is the window length. $f_{Nq}$ indicates the Nyquist frequency.

The coefficients of the normalized power spectrum were computed as:

$$\text{mean}, \mu = \sum_{k=0}^{N/2} p(f_k)f_k$$

$$\text{standard deviation} \sigma = \sqrt{\sum_{k=0}^{N/2} (f_k - \mu)^2 p(f_k)}$$

$$\text{skewness} = \sum_{k=0}^{N/2} \left(\frac{f_k - \mu}{\sigma}\right)^3 p(f_k)$$

$$\text{kurtosis} = -3 + \sum_{k=0}^{N/2} \left(\frac{f_k - \mu}{\sigma}\right)^4 p(f_k)$$

To separate speech segments from background noise, we used two steps. In the first step, assuming that during the first 100 ms of the recording interval there is no speech present; some background silence statistics can be measured. Such statistics include the spectral centroid, standard deviation and skewness of the silence. These measurements are used to set two thresholds: mean threshold $\text{Th}_C$ and skewness threshold $\text{Th}_S$.

The second step consists of searching for the beginning point $N_1$ and the ending point $N_2$ of the utterance: By finding the first point at which the skewness exceeds $\text{Th}_S$, and then the mean exceeds $\text{Th}_C$.

A similar approach is used to define the endpoint of the utterance.
III. EXPERIMENTATION AND EVALUATION

Participants were ten Moroccan adult speakers of Arabic Modern Standard. The group included 5 female and 5 male subjects ranged in age from 19 to 25 years. They were asked to produce CVC (C: consonant and V: vowel) phonemes. The used consonants are fricatives /\fahrain/: /f/, /\zain/: /z/, /\saad/: /s/, /\nun/: /n/, /\zahid/: /z/, /\qinun/: /q/, /\lajnah/: /l/, /\hijz/: /h/ and /\yain/: /y/ with vowels /\alaf/: /a/, /\laam/: /i/ et /\waaw/: /u/.

The use of CVC sequences allows testing the ability of our algorithm to determine both the beginning and the ending of speech when using fricatives. The beginning and ending points of the speech utterance were labelled manually using both the time waveform and the spectrogram of utterances. This manual detection information is used as a reference to determine the accuracy of those detected by the proposed algorithm. The performance of the proposed algorithm is compared against that of the well-established and widely used algorithm of Sambur-Rabiner [15]. The speech segmentation boundaries, estimated using MATLAB by the proposed and reference algorithms, were compared against the boundaries obtained via manual segmentation. As a quantitative comparison metric, we have used the Mean error (ME) defined as given by:

$$ME = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\frac{(N_1(i) - N_2(i))^2 + (N_2(i) - N_3(i))^2}{N}}$$

(6)

Where \(N\) is the number of estimates and \(\hat{N}_1\) and \(\hat{N}_2\) are the estimate of \(N_1\) and \(N_2\).

We apply our algorithm and that of Sambur-Rabiner to CVC records in a clean environment. Table 1 summarizes the results for each fricative consonant. We can see that our algorithm has better performance than that of Sambur-Rabiner in terms of overall mean error. Figure 3 presents the performance of our algorithm and that of Sambur-Rabiner in the detection of the beginning and the ending of phonemes. The accuracy of our algorithm is due to the fact that although the amount of energy contained in the fricatives is low, their spectral mean remains significant. Then, the speech can be separated efficiently from noise.

<table>
<thead>
<tr>
<th>Consonants</th>
<th>The proposed algorithm</th>
<th>Sambur-Rabiner algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME %</td>
<td>N1 N2 Overall</td>
<td>N1 N2 Overall</td>
</tr>
<tr>
<td>/\fahrain/</td>
<td>0.6 0.8 0.7</td>
<td>5.7 6.7 6.2</td>
</tr>
<tr>
<td>/\zain/</td>
<td>1.2 1.5 1.35</td>
<td>6.2 5.9 6.05</td>
</tr>
<tr>
<td>/\saad/</td>
<td>1.4 2 1.7</td>
<td>8.2 6.4 7.3</td>
</tr>
<tr>
<td>/\nun/</td>
<td>0.9 0.7 0.8</td>
<td>7.5 5.3 6.4</td>
</tr>
<tr>
<td>/\zahid/</td>
<td>1.1 1.5 1.3</td>
<td>6.8 6.2 6.5</td>
</tr>
<tr>
<td>/\qinun/</td>
<td>0.8 1.1 0.95</td>
<td>5.8 5.4 5.6</td>
</tr>
<tr>
<td>/\lajnah/</td>
<td>0.7 1.3 1</td>
<td>4.6 5.2 4.9</td>
</tr>
<tr>
<td>/\hijz/</td>
<td>0.9 1.2 1.05</td>
<td>4.2 3.8 4</td>
</tr>
<tr>
<td>/\yain/</td>
<td>1.3 1.7 1.5</td>
<td>7.2 6.8 7</td>
</tr>
<tr>
<td>/\ch/</td>
<td>1.5 1.6 1.55</td>
<td>6.8 7.4 7.1</td>
</tr>
</tbody>
</table>
As the next step, the data signals were contaminated by various levels of three different types of measurement noises: white, car and conversation noises with SNR value in the range from 0dB to 30dB. The results obtained in the case of white noise contamination are summarized in Table 2. Our algorithm is observably superior to the reference one, especially at low SNRs when detecting the starting points and at high SNRs when detecting the ending points. The overall error has also been measured and visualized in Table 2. It can be seen that our algorithm outperforms the competitive algorithm at all SNR levels.

Tables 3 and 4 show the results of the conducted experiments in case of car noise and conversation noise, respectively. As evident from tables, our algorithm outperforms the reference method, especially in detecting the ending points. As one can see, the proposed algorithm has the best overall ME values, indicating a better performance.
IV. Conclusion

This paper proposed a new VAD technique to improve automatic speech recognition. The proposed feature extraction technique, based on the spectral moments and the MBSS speech enhancement approach was shown to be efficient and yet reliable. Experiments conducted in clean environment confirmed that the proposed algorithm is superior to the conventional algorithm. In addition, the proposed algorithm confirms its efficiency even in presence of different type of noises with SNR value ranging from 0 to 30 dB.

REFERENCES


