World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:9, No:1, 2015

Automatic Segmentation of the Clean Speech Signal

M. A. Ben Messaoud, A. Bouzid, N. Ellouze

***Abstract*—**Speech Segmentation is the measure of the change point detection for partitioning an input speech signal into regions each of which accords to only one speaker. In this paper, we apply two features based on multi-scale product (MP) of the clean speech, namely the spectral centroid of MP, and the zero crossings rate of MP. We focus on multi-scale product analysis as an important tool for segmentation extraction. The MP is based on making the product of the speech wavelet transform coefficients (WTC). We have estimated our method on the Keele database. The results show the effectiveness of our method. It indicates that the two features can find word boundaries, and extracted the segments of the clean speech.

***Keywords*—**Speech segmentation, Multi-scale product, Spectral centroid, Zero crossings rate.

I. Introduction

SPEECH segmentation is the problem of detecting word limits in vocal speech when the underlying vocabulary is still unrecognized. It is a vitally important task in many applications like information extraction, automatic transcription [1], speaker identification [2], and automatic speech recognition [3].

The features of segmentation influence the recognition quality in different ways. For best performance of language model, the segment extremity must accord to extremity of sentence like units. Furthermore, the silence, the noise regions produce inaccuracies and it should be removed. Also, the overlapped speech regions must be segregated to reduce the errors on the recognition of surrounding frames. Certainly, segment boundaries arranged inside a word can degrade the recognition performance. For the isolated word, the task is summarized to the elimination of the sound artefacts and the estimation of the correct word boundary. For the continuous speech case, the task is to reject of silences regions in addition to the artefacts.

Numerous speech segmentation methods have been proposed and are generally classified into three categories: model-based [4], [5], metric-based [6], [7], hybrid techniques [8], and decoder–guided [9]. The model based method is founded on a set of models. It is derived trained for many speaker categories from a training corpus.

The models are based on a support vector machines (SVM), hidden Markov models (HMMs) or Gaussian mixture models (GMM). It consists to locate the modification in the acoustic environment.

In the metric-based methods, we define the acoustic distance criterion and then the similarity of distance between two adjacent windows is estimated and a distance curve is formed. The feature information in each of the two adjacent windows is adopted to follow some probability density. The Bayesian information criterion, and KL distance have been used to compute the distance.

The Hybrid based techniques combine model and metric based methods.

In the decoder-guided, the speech is decoded, and then we cut off the speech at the silence regions, also apply the gender information to determine the segments.

For a continuous speech, many methods used the spectral energy to determine the endpoint**.** Typically a fixed threshold is applied based on the features of the energy to make a distinction between the voiced speech segments and the unvoiced or silence segments. It's not an efficient approach, as it tends to interrupt the ends of some voiced segments. The energy being very delicate to the amplitude of the speech sound will not result satisfactory results in the voiced/unvoiced decision. Then we decide to use the zero-crossings rate instead of the energy feature.

In this paper, we present a simple method for the detection of speech segmentation. The method is based on two features vectors. The first feature of extraction is the zero-crossings rate (ZCR) of the multi-scale product (MP), and the second is the spectral centroid of the MP.

This paper is organized as follows: the proposed segmentation method is described in Section II. Experimental results are presented in Section III, and conclusions are drawn in Section IV.

II. Proposed Approach

We propose a new method for segmentation of speech signal in the case of a single speaker. We apply two features, namely the spectral centroid of MP, and ZCR of MP. Both feature vectors are used for this purpose.

Our method can be decomposed into three steps, as shown in Fig. 1. The first step consists of calculating the product of the clean speech wavelet transform coefficients (WTC) at successive scales. We use the quadratic spline function wavelet at scales $s1=2^{-1}$, $s2=2^{0}$ and $s3=2^{1}$. It is a smooth function with property of derivative. The second step consists of computing the spectral centroid of MP, and the third step consists to calculate the zero crossings rate of the obtained signal.

M. A. Ben Messaoud, A. Bouzid and N. Ellouze are with the Electrical Engineering Department, Université de Tunis El Manar, Ecole Nationale d'Ingénieurs de Tunis, LR11ES17, Tunis, CO 1002 Tunisia (e-mail: anouar.benmessaoud@yahoo.fr, bouzidacha@yahoo.fr, n.ellouze@enit.rnu.tn).

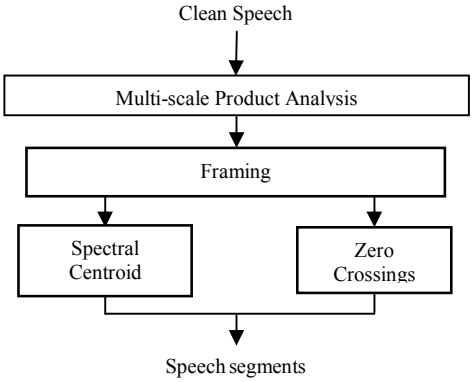

Fig. 1 Block diagram of the proposed method for speech segmentation

A. Multi-Scale Product Analysis

Wavelet transform (WT) was introduced as an alternative method for analyzing non stationary speech. It provides a new way for describing the speech signal into well be have expression that yields useful properties. Dyadic Wavelet Transform is the special case of continuous wavelet transform when the scale parameter is discretised along the dyadic grid (2^j), j=1, 2….

According to [10], the WT has shown excellent capacities for the detection of signal singularities. When the wavelet function has specific selected properties, WT acts as a differential operator. An appropriately chosen wavelet for discontinuity detection is a wavelet that is the second derivative of a smoothing function corresponding to the quadratic spline function.

The multi-scale (MP) analysis consists to producing the product of the WTC of the clean speech signal at three scales. The wavelet used in this analysis is the quadratic spline function at scales $s_1=2^{-1}$, $s_2=2^0$ and $s_3=2^1$. This step is described in our approach reported by [11].

The product $mp(k)$ of wavelet transforms coefficients of the speech $x(k)$ at some successive dyadic scales is given as follows:

$$mp(k) = \prod_{j=-1}^{j=1} W_{2^j} x(k) \quad . \tag{1}$$

where $W_{2^j} x(k)$ is the wavelet transform of the speech frame x at scale 2^j.

For the second step, the product $mp(k)$ is split into frames of N length by multiplication with a hamming window $h[k]$:

$$mp_h[k,i] = mp[k]h[k - i\Delta k]. \tag{1}$$

where i is the window index, and Δk the overlap.

Fig. 2 summarizes the steps of MP.

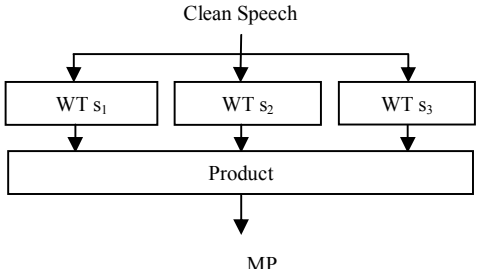

Fig. 2 Scheme of the speech multi-scale product

Fig. 3 presents the MP of a voiced speech frame.

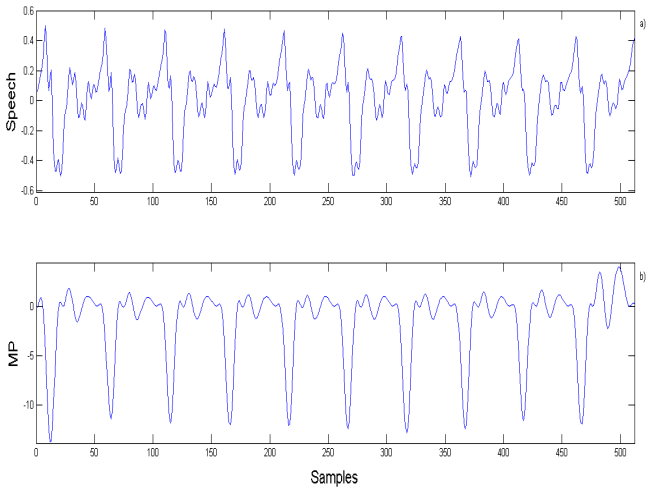

Fig. 3 Voiced speech (a) Voiced clean speech (b) Its multi-scale product

It indicates that the voiced speech MP has a periodic structure. It has a structure that remembers the derivative laryngograph signal. So, a spectral centroid analysis can be implemented on the MP of obtained speech.

B. Spectral Centroid of MP

The MP spectrum is the Fourier transform of the MP signal. The spectral centroid of multi-scale product (SDMP) of the k^{th} frame is described as the gravity of its spectrum MP, It's described as follows:

$$SDMP_k = \frac{\sum_{i=1}^{N} (i+1) P_k(i)}{\sum_{i=1}^{N} P_k(i)} . P_k(i) \tag{3}$$

where $i = 1,...,N$, N is the frame length, and $P_k(i)$ is the Fast Fourier Transform coefficients of the k^{th} frame of multi-scale product. This feature is a measure of the spectral position of MP.

Fig. 4 represents a voiced clean speech signal followed by its MP and the corresponding spectrum centroid of MP.

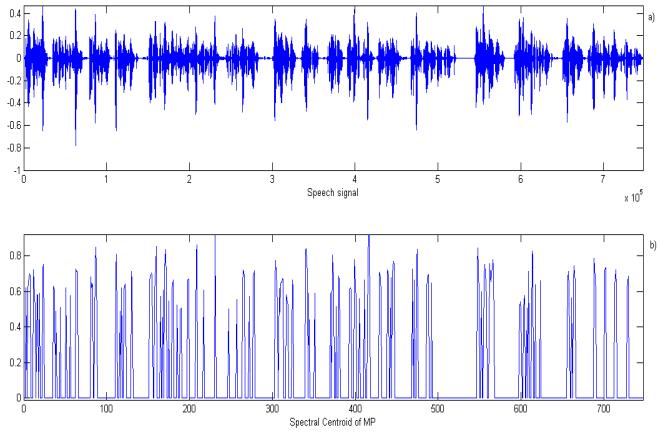

Fig. 4 Spectral centroid speech (a) Speech signal, (b) Spectral centroid of multi-scale product

We apply a simple threshold in order to remove the silence regions in the speech signal.

C. Zero Crossings Rate of MP

We have applied the MP because it gives a derived speech signal which is simpler to be analyzed. Then, we calculate the zero crossings rate of the MP. It can be defined as:

$$Z_i = \sum_{k=0}^{N-1} abs\left(\text{sgn}\left[mp(k)\right] - \text{sgn}\left[mp(k-1)\right]\right) h(i-k) \quad (4)$$

with

$$\text{sgn}\left[mp(i)\right] = \begin{cases} 1, mp(i) \geq 0 \\ 0, mp(i) < 0 \end{cases}$$

and

$$h(i) = \begin{cases} 1/2N, 0 \leq i \leq N-1 \\ 0, otherwise \end{cases}$$

The threshold of ZCR is determined experimentally.

III. Experiments

To evaluate the performance of our method, we employ the Keele database [12]. It contains ten English speakers (five female and five male English speakers) with duration between about 40 seconds. It includes reference files containing a pitch estimation of 51.2 ms segments with no-overlapping. The product MP is decomposed into frames of 1024 samples without overlapping at a sampling frequency of 20 kHz.

Fig. 5 depicts the original speech signal pronounced by a women followed by its multi-scale product (MP), its spectral centroid of MP, and finally the all segmented speech are formed by a successive frames with an example of detected voiced segments is presented on magenta colors.

Fig. 6 depicts the original speech signal pronounced by a men followed by its multi-scale product (MP), its spectral centroid of MP, and finally the all segmented speech are formed by a successive frames with an example of detected voiced segments is presented on magenta colors.

For the evaluation of the speech/silence detection method, we calculate the error decision probabilities that comprise speech content frames detected as silence noted (Active Error), and silence detected as speech content noted (Inactive Error).

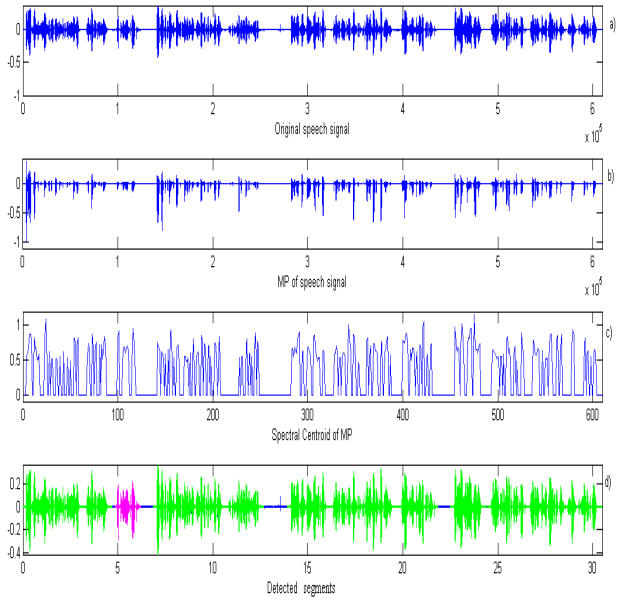

Fig. 5 Segmentation of speech signal (a) Speech signal pronounced by women "F3", (b) Multi-scale product, (c) Spectral centroid of multi-scale product (d) Detected segments

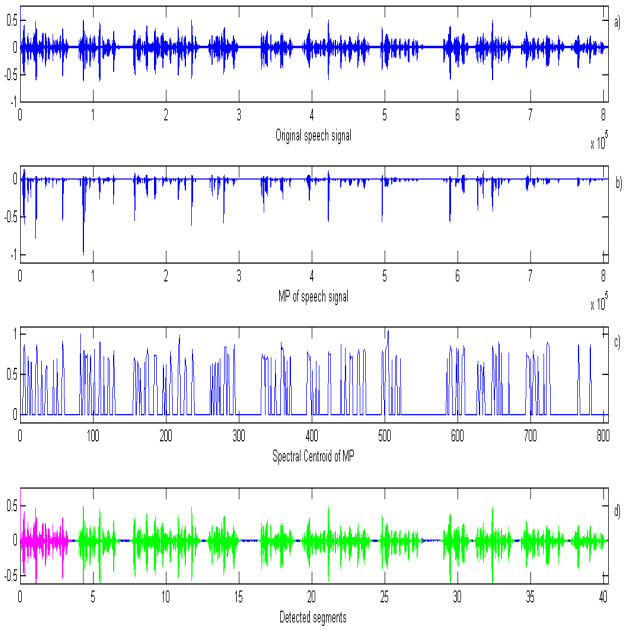

Fig. 6 Segmentation of speech signal (a) Speech signal pronounced by men "M5", (b) Multi-scale product, (c) Spectral centroid of multi-scale product (d) Detected segments

The table illustrates the speech/silence estimation results in clean speech.

TABLE I
Evaluation of the Proposed Method for Speech/Silence Estimation

Speakers	Active Error (%)	Inactive Error (%)
Females	1.23	0.35
Males	2.08	0.50

As reported in Table I, the proposed method shows the accurately classifying of speech frames as speech/silence classification.

IV. Conclusion

In this paper, we present a speech segmentation method that relies on the MP analysis of the speech. The proposed approach can be summarized in three essential steps. First, we make the product of speech WTC at three successive dyadic scales. Second, we calculate the spectral centroid of the speech MP. Thirdly, we compute the zero crossings rate of the MP. The experimental results show the efficiency of our method to remove the silence regions, and to extract the segments for clean single speaker speech. Future work concerns the extension of the proposed method for the segmentation in the presence of noise.

References

[1] F. Kubala, T. Anastasakos, H. Jin, L. Nguyen, and R. M. Schwartz. "Transcribing radio news," in Proc. *ICSLP,* 1996.
[2] L. Zhang, H. J. Lu, "Speaker change detection and tracking in real time news broadcasting analysis," in Proc. *ACM Multimedia,* 2002, pp. 602-610.
[3] S. E. Tranter, K. Yu, G. Evermann, and P. C. Woodland. "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," in Proc. *ICASSP,* Canada, 2004, pp. 753-756.
[4] J. Wang, H. Sung, and P. Lin, "Unsupervised change detection using SVM misclassification rate," *IEEE Trans. Computers,* vol. 56, pp. 1234–1244, 2009.
[5] I. McCowan, H. Bourland, and J. Ajmera, "speech/music segmentation using entropy," *Speech Comm.,* vol. 40, pp. 351–363, 2003.
[6] D. Wang, R. Vogt, M. Mason, and S. Sridharan, "Automatic audio segmentation using the GLR," in Proc. *International Conference on Signal process. Comm. Systems,* Australia, 2008, pp. 1-5.
[7] J. Hansen, and B. Zhou, "Unsupervised audio stream segmentation via the BIC," in Proc. *ICSLP,* 2000, pp. 714-717.
[8] D. Elter, T. Sikora, and H. Kim, "Hybrid speaker based segmentation system using MLC," in Proc. *International Conference on Acoustics, Speech and Signal Processing,* 2005, pp. 745-748.
[9] S. Tranter, and D. Reynolds, "Speaker diarization for broadcast news," in *the Speaker and Language Recognition Workshop*, *ODYSSEY'04*, 2004, Spain.
[10] S. Mallat, *A Wavelet Tour of Signal Processing The Sparse Way*. 3rd ed., Academic Press Elsevier, 2008.
[11] M. A. Ben Messaoud, A. Bouzid, and N. Ellouze, 2013. "An efficient method for fundamental frequency determination of noisy speech," in *LNAI 7911*, T. Drugman, T. Dutoit, Eds. Verlag Berlin Heidelberg: Springer, pp. 33–41.
[12] G. Meyer, F. Plante, and W. A. Ainsworth, "A pitch extraction reference database," in Proc. *EUROSPEECH,* Madrid, 1995, pp. 837–840.

Ben Messaoud Mohamed Anouar received the Ph. D. degree in Electrical Engineering from the National school of Engineer of Tunis in 2011. His research interests include topics in speech analysis and applications to engineering and computer science, particularly to pitch estimation, voiced decision, speech separation, and also speech enhancement.

Bouzid Aicha received Ph.D in Electrical Engineering from the National school of Engineer of Tunis (ENIT) Tunisia in 2004. His research interests include Speech Processing. Currently she is working as Professor in Department of Electronics Engineering, ENIT. His current research interests include Speech Processing and Image Processing.

Elllouze Noureddine is the laboratory's founder of Signal, Image, and Technology Information Labo (LSITI). His research interests include pattern recognition, signal processing and image processing applied in biomedical, and communication.