International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 10 5822 - 5824

Frame Based Single Channel Speech Separation using Sumary Autocorrelation Function

Raghi E.R
Electronics and Communication Department
Ilahia College of Engineering and Technology
Muvattupuzha, India
e-mail: raghi.er@gmail.com

Lekshmi M.S
Electronics and Communication Department
Ilahia College of Engineering and Technology
Muvattupuzha, India
e-mail: *lekshmims@gmail.com*

***Abstract*—** Single channel speech separation system is widely used in many applications. Pre-processing stage of Automatic speech recognition system, telecommunication system and the hearing aid design require the speech separation system to enhance the speech. This paper proposes a separation system that separates the dominant speech from the noisy environment, based on summary autocorrelation function (SACF) analysis pitch range estimation in the modulation frequency domain. Performance evaluation of the proposed system shows a better response compared to the existing methods**.**

Keywords- *Summary autocorrelation function; Modulation frequency domain; dominant speech*

*****

I. INTRODUCTION

In a natural environment, speech signal usually get mixed with background noise speech. In order to improve the quality of the speech, the speech separation system is often required in the pre-processing stage of many applications such as speech coding, automatic speech recognition system, telecommunication system, hearing aid design etc. The noise removal from the target speech is very critical for many applications and the need for a speech separation system is inevitable.

Separating target speech from the interference is very challenging topic in signal processing. Many works have been done in the single channel speech area for preserving quality of the target speech [1, 2].Hidden Markov models [3] is such a system which is capable to model both target and interference. But these works usually make assumption regarding the properties of target and interference speech which impacts the quality of the output.

Human listeners have the ability to recognize speech even in the presence of other interfering speech. The process behind this ability was named "auditory scene analysis" (ASA) by Bregman. Principle of ASA gave enlightenment in single channel speech separation. That is "computational auditory scene analysis" (CASA) [4] Many researchers were devoted in developing CASA system for single channel speech separation [4, 5].Wang and Brown model is such an example [6] which is based on the oscillatory correlation. But the system is not capable of handling the unresolved portion of the speech.Hu and Wang model [7] employed different method to segregate resolved and unresolved harmonics of the target speech and but the model is limited to segregate only the voice speech. Mahmoodzadeh model [8] uses onset and offset for pitch estimation. This system decomposes the acoustic mixture into time-frequency frames and create frequency mask for removing the interference part of the mixture. All these techniques require accurate pitch estimation method.

Pitch is an important attribute of voiced speech. The pitch determination is the most critical part of the speech separation system. Accuracy of the speech separation system is mainly concentrated in pitch estimation method. Pitch of resolved portion of the speech can be easily obtained. But the unresolved part of the speech is seemed to be very difficult to extract. This paper proposes an efficient method that estimates the pitch range of both resolved and unresolved portion of the speech by using summary autocorrelation function (SACF). And by using this pitch; the system segregates the target speech.

This paper is organized as follows. In section II, we first give a brief description of our system and then present the details of each stage. The results of the system are reported in section III. The paper concludes with a discussion in section IV.

II. SYSTEM DESCRIPTION

The main idea of the system is to remove the interference portion of the mixture signal to extract the target speech. Thereupon, at first modulation frequency of the acoustic mixture signal is computed and the pitch range of both target and interference speech is estimated with the help of SACF. Finally, by using this pitch range, a mask is created to separate the target speech. The block diagram of the proposed system is shown in figure.1. The system contains mainly four steps: T- F decomposition, modulation transform, pitch range estimation and speech separation. And the detailed description of the each block is as follows.

A. T-F Decomposition

The acoustic noisy input is a broadband signal. For the analysis, we want to convert this broadband signal into narrowband subband signal. At the T-F decomposition stage, input mixture signal is decomposed into narrowband signal. Here Short Time Fourier Transform (STFT) is used for decomposition. In STFT, short time signal is obtained by windowing the long time speech signal. Then fast fourier transform (FFT) of each windowed segment is computed.

$$X(m,k) = STFT\{x[n]\} \quad (1)$$

$$= \sum_{n=0}^{K-1} (x(n) w(mM - n) \exp(-j2\pi nk/K))$$

Where K is the DSTFT length, w (·) is the acoustic frequency analysis window with length L and M is the decimated factor. So after STFT, wideband input speech signal is divided into number of channels or frames.

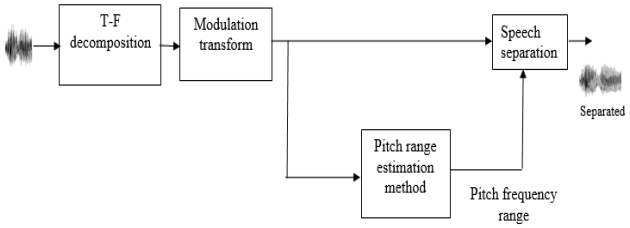

Figure.1 Basic block diagram of the proposed system

B. *Modulation Transform*

After the T-F decomposition, each channel can be regarded as the multiplication of two processes: a high frequency carrier and low frequency modulator. Narrowband frequency subband from the T-F decomposition module is divided into carrier and modulator signal. That is,

$$X(m, k) = M(m, k)\, C(m, k) \quad (2)$$

The modulator signal M(m,k) is obtained by applying an envelope detector to each channels, as

$$M(m, k) \triangleq D\{X(m, k)\} \quad (3)$$

Where D is the operator of the envelope detector. This modulator signal is used for further processing. Then, the time index m in the modulator signal is transformed into frequency index by taking Fourier transform.

C. *Pitch Range Estimation*

The correlogram is computed in modulation frequency domain by performing an autocorrelation at the output of each channel. correlogram detects the periodicities present in the each frames. It is an effective means for pitch estimation. The overall stages of pitch range estimation are shown in the figure 2.

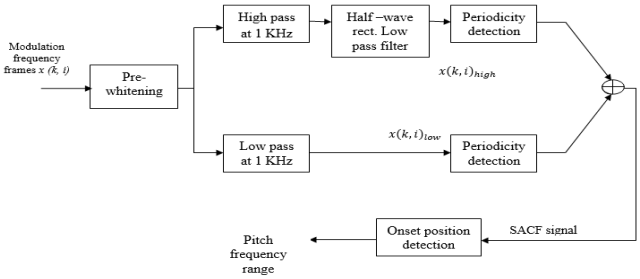

Figure.2 block diagram of the proposed system's pitch range estimation method

Most of the computational auditory scene analysis systems further process each frames in order to extract feature that are useful for grouping. Here each modulation frequency frame of mixture signal from the previous stage is processed by pre-whitening filter. Pre –whitening removes the short – time correlation of the noisy signal. The pre-whitening filter is implemented by using warped linear prediction (WLP). Next, Pre-whitened modulation frequency frames of mixture signal are splitting in to two bands, below and above 1 kHz. This is done with low pass and high pass filters. Thereby, resolved and unresolved part of the signal is separated and is treated differently. These filters have 12 dB/octave attenuation in stop band .Next step is half-wave rectification of high frequency signal. To obtain envelope, resulting high frequency signal is lowpass filtered. A generalized autocorrelation is then computed for the low-frequency band and the envelope of the high frequency band for periodicity detection. The autocorrelation can be performed in the modulation frequency domain by means of the discrete Fourier transform (DFT) and its inverse transform (IDFT). Speed of the computation is increased by using fast Fourier transform (FFT) and its inverse (IFFT). Then, the summation of the generalized autocorrelation of both high and low frequency frames corresponds to the SACF. That is,

$$x_2 = IDFT(|DFT(x_{low}|)^k + IDFT(|DFT(x_{high}|)^k \quad (4)$$

Where k determines the frequency domain compression and its value is smaller than 2. The SACF exhibits peaks at the period of each fundamental frequency. So the highest peak in resulting SACF signal corresponds to the pitch of the dominant speech. So the highest and lowest onset position detection determines the pitch range of both target and interference speech.

D. *Speech Separation*

After finding pitch range of target and interference speaker, system separates the target speech by using frequency masking. Here the system masked out the interfering speaker. For generating frequency mask, first we have to evaluate the mean of modulation spectral energy over the pitch frequency of both the target and interference signals. They can be represented as,

$$E_t^k = \frac{\sum_{i \epsilon Q^k} (|X(k,i)|)^2}{target\ \ pitch\ range} \quad (5)$$

$$E_I^k = \frac{\sum_{i \epsilon Q^k} (|X(k,i)|)^2}{Interferance\ \ \ pitch\ range} \quad (6)$$

Then compare the modulation spectral energy of the target and interference speakers by using this equation,

$$F^k = \frac{E_t^k}{E_t^k + E_I^k} \quad (7)$$

The resulting the frequency masking function is not applied in the modulation frequency domain directly. There are artifacts associated with it. So, frequency mask is transformed into the time domain by taking inverse FFT.ie,

$$f^k(m) = IFFT(F^k) \quad (8)$$

Then the speech is separated by convoluting the obtained filter $f^k(m)$ with the modulator signal of the mixture signal. And then, original target speech is reconstructed by multiplying the carrier signal.ie,

$$\tilde{X}(m, k) = [M(m, k) * f^k(m)]C(m, k) \quad (9)$$

From this target signal in the modulation frequency domain is obtained. To get separated target signal in the time domain, take the inverse STFT of $\tilde{X}(m, k)$.

III. EXPERIMENTAL RESULTS

A series of experiments have been conducted to evaluate the accuracy of the proposed method. We have taken samples

mixtures containing two male voices, two female voices, a male and a female voice with female dominant and male and female with male dominant. Separation performance was measured with signal-to-noise ratio (SNR), Perceptual Evaluation of Speech Quality (PESQ),and the overall quality .Also the proposed system was compared with the Mahmoodzadeh model. Table I show the SNR of mixture speech and separated speech of Mahmoodzadeh model and proposed model. Table II and III show PESQ and overall quality of Mahmoodzadeh model and proposed model respectively. All results show that the proposed system yields better performance than the Mahmoodzadeh model and the SNR of segregated speech is improved from the mixture. From the Welch power spectral density of separated speech in figure.3, it is clear that the dominant speech can be separated without loss of information.

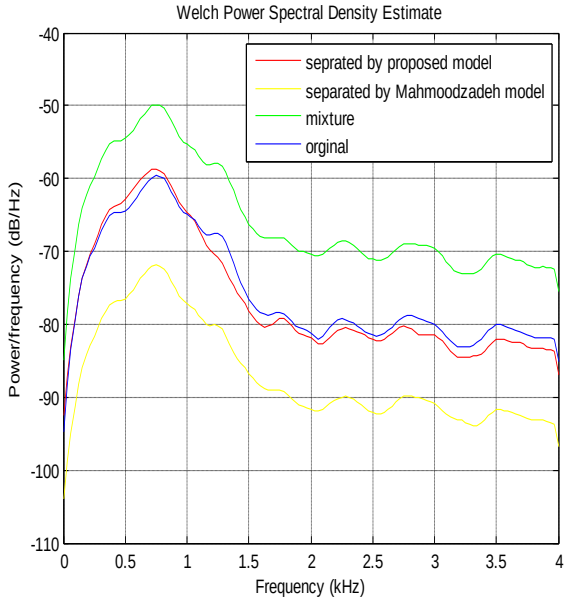

Figure.3 Welch power density of original, mixture and separated speech

A- Mixture of Two female speakers.
B- Mixture of Two male speakers.
C- Mixture of male & female speakers with female dominant.
D- Mixture of male & female speakers with male dominant.

TABLE I
SNR RESULTS FOR SEPARATED AND ORIGINAL MIXTURES

Type of mixture	Signal to Noise Ratio(dB)		
	Mixture	Mahmoodzadeh model	Proposed system
A	-6.6295	2.255	7.8569
B	-6.4927	3.4619	9.3763
C	-6.1279	2.1866	7.7758
D	-8.5507	1.3483	4.9985

TABLE II
PESQ OF SEGREGATED SPEECH OF MAHMOODZADEH MODEL AND PROPOSED MODEL

Type of mixture	PESQ	
	Mahmoodzadeh model	Proposed system
A	2.3142	2.3736
B	2.5326	2.9021
C	2.5204	2.5339
D	2.6588	2.7578

TABLE III
OVERALL QUALITY OF SEGREGATED SPEECH OF MAHMOODZADEH AND PROPOSED MODEL

Type of mixture	Overall Quality	
	Mahmoodzadeh model	Proposed system
A	1.4446	4.0110
B	4.3196	4.4579
C	4.0355	4.0908
D	3.0810	4.1378

IV. CONCLUSION AND DISCUSSION

In this paper, we presented a new single channel speech separation system that estimates the pitch range by analyzing summary autocorrelation function in modulation frequency domain. It results good estimate of pitch range to produce frequency mask and separate the target speech from the interfering speech. From the signal to noise ratio, PESQ and overall quality, it is clear that the proposed system is superior to the existing Mahmoodzadeh model.

REFERENCES

[1] G Hu, D Wang, "A Tandem algorithm for pitch estimation and voiced speech segregation," IEEE Trans Audio Speech Lang Process. 18(8),2067–2079 (2007).

[2] JJ. Jensen and J. H. L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model,"IEEE Trans. Speech and Audio Process., vol. 9, pp. 731-740, 2001.

[3] P. Varga and R. K. Moore, "Hidden Markov model decomposition ofspeech and noise," Proceedings of ICASSP, pp. 845-848, 1990.

[4] Brown GJ, Wang DL (eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Applications (Wiley & IEEE, Hoboken, NJ, 2006)

[5] M Buchler, S Allegro, S Launer, N Dillier, "Sound classification in hearing aids inspired by auditory scene analysis",*EURASIP J Appl Signal Process*. 18, 2991–3002 (2005) .

[6] D.L. Wang and G.J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Net*., Vol. 10, pp. 684-697, 1999.

[7] G. Hu, and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135-1150, 2004.

[8] A.Mahmoodzadeh, H.R Abutaledi ,H.Soltanian,H.Sheikhzadeh,"Single channel speech separation with a frame – based pitch range estimation method in modulation frequency,"*IEEE Trans Audio Speech Lang Process*,2012