Single Channel Speech Separation Based on Modulation Frequency Domain

Farhana Yousaf[1] Lekshmi M.S[2]
[1,2]PG Student
[1,2]Department of Electronics & Communication Engineering
[1,2]Ilahia College Of Engineering And Technology, Ernakulam

Abstract— Computational auditory scene analysis (CASA) is the study of auditory scene analysis (ASA) by computational means .In recent literature CASA focus for speech separation from monaural mixtures. It is based on the cochlear modeling using short time Fourier transforms (STFT).While the segregation efficiency and perceptual quality is not good; hence it is not applicable for an efficient hearing aid. The proposed system decomposes the input signal into segments by using Wavelet Packet Transform (WPT).Then the pitch range in each frame is calculated by Onset and Offset algorithm. Speech separation is performed by a mask extracted from the modulation spectrogram. Systematic evaluations show that the proposed system outperforms the previous system.

Key words: ASA, CASA, Onset-Offset, Short Time Fourier Transform, Wavelet Packet Transform

I. INTRODUCTION

The hearing system, even in front of complex auditory scenes and in unfavorable conditions, is able to separate and recognize auditory events accurately. A great deal of effort has gone into the understanding of how, after having captured the acoustic data, the human auditory system processes them. There are some problems related to Hearing Impaired people wearing hearing aid. ie, difficulty in understanding speech contaminated by speech from other talkers and difficulty in understanding the speech if speed of the speech is more. This paper aims to solving the first problem; it is treated as "Cocktail party effect". So if we can separate the dominant speech from the mixture and then amplify it. It will helpful for people having hearing impairment. Many works have been done for retaining the quality of speech in single channel speech separation area [1]-[2]. The main principle behind computational auditory scene analysis (CASA) is Auditory Scene Analysis (ASA). CASA systems are "machine listening" systems that aim to separate the mixture of sound sources. For monaural speech enhancement there are some algorithms are developed [3]-[4] and these are related on some analysis of speech or interference and speech amplification or noise reduction. Another method of speech enhancement by using Eigen-decomposition [5] on an acoustic mixture and then subspace analysis for removing interference. Wang and Brown [6] is another model for speech segregation which is based on the oscillatory correlation, but the main problem of this system not capable of handling the unresolved components of the speech. Hu and Wang developed another model [7] which can segregate both resolved and unresolved harmonics of target speech, but this model limited for segregation of voiced speech. An onset-offset based speech segregation technique was implemented by Mahmoozadeh [8]. By determining the onset and offset fronts from the onset-offset values, and these fronts are used for segmentation and grouping. This paper organized as follows. In section 2 gives a short description of our proposed system and details of each stage. The simulated results are reported in section 3. This paper concludes with a discussion in section 4.

II. SYSTEM DESCRIPTION

Aim of our proposed system is to remove interference from the mixture signal to extract the target speech. For that, at first modulation frequency of the acoustic mixture is computed target and interference pitch was calculated by using onset and offset algorithm. The system mainly contains four steps: Wavelet Packet Transform (WPT), modulation transform, pitch range estimation and speech separation. The block diagram of our proposed system is shown in Fig 1.The detailed description of the each block is given as follows.

A. Wavelet Packet Decomposition

The acoustic noisy input is a broadband signal. For analysis we convert this broadband signal into narrowband subband signal by using Wavelet Packet Transform (WPT). In WPT, the decomposition of input signal is performed by passing the signal through different filters. ie, discrete-time low and high pass quadrature mirror filters. For n levels of decomposition WPT produces 2^n different sets of coefficients (or nodes). The discrete wavelet packet transform of the input signal can be expressed as follows

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

B. Modulation Transform

The input signal X(m,k) can be represented as the product model of the Modulator Signal M(m,k) and the Carrier Signal C(m,k) is defined as

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

The modulator signal can be obtained by applying an envelope detector to this signal as

$$M(m,k) \cong ev\{X(m,k)\} \tag{3}$$

Where 'ev' is an operator of the envelope detector, which is an incoherent detector, is able to create a modulation spectrum that has large area covered in modulation frequency domain. Here Hilbert envelope detector is used for envelope detection. The modulator of the complex signal X(m,k) is defined as

$$M(m,k) \cong |X(m,k)| \tag{4}$$

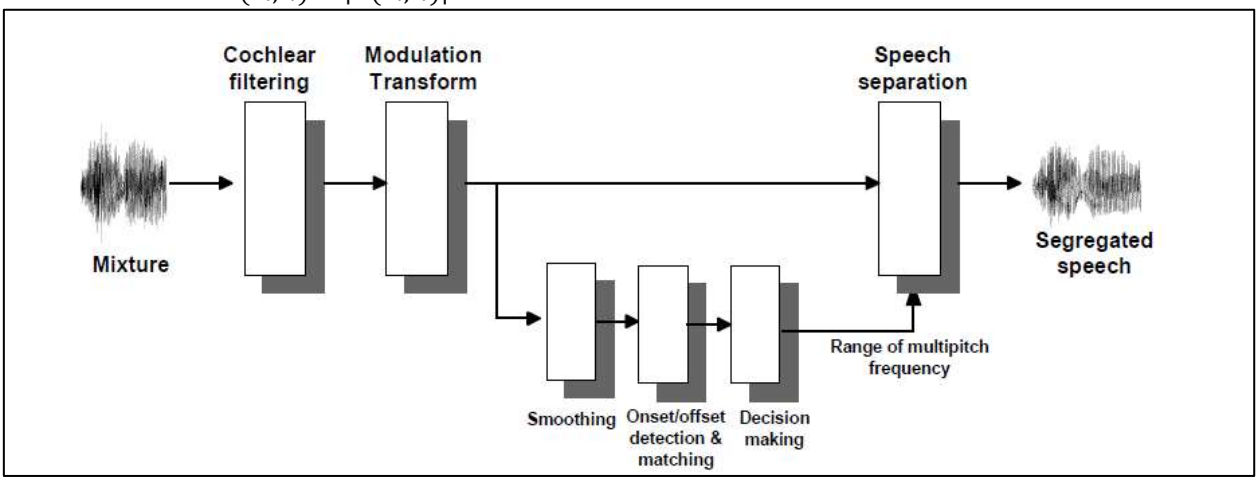

Fig. 1: Basic block diagram

The modulation frequency domain of the acoustic mixture signal is obtained by applying WPT followed by a sub band envelope detector and then a frequency analyzer of the sub band envelopes (the DFT).The discrete short-time modulation transform of the signal x(n) is defined as

$$X(k,i) = DFT\{ev\{WPT\{x(n)\}\}\}$$
$$= \sum_{m=0}^{I-1} M(m,k)e^{\frac{-j2\pi mi}{I}} \qquad \text{i} = 0, \ldots\ldots, \text{I} - 1, \tag{5}$$

C. Smoothing

The intensity of modulation spectrogram of the input signal is first smoothed over the modulation frequency by using low-pass filter. The intensity fluctuations can be reduced by smoothing.ie, the local details of fluctuations become blurred and the major intensity changes still preserved. The smoothed intensity for the output of modulation transform stage X(k,i) is expressed as follows

$$X_s(k,i) = X(k,i) * g_s(i) \tag{6}$$

Where $g_s(i)$ is a low-pass filter with small no of coefficients having pass-band [0,s] in Hz. Next step is to find the onsets and offsets from the smoothed intensity over the modulation frequency .Onsets and offsets are sudden intensity changes, in order to find this we take the partial derivative of the smoothed modulation spectrogram intensity is obtained as

$$\frac{\delta}{\delta i}(X_s(k,i)) = \frac{\delta}{\delta i}(X(k,i) * g_s(i)) \tag{7}$$

D. Onset And Offset Detection And Matching

Peaks and valleys of the signal equation (7) are, respectively, denoted as onset and offset candidates. The peaks corresponding to the true onsets are usually significantly higher than other peaks. Because of this reason a threshold $\theta_{on} = \mu+\sigma$ is selected, where μ and σ are mean and standard deviation of all the onset candidates, respectively. The onset candidates with peaks bigger than threshold θ_{on} are accepted and other onset candidates are eliminated. The same procedure is applicable for corresponding offset candidates. If there are multiple offset candidates in between two onsets, the offset candidate with the largest intensity decrease (ie, smallest $\frac{\delta}{\delta i}(X_s(k,i))$ is chosen.

After finding the onsets and offsets, those with close modulation frequencies may be corresponds to same source. These are connected to the onset and offset fronts. Our system connects an onset candidate from a filter channel to an onset candidate in the above filter channel, if their distance in modulation frequency may be less than a certain threshold compared with latter filter channel. In every filter channel, this threshold is defined by mean of the distances in the modulation frequency direction between two adjacent onsets. The same procedure applies to offset candidates. The next step is to form segments by matching individual onset and offset fronts [9].

E. Frequency Masking

A frequency mask is created for speech separation in the modulation spectrogram domain by assuming the pitch ranges of target and interference are known and these ranges are same in every subband. The proposed system provides the value of mask in each filter channel, which depends on the estimated pitch range of that filter channel. Assume the input signal x(n) sampled at rate fs is a mixture of both target signal $x_{ts}(n)$ and interference signal $x_{is}(n)$, which is represented as

$$x(n) = x_{ts}(n) + x_{is}(n) \quad (8)$$

For generating the frequency mask, first we have to calculate the mean of modulation spectral energy over the pitch frequency of both target and interference signals. It can be represented as

$$G_T(k) = \frac{\sum_J |X(m,k)|^2}{target\ pitch\ range} \quad (9)$$

$$G_I(k) = \frac{\sum_J |X(m,k)|^2}{Interference\ pitch\ range} \quad (10)$$

Then the frequency mask calculated as,

$$F^k = G_T(k)/[G_T(k) + G_I(k)] \quad (11)$$

The resulting frequency mask is not directly applicable in the modulation frequency domain. There is some artifacts associated with it. By taking inverse FFT the frequency mask is transformed into time domain is defined as

$$f^k(m) = IFFT(F^k) \quad (12)$$

Then the target speech signal is separated by convoluting the obtained filter $f^k(m)$ with the modulator signal of the mixture signal and then multiplied with the carrier signal is expressed as

$$\tilde{X}(m,k) = [M(m,k) * f^k(m)]\, C(m,k) \quad (13)$$

In order to get the separated target signal in time domain by taking inverse STFT of $\tilde{X}(m,k)$

III. EXPERIMENTAL RESULTS

The proposed algorithm uses five level wavelet packet transform for decomposing the input mixture into different subbands. A series of experiments have been conducted to determine the accuracy of our proposed system by taking the sample of mixtures containing male and female voices, mixture of female voices, mixture of male voices and mixture of three voices. Matlab programs with the existing cochlear and the proposed Wavelet packet filter bank in the analysis phase are simulated, similar database are used for existing and proposed models respectively. In all the cases of mixtures, the dominant speech can be separated without loss of information. Separation performance was evaluated with signal-to-noise ratio (SNR), Perceptual Evaluation of Speech Quality (PESQ). it is the numerical indication of the perceived quality of reconstructed signal after separation. PESQ is expressed by a single number in the range 1 to 5. Where 1 is the lower perceived quality and 5 is the higher perceived quality.

Signal-to-Noise ratio compares the level of a desired signal to the level of background noise. The higher the ratio, the less obstructive the background noise is. SNR is defined by equation

$$SNR = 10 \log \Sigma x(n)^2 / \Sigma (x(n) - \widehat{\widehat{x(n)}})^2 \quad (14)$$

Where x(n) is the original signal before mixing and $\widehat{x(n)}$ is the reconstructed speech from the mixture. The SNR and PESQ of mixture speech, separated by existing model and proposed model are shown in Table 1 and 2 respectively.

Signal to Noise Ratio(dB)			
Type of mixture	Mixture	Existing system	Proposed system
A	-9.0595	5.2271	7.6960
B	-9.6290	3.5565	9.8744
C	-9.0413	0.4680	2.1711
D	-9.6391	5.3651	10.3373
E	-9.0563	-0.3233	2.4564

Table 1: SNR results for separated and original mixtures

- Mixture of Two female speakers
- Mixture of Two male speakers
- Mixture of male and female speakers with female dominant
- Mixture of female and male speakers with male dominant
- Mixture of three female speakers

All results show that the proposed system yields better performance and SNR of segregated speech is increased from mixture. By analyzing the Waveforms ie, Welch power spectral density, Spectrogram and T-F plot of mixture speech, separated speech by using two cochlear models shown in fig 2,3 and 4 respectively, it is clear that the dominant speech can be separated without loss of information.

PESQ		
Type of mixture	Existing system	Proposed system
A	2.429	2.978
B	2.984	3.287
C	2.167	3.038
D	2.408	3.205
E	2.303	2.849

Table 2: PESQ of segregated speech of existing and proposed model

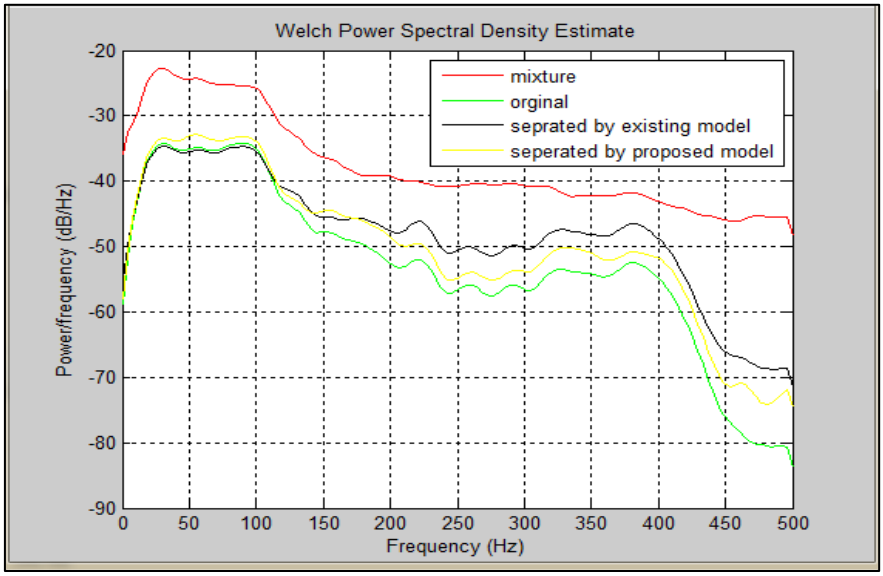

Fig. 2: Welch power density of original mixture and separated speech

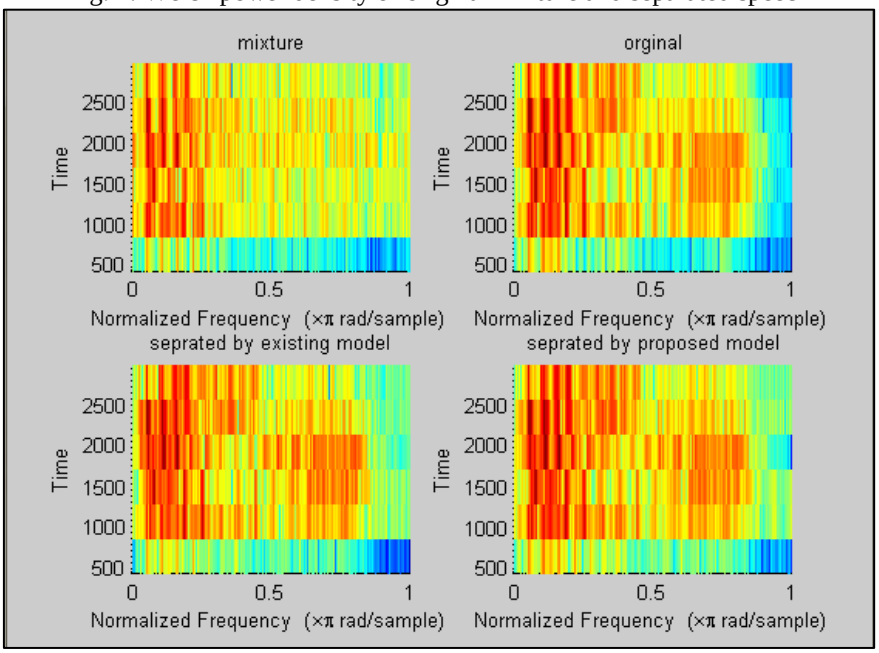

Fig. 3: spectrogram of signals

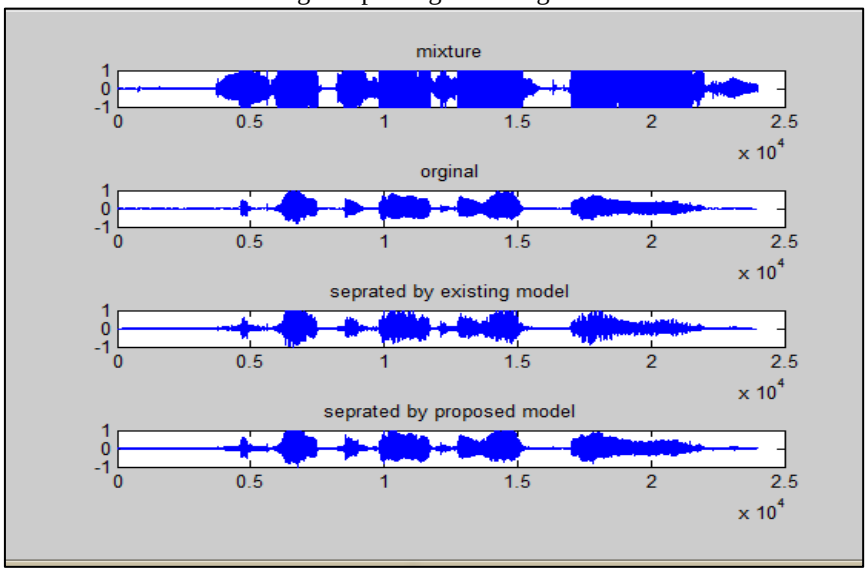

Fig. 4: T-F plot of signals

IV. CONCLUSION

In this paper, we presented a new single channel speech separation system based on wavelet packet transform. By analyzing the values of signal to noise ratio, perceptual evaluation of quality (PESQ), it is clear that the proposed wavelet filter bank is superior to the existing STFT. And a wavelet packet filter bank operates faster than the STFT filter bank. As we implemented this system as the preprocessing stage of a digital hearing aid, real time realization is very important and hence wavelet packet filter bank can be used in the analysis phase. The efficiency of the system depends on the pitch estimation algorithm.

REFERENCES

[1] 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.
[2] G Hu, D Wang, "A Tandem algorithm for pitch estimation an voiced speech segregation," IEEE Trans Audio Speech Lang Process. 18(8),2067–2079 (2007).
[3] J. Benesty, S. Makino, and J. Chen, Ed., Speech enhancement, NewYork: Springer, 2005.
[4] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, "A brief survey of speech enhancement," in The Electronic Handbook, CRC Press, 2005.
[5] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancement," IEEE Trans. Speech and Audio Process., vol. 9, pp. 87-95, 2001.
[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. Abutalebi, H. Soltanian-Zadeh, H. Sheikhzadeh "Single Channel Speech Separation with a Frame-based Pitch Range Estimation Method in Modulation Frequency"
[9] A. Mahmoodzadeh , H. R. Abutalebi , H. Soltanian-Zadeh , H. Sheikhzadeh "Single channel speech separation in modulation frequency domain based on a novel pitch range estimation method".