# Dominant voiced speech segregation based on Onset Offset analysis

Shibani .H

**Abstract- Computational Auditory Scene Analysis (CASA) has been the focus in recent literature for speech separation from monaural mixtures .The recent literature is based on the cochlear modeling using gamma-tone filter bank .While the computational complexity associated with gamma-tone filter bank is high; hence it is not applicable for an efficient hearing aid.**

**Index Term- Cochlear filter, Frequency Mask, Monaural speech, Ideal Binary Mask, Onset-Offset , Segregation.**

## 1 INTRODUCTION

IN natural environment, speech from a single source undergoes continuous acoustic deterioration such as, additive noises from other channels, reverberations from surface reflections etc. While many of the applications in audio signal processing such as Automatic speaker recognition, telecommunication, and Hearing aid applications etc requires an effective way to segregate the target speech from the monaural mixtures. The human have the ability to automatically segregate the speech and can focus to the target speaker even with one year. This perceptual property is known as Auditory Scene Analysis (ASA). Research and development in ASA will leads to the development of Computational Auditory scene analysis (CASA).

Various algorithms have been proposed for monaural speech enhancement, [1][2]and they are generally based on some analysis of speech or interference and subsequent speech amplification or noise reduction. Another method in dealing with speech separation is to perform Eigen-decomposition [3]on an acoustic mixture and then apply subspace analysis to remove interference. Hidden Markov models have been used to model both speech and interference and then separate them [4][5].All these technique requires very accurate pitch estimation, which is a difficult task.

An onset-offset based speech segregation technique is employed in Mahmoodzadeh [6] method. This algorithm determines onset and offset fronts from the onset –offset values, and these fronts are used for segmentation and grouping.

- *Author Shibani.H is currently pursuing masters degree program in Applied electronics engineering in Ilahia college of engineering &tech., Mahathma Gandhi University, India, E-mail: Shibani.h@gmail.com*

This paper proposes an incoherent modulator signal analysis and onset offset based approach for target speech signal separation from monaural mixtures. Also, the computational complexity associated with the gamma – tone filter can be avoided here by replacing it with discrete modulation Transform.

## 2 SYSTEM DESCRIPTIONS

The main aim of the proposed system is to produce a mask for single channel speech separation. Thereupon, at first the modulation spectrum of the speech signal is calculated Discrete Short Time Modulation Transform (DSTMT)[7]. Then the pitch frequency range of the Target and interference signals are calculated by means of onset offset based binary masking, and the pitch frequency range is used for the generation single channel speech segregation. The proposed multistage system is in fig: 1

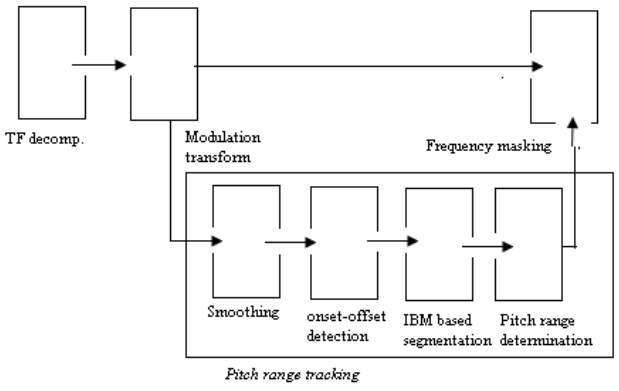


Fig1:Basic block diagram

### 2.1 T-F Decomposition.

The T-F Decomposition achieved from STFT (short Time Fourier Transform) , In this case, the data to be transformed

could be broken up into chunks or frames . Each chunk is Fourier transformed, and the complex result is added to a matrix, which records magnitude and phase for each point in time and frequency. This can be expressed as:

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

$$=\sum(s[n]\omega[n-m]\exp(-j\omega n)\,.$$

S(m, k) is a T-F transformed narrowband signal (with the time index m) coming out of the k th channel. Where s[n] represents signal and that of window is w[n].

### 2.2 Modulation Transform

The signal S(m,k) can be represented as the product of Modulator Signal M(m,k) and Carrier Signal C(m,k).

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

The modulator signal of S(m,k) can be determined from the signal itself by the analysis of envelop detection.

$$M(m, k) \cong ev\{S(m, k)\} \quad (3)$$

Where 'ev' is an operator for envelop detection.
Envelope detector is the incoherent detector based on Hilbert envelope [8 ], since it is able to create a modulation spectrum that has a large area covered in the modulation frequency domain. for complex-valued sub bands ,it acts as a magnitude operator as in eq (4).

$$M(m, k) \cong |S(m, k)| \quad (4)$$

Then the information regarding modulation frequency can be obtained by evaluating the Fourier transform of the modulating signal M(m,k).Then the Discrete Short time Modulation Transform of the signal s(n) can be defined as,

$$S(k, i) = DFT\ \{ev\{STFT\{s(n)\}\}\}$$

$$=\sum_{n=0}^{K-1} M(m,k)e-\frac{j2\pi mi}{I} \quad (5)$$

### 2.3 Onset-Offset Position analysis

Many of the CASA algorithms are generally based on some analysis of speech or interference and subsequent speech amplification or noise reduction .While all these technique requires very accurate pitches estimation, which is a difficult task in itself for single speaker, even more complex in the presence of interfering speaker. This problem can be avoided by the onset –offset based algorithm.
In this approach at first the signal after modulation transform is smoothed using a low pass filter. Then its partial derivative with respect to modulating frequency will helps to easily determine the peaks and valleys of the signal referred as onset position and offset position respectively.

### 2.4 Binary Mask Segmentation

The next step is to form segments by matching Onset and offset positions. It can be achieved by means of an ideal binary mask. The ideal binary mask can be defined as,

$$IBMt,f)=1\ ;\ \text{if } f = f_{on} + \left(\frac{\rho fs}{N}\right) \quad (6)$$

Where, $f_{0n}$ is onset position obtained from onset offset analysis. ρ takes values from -10 to 10.Then the masked signal can be represented as

$$S_{IBM}(t,f)=\begin{cases} s(t,f)\text{if} & \text{if } f = f_{on} \\ 0 & \text{else} \end{cases} \quad (7)$$

The pitch range of the dominant signal can be determined from this masked signal. Similarly the pitch range of interference can be determined from the remaining part of the mixture. Using these pitch ranges we can estimate a proper mask for segregating the target signal from the interference signal.

### 2.5 Frequency Masking

Assume the input signal s(n) sampled at rate fs is a mixture of both the target signal $s_t$ (n) and the interference signal $s_i$ (n).

$$s(n)= s_t(n)+ s_i(n) \quad (8)$$

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

$$X_T(k)=\Sigma(S(m,k)^2)/(\text{target pitch range}) \quad (9)$$

$$X_I(k)=\Sigma(S(m,k)^2)/(\text{Interference pitch range}) \quad (10)$$

Then the frequency mask is calculated as,

$$F(k,i)= X_T(k)/[\,X_T(k)+ X_I(k)] \quad (11)$$

The filter can be designed by taking the inverse Fourier transform followed by the multiplication of the phase response. The obtained filter is used to separate the target speech by convolution.

$$S_t(k,m)=S(k,m)*F(k,m) \quad (12)$$

## 3 RESULTS

In the proposed algorithm were set at K = 512 and I = 512 , and h (n ) and g (m ) were a 48-point and 78-point Hanning windows. The separation performance of the modulation masks was measured with the signal-to-distortion ratio (SDR).

$$SDR=10 \log \Sigma S_t(n)^2/\Sigma(S(n)-S_t(n))^2 \quad (13)$$

TABLE I.
RESULT BASED ON SDR

| | | | | | |
|---|---|---|---|---|---|
| SDR (mixture) | 11.4671 | 13.1495 | 15.2378 | 17.992 | 22.0508 |
| SDR (separated) | 21.2584 | 24.0134 | 28.0714 | 35.935 | 42.4489 |

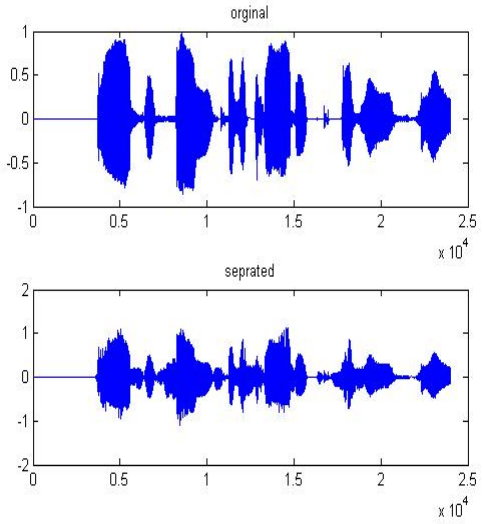


Fig3 : Original and target signal s along time axis.

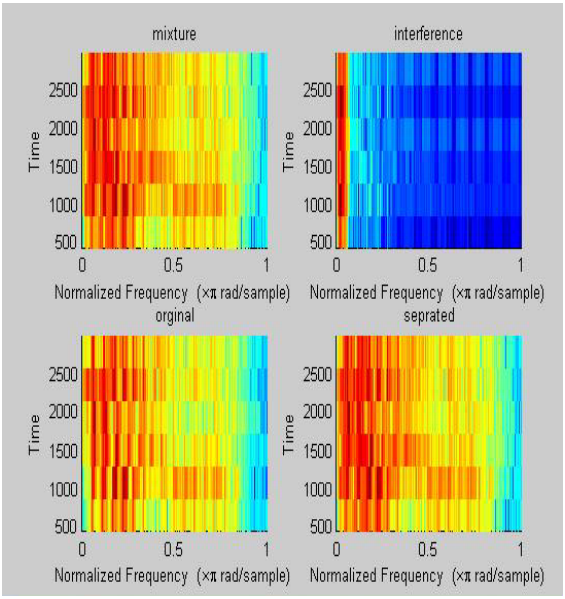


Fig4: time –frequency plot of mixture, Interference ,Original and Target signals

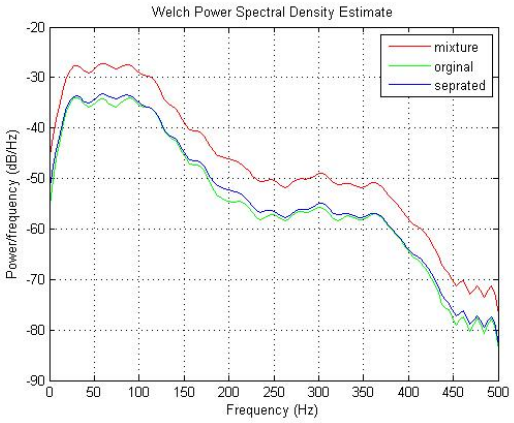


Fig2. Welch Power Spectral Density estimate of mixture, target signal and separated signal

## 4 CONCLUSIONS AND DISCUSSION

In this paper, we presented a new approach for monaural speech segregation based on onset offset analysis and ideal binary mask based segmentation. The proposed method is simple with reduced computational complexity and higher signal to noise ratio.

## ACKNOWLEDGMENT

The author wish to thank Mrs.Lekshmi.M S (Asst.Professor, ICET) ,for the valuable guidance and assistance throughout the work.

## REFERENCES

[1] J. Benesty, S. Makino, and J. Chen, Ed., Speech enhancement, NewYork: Springer, 2005.
[2] Y. Ephraim, H. Lev-Ari, and W. J. J. Roberts, "A brief survey of speech enhancement," in The Electronic Handbook, CRC Press, 2005.
[3] A. Rezayee and S. Gazor, "An adaptive KLT approach for speech enhancment," IEEE Trans. Speech and Audio Process., vol. 9, pp. 87-95, 2001.
[4] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," Proceedings of ICASSP, pp. 845-848, 1990.
[5] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," IEEE Trans. Speech and Audio Process., vol. 6, pp. 445-455, 1998.
[6] 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
[7] A. Mahmoodzadeh , H. R. Abutalebi , H. Soltanian-Zadeh , H. Sheikhzadeh".Single Channel Speech Separation with aFrame-based Pitch Range Estimation Method inModulation Frequency", EURASIP Journal on Advances in Signal Processing 2012
[8] R Drullman, JM Festen, R Plomp, Effect of temporal envelope smearing on speech reception. J Acoust Soc Am. 95, 1053–1064