International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

# Least Squares Fitting Approach For Monaural Speech Segregation

**Fayisa Yousef[1], Robin Abraham[2]**

*[1]PG Scholar,Dept.of Electronics and Communication Engineering,Ilahia College of Engineering and Technology,Kerala,India*
*[2] Asst.Professor &HOD of Electronics and Communication Engineering,Ilahia College of Engineering and Technology,Kerala,India*

---

**Abstract -** *Speech segregation is the separation of a desired speech signal from a mixture of environmental signals. This paper introduces an algorithm to isolate speech streams from a single-channel speech mixture. Current speech segregation algorithms assign speech regions to participating speakers depending on which speaker dominates in which spectro-temporal region. The proposed method is a different approach to speech segregation. In this paper wavelet transform and Lease-Squares fitting approach are combined for single channel speech separation. The proposed system decomposed the input speech signal into a number of channels by using Wavelet Packet Transform (WPT).Then the multipitch tracking in each frame is calculated by Enhanced Summary Auto-Correlation Function(ESACF) algorithm.*

***Key Words***: **speech segregation, co-channel speech, monaural speech, auditory scene analysis, enhanced summary autocorrelation function ( ESACF), Wavelet packet transform(WPT).**

## 1. INTRODUCTION

Speech signals are rarely available in its pure form for speech processing applications, and are frequently corrupted by acoustic interference like background noise, distortion, simultaneous speech from another speaker etc. In such states, it becomes essential to first separate the desired speech signal from the background. In specific, the task of separating overlapping speech from numerous speakers, called Speech Segregation, is particularly exciting since it includes separating signals having very similar statistic and acoustic characteristics. The ideal approach to segregation should recognize the perceptually significant features of the participating streams, and preserve all those features throughout segregation. Recent methods have attained this objective to an extent [1]. However, they do not totally recreate all portions of the participating speech streams.

In both model-based ( [2]) and feature-based ( [3]) approaches to speech segregation, the input mixture signal is first decomposed into a number of channels using a filter-bank over all time frames, giving a group of Time-Frequency Units (TFUs). Features are mined from each TFU for analysis. In model-based approaches, each TFU is allocated to the speaker whose model has the maximum likelihood of generating its feature. In feature-based approaches, the features of the TFU are analyzed to recognize which of the sources the features match better with, and the TFUs are accordingly allocated to that source. In both cases, each TFU is assigned to one of the two speakers using some predefined criteria. The signals within the TFUs parallel to each source are then used to reconstruct that source. These two approaches adopt that TFUs hold energy from one of the two speakers. This assumption grounds leakage errors and speech missing during reconstruction, because most TFUs characteristically take speech from two speakers.

In this paper, we propose a new method towards segregation, and propose an algorithm that performs better than a contemporary algorithm [4].This paper, focus on and discuss the two-speaker co-channel segregation problem. Proposed system combine the wavelet packet decomposition and least squares fitting approach for single channel speech separation. Multipitch tracking is performed by enhanced summary autocorrelation function (ESACF). ESACF is a computationally well-organized model for multipitch tracking of complex audio signals. This model divides the speech signal into two channels, computes a "generalized" autocorrelation of the low channel signal and of the envelope of the high channel signal, then sums the autocorrelation functions. The summary autocorrelation function (SACF) is again processed to get an enhanced SACF (ESACF). The SACF and ESACF representations are used for finding the pitch.

## 2. SYSTEM DESCRIPTION

Aim of our proposed system is to achieve single channel speech segregation. For that, least squares fitting approach is used. The proposed algorithm performs speech segregation through modeling each time frequency unit as a combination of complex sinusoids which are harmonics of the pitch frequencies of the speakers. Therefore, it needs a multi-pitch detector [5]. Proposed algorithm is different from previous algorithms using sinusoids [6] which are spectrum-based and calculate only the amplitudes of the sinusoids. Such algorithms are vulnerable to the effects of windowing. Our algorithm, instead, directly models the time series and thus is not exaggerated by window parameters. It also estimates both the amplitudes & phases of the sinusoids.

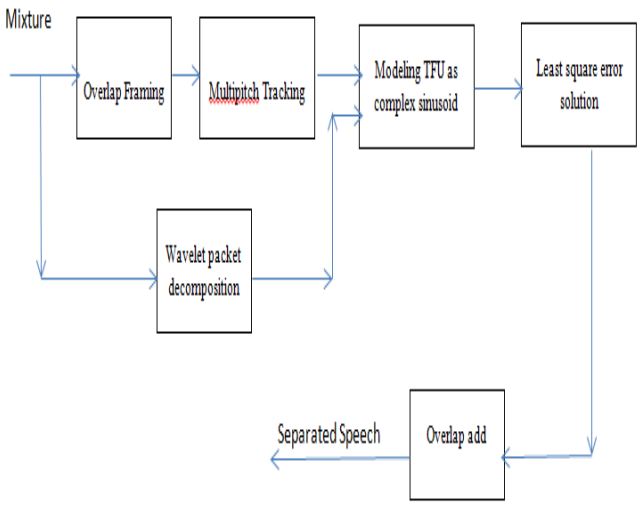


**Fig -1**: Basic Block Diagram

## 2.1 Wavelet Packet Decomposition

Wavelet Packet Decomposition also called Optimal Subband Tree Structuring. The broadband input signal is converted in to narrowband sub band signal by using wavelet packet transform. In wavelet packet transform(WPT), the decomposition of input speech signal is performed by passing the signal through different filters. For n levels of decomposition WPT produces $2^n$ different groups of coefficients. When the decomposition is applied to together the approximation coefficients and the detail coefficients, the process is called wavelet packet decomposition.

## 2.2 Multipitch Tracking

Multipitch tracking is performed by enhanced summary autocorrelation function (ESACF).ESACF is a computationally effective model for multipitch tracking of complex speech signals. This model divides the input speech signal into two channels, below and above 1000 Hz and computes a "generalized" autocorrelation of the low channel signal and of the envelope of the high channel signal, then sums the autocorrelation functions. The summary autocorrelation function (SACF) is again processed to obtain an enhanced SACF (ESACF). The SACF and ESACF representations are used for finding the pitch. The proposed pitch tracking method used in complex audio signal processing applications, such as sound source separation, computational auditory scene analysis(CASA), and structural representation of speech signals.

A block diagram of the proposed two-channel pitch tracking model is shown in figure 2. The system mainly contains four steps: Pre-whitening, Signal separation, Periodicity detection and SACF Enhancer.

The first block is a pre-whitening filter that is used to eliminate short-time correlation of the speech signal. The whitening filter is executed using warped linear prediction (WLP).

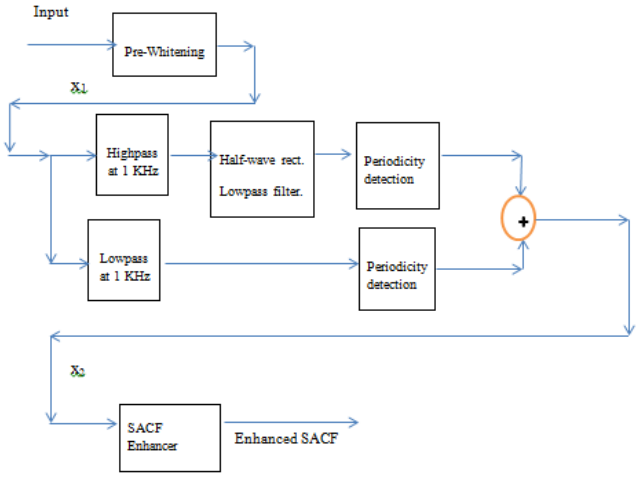


**Fig -2**: Multipitch Tracking Block Diagram

The WLP technique works as ordinary linear prediction and can be used to decrease the filter order significantly. In the middle part of the block diagram the signal is separated into two channels, below and above 1 kHz. The channel separation is carried out with filters with 12dB/octave attenuation in the stop-band. The high-channel signal is half-wave rectified and low-pass filtered with a similar filter that is used for separating the low channel.

The periodicity detection is generally based on generalized autocorrelation i.e. the calculation contains of a discrete Fourier transform (DFT), magnitude compression of the spectral representation, and an inverse transform (IDFT).

$$x_2 = \mathrm{IDFT}\left(|\mathrm{DFT}(X_{low})|^k\right) + \mathrm{IDFT}\left(|\mathrm{DFT}(X_{high})|^k\right)$$

$$= \mathrm{IDFT}\left(|\mathrm{DFT}(X_{low})|^k + |\mathrm{DFT}(X_{high})|^k\right) \quad (1)$$

where $X_{low}$ and $X_{high}$ are the low channel signals and high channel signal before the periodicity detection blocks in Fig. 2. The constraint $k$ determines the frequency domain compression. For normal autocorrelation it is beneficial to use a value lesser than 2. Autocorrelation function makes peaks at all integer multiples of the fundamental period.

## 2.3 Segregation : A Least Squares Problem

The proposed algorithm performs speech segregation through modeling each time frequency unit as a combination of complex sinusoids which are harmonics of the pitch frequencies of the speakers. Therefore, it needs a multi-pitch detector [5]. Proposed algorithm is different from previous algorithms using sinusoids [6] which are spectrum-based and calculate only the amplitudes of the sinusoids. Such

algorithms are vulnerable to the effects of windowing. Our algorithm, instead, directly models the time series and thus is not exaggerated by window parameters. It also estimates *both* the amplitudes & phases of the sinusoids. Fourier series representation of a stationary periodic signal *x[n]* is

$$x[n] = \sum_{i=1}^{N}(\alpha_i^+ exp(j\omega_0 in) + \alpha_i^- exp(-j\omega_0 in)) \quad (2)$$

i = 1, 2, .., N are the N harmonics of the fundamental frequency and $\alpha_i$ is the amplitude of the $i^{th}$ harmonic it is complex + and – indicates $\alpha_i$ for positive and negative frequencies. For a sequence x[n],the unknown amplitudes can be estimated by using M>2N different values of x[n].By giving the values n=1 ,2…….M in equation(2) we get the result as M equations with N unknown coefficients, in vector form

$$\underline{x} = [V^+ V^-]\,\underline{\alpha} = A\,\underline{\alpha} \quad (3)$$

If *M* > 2*N*, this provides an over determined system of equations. The least square error solution of (3) is

$$\underline{\alpha} = A^P \underline{x} \quad (4)$$

$$A^P = (A^H A)^{-1} A^H \quad (5)$$

$A^P$ is the pseudo-inverse of A. Since $\underline{\alpha}$ is complex, *both* the amplitudes *and* phases of the complex exponentials are calculated. The input mixture signal undergo wavelet decomposition that decomposes the input into a number of channels. Analysis is carried on a frame-wise basis with overlapping frames, giving a set of TFUs. For each TFU, if the energy is below a threshold value, the TFU is considered as silent and not processed further. For all non-silent TFUs, the pitch is used to obtain the two streams as described below.

## 2.4 Segregation Of Voiced-Voiced Speech

If the pitch of both speakers is non-zero, the input mixture signal being processed is the sum of two periodic signals, *sA[n]* and *sB[n]*.

$x_{TF}[n]$ is the input mixture signal and the pitch values be $\omega_A$ and $\omega_B$ .

$$x_{TF}[n] = S_{A,TF}[n] + S_{B,TF}[n] = \sum_{i=1}^{N_A}(\alpha_i^+ exp(j\omega_A in) + \alpha_i^- exp(-j\omega_A in)) + \sum_{k=1}^{N_B}(\alpha_i^+ exp(j\omega_B kn) + \alpha_i^- exp(-j\omega_B kn)) \quad (6)$$

Set of parameters $\{\underline{\alpha}\}$ =

$$[\alpha_1^+ \ \alpha_2^+ \ \alpha_3^+ \ \ldots\ldots\ldots \alpha_{N_1}^+ \ \alpha_1^- \ \ \alpha_2^- \ \alpha_3^- \ \ldots\ldots\ldots\ldots \alpha_{N_2}^-]$$

corresponds to the voiced component of speaker A and the parameters

$$\{\underline{\beta}\} = [\beta_1^+ \ \beta_2^+ \ \beta_3^+ \ \ldots\ldots\ldots \beta_{N_1}^+ \ \beta_1^- \ \beta_2^- \ \beta_3^- \ \ldots\ldots\ldots\ldots \beta_{N_2}^-]$$

corresponds to the voiced component of speaker B. *NA* & *NB* are the number of harmonics prevailing between 0 and FS/2, FS is the sampling frequency. By choosing the length of a time frequency unit as *M* > 2(*NA* + *NB*), {$\alpha$} & {$\beta$} are obtained as the least squares solution to the set of equations:

$$\underline{x} = [V_A^+ V_A^- \ V_B^+ V_B^-][\underline{\alpha}^T \underline{\beta}^T]^T \quad (7)$$

reconstruct both the signals *sA,TF[n]* & *sB,TF[n]* that composed the mixture,

$$S_{A,TF}[n]' = V_A^T[n]\alpha' = \sum_{i=1}^{N_A}(\alpha_i'^+ exp(j\omega_A in) + \alpha_i'^- exp(-j\omega_A in)) \quad (8)$$

$$S_{A,TF}[n]' = V_A^T[n]\alpha' = \sum_{k=1}^{N_B}(\alpha_i'^+ exp(j\omega_B kn) + \alpha_i'^- exp(-j\omega_B kn)) \quad (9)$$

## 2.5 Presence Of Unvoiced Speech

If one of the speakers is unvoiced ($\omega_B = 0$) then the mixture signal can be modeled as

$$x_{TF}[n] = S_{A,TF}[n] + S_{B,TF}[n] = \sum_{i=1}^{N_A}(\alpha_i^+ exp(j\omega_A in) + \alpha_i^- exp(-j\omega_A in)) \quad (10)$$

## 3. EXPERIMENTAL RESULTS

The proposed algorithm uses wavelet packet transform for decomposing the input mixture into different sub bands. Numbers of experiments have been led to govern the accuracy of the proposed system by taking the sample of mixtures containing male and female voices. Matlab programs with the existing gamma tone filter bank and the proposed Wavelet packet filter bank in the analysis phase are simulated, similar database are used for existing and proposed models correspondingly. In all the cases of mixtures, the participating speech streams can be separated without loss of information. Separation performance was evaluated with signal-to-noise ratio (SNR). Signal-to-Noise ratio relates the level of a wanted signal to the level of background noise. The higher the value of SNR, the N less disruptive the background noise is. SNR is defined by

equation

$$SNR = 10\log \Sigma x(n)^2 / \Sigma (x(n) - \bar{x}(n))^2 \quad (11)$$

x(n) is the original signal formerly mixing and $\bar{x}(n)$ is the reconstructed speech stream from the mixture

**Table -1:** SNR results for separated and original mixtures

| | SNR | |
|---|---|---|
| | Existing System | Proposed System |
| Mixture | -5.6175 | -5.6175 |
| Speaker A | 3.3390 | 6.5743 |
| Speaker B | 1.3191 | 3.9876 |

From these SNR results we can understand that the proposed system yields better performance than the existing system.

By analyzing the Waveforms ie, Spectrogram and T-F plot of mixture speech, and separated speech shown in figure 3 and 4 respectively, it is clear that the participating speaker streams can be separated without loss of information.

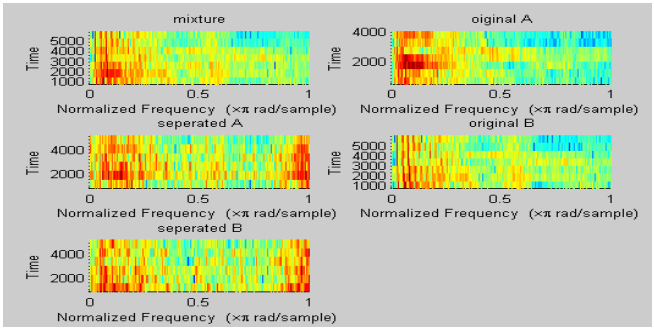


**Fig -3**: spectrogram of signals

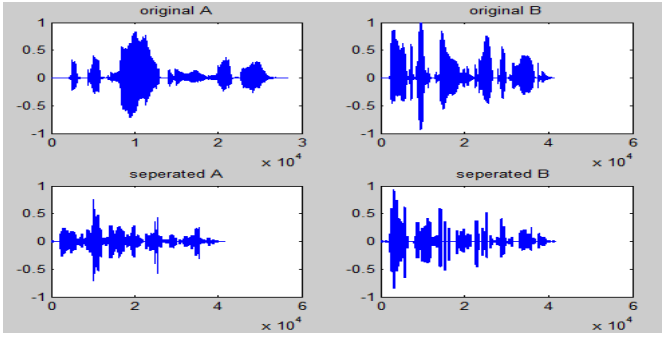


**Fig -4**: T-F plot of signals

## 4. CONCLUSIONS

In this paper, we presented a new method for monaural speech segregation. In this system we used the wavelet packet decomposition in analysis phase and speech segregated through least square solution. Multipitch tracking in each frame is calculated by enhanced summary autocorrelation function (ESACF) algorithm. By analyzing the values of signal to noise ratio, it is clear that the proposed wavelet filter bank is superior to the existing gamma tone filter bank. The efficiency of the system depends on the pitch estimation algorithm.

## REFERENCES

[1] D. L. Wang & G. Brown, eds. (2006), *Computational Auditory* Scene Analysis: Principles, Algorithms and Applications, Wiley/IEEE Press, Hoboken, NJ, 2006.

[2] G. J. Brown & M. P. Cooke, "Computational auditory scene analysis", Comput. Speech and Language, 8, pp. 297-336, 1994.

[3] G. Hu & D. L. Wang, "An auditory scene analysis approach to monaural speech segregation," in *Topics in Acoustic Echo andNoise Control*, edited by E. Hansler & G. Schmidt, Springer, Heidelberg, Germany, pp. 485–515., 2006.

[4] Srikanth Vishnubhotla, Carol Y Espy-Wilson,"An algorithm for speech segregation of co-channel speech" *Institute for Systems Research & Department of Electrical & Computer Engineering,University of Maryland, College Park, MD, USA,* 2009 IEEE

[5] Tero Tolonen, *Student Member, IEEE,* and Matti Karjalainen, *Member, IEEE,"* A Computationally Efficient Multipitch Analysis Model" IEEE transactions on speech and audio processing, vol. 8, no. 6, november 2000

[6] T.F Quatieri & R.G. Danisewicz,, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," IEEE Transactions on Acoustics, Speech and Signal Processing, , vol.38, no.1, pp.56-69, Jan 1990