www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3, Issue 10 October, 2014 Page No. 8978-8982

Co-Channel Speech Separation by Cochlear Filtering And Binary Masking.

Ligin George[1], Lekshmi M.S[2]

[1]Ilahia College of Engineering and Thechnology, MG University,
Muvattupuzha, kerala, India
ligin88@gmail.com

[2]Ilahia College of Engineering and Technology, MG University,
Muvattupuzha, Kerala, India
lekshmims@gmail.com

Abstract: ***Human speech undergoes much interference in a medium. These distortions in the speech signal may leads many disadvantageous in the hearing aid application, speech separation and synthesis. So an effective application can make right turn in field of speech separation. The development of CASA (Computational Auditory Scene Analysis) is trying to reduce these defects and improve the speech applications. The possibility of separating the dominant speech from a mixture and amplifying that may be used in the hearing aid applications. In this paper, we are introducing cochlear design of filters as the channels. Segregation and the grouping are the main methods implemented in this paper. Pitch determination is done based on the response of the cochlea model and combining them using the periodicity detection. Frequency domain analysis done based on the STFT (Short Time Fourier Transform) method. Separation of the dominant speech is done by masking. This method is computationally less complex and we can obtain the better SNR (Signal to Noise Ratio) compare to other related methods available in this literature.***

Keywords: Detection, cochlear filtering, speech separation.

1. Introduction

A speech signal is a representation of sound, also an electrical voltage. Audio signals have frequencies in the audio frequency range of 20 to 20,000 Hz (the limits of human hearing). Signal processing is the mathematical manipulation of an information signal to modify or improve it in some way. Speech processing is used in artificial intelligence and linguistics concerned with the interactions between computers and human (natural) languages. Hence the human speech processing may be more challengeable in case of those distortions. Speech from a single source undergoes continuous acoustic deterioration such as, additive noises from other channels, reverberations from surface reflections etc. CASA [3](Computational Auditory Scene Analysis) is the study of sounds in artificial auditory analysis. It leads with the analysis of sound signals, there computations etc. These algorithms based on CASA may be very beneficial in case of hearing aid applications and automatic speech separation and synthesis. Various algorithms are implemented for monaural speech separation [4] and co channel speech separation. These all methods are based on the analysis of the speech signals as well as the distortion signals. Hence the time level and the frequency analysis is done for the estimation of the ranges. The normal environmental sounds and externals noise [12] leads a main problem in case of the speech separation. Certain algorithms perform the decomposition of the speech signals and separate the acoustic signals. The algorithms providing Hidden Markov Model [8] may lead better quality in the model based speech separation. An onset-offset based speech separation [13] shows a better signal to noise ratio in separated signals.

This paper proposes an unsupervised method of co-channel speech separation. Here STFT (Short Time Fourier Transform) is the signal transformation method used. And along with STFT [5] the pitch determination is also done. The speech signal may be separated by using the masking method and we get the separated dominant signal as well as the interference signal.

WHAT IS CO-CHANNEL SPEECH?

Two speakers are talking simultaneously through a single channel is known as cochannel speech. From the diagram it is shown that the speakers are talking throuugh a single microphone. And after a speech separation process the mixed spech may get separated.

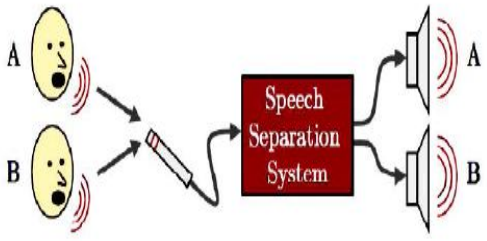

Figure 1: Cochannel speech

Hence comparing the single channel speech separation the cochannel is more complex. It may have the multiple pitch ranges for the speech signals. So the performance of the algorithm is also poor.

2. Related Works

New research on CASA (computational auditory scene analysis) has been started before few years. CASA [3] may provide better solutions for the speech separation. And it avoids most of the interference too. The main advantage of using CASA is to provide the segregation of the speech signals. Blind source separation [6] is also a problem facing in the field of speech synthesis. In blind source separation we take an assumption that the signal is linear and time limited and equalization of the time signal is doing there.

In CASA system, the model based methods like GMM (Gaussian Mixture Model)[7], HMM (Hidden Markov Model) [8] and NMF (Non-negative Matrix Factorization) [10] methods are mostly used. In case of these methods the spectral parameters of the wanted signals are separated and stored, based on these stored information the retrieval of the original signal is done. The effective segregation of the speech signal may occur by the initial analysis and the exact channel decomposition methods. Gamma tone Filter Bank is used as the channel estimation, but it leads high computational complexity.

Hence a new method that is efficient and less complex that compare to the filter bank [6] is proposed. This model is designed using by the cochlear modelling of the filters. Cochlea is a critical band in the human ear that predicts the frequencies in different spots. The paper is organized as follows: section 3 describes the cochlear filtering of channels. In section 4 the complete speech segregation system is explained. Ideal Binary Masking [11] is used for the separation of the dominant signal. Section 5 describes the quality of the separated signal using some quantitative measure and perspective measure.

3. Cochlear Filter Design

Cochlea is a part of human ear that may predict every range of frequencies. Critical bands are in cochlea that may responds to all ranges of frequency. The critical bands are divided into 18 channels. The decomposition of channels is given in figure 1. The whole channel may decomposed by using filters of having 20 ms of frame length. The decomposition of (0-8000) Hz is done through this method. The proposed method is implemented in MATLAB.

After the decomposition we get 18 separated channels as output. The results of these 18 channels are combined using the periodicity detection method. Thus we are getting combination of decomposed signal that may successfully produce a response. The Fig.3 shows the block diagram representation of the system. Here the speech mixture is processed for the pitch determination process and the (STFT) Short Time Fourier Transform that convert to frequency domain for better evaluation. And the resulted signal is segmented to and the speech separation is done and (ISTFT) Inverse Short Time Fourier Transform is done to convert back to time domain to get the speech signal.

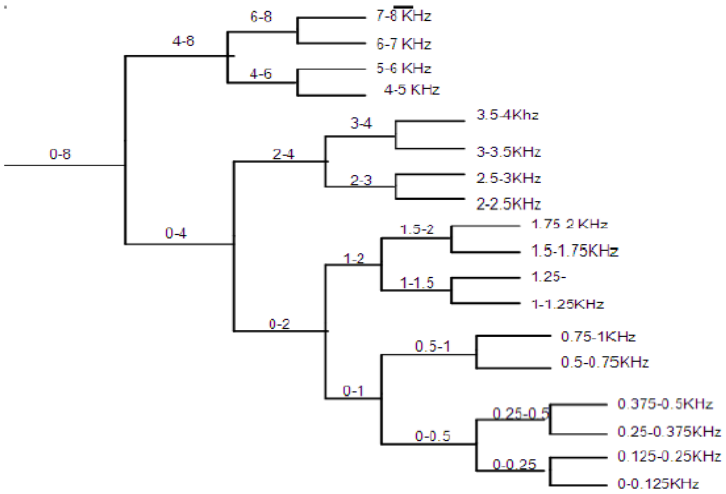

Figure 2: Cochlear model

4. System Description

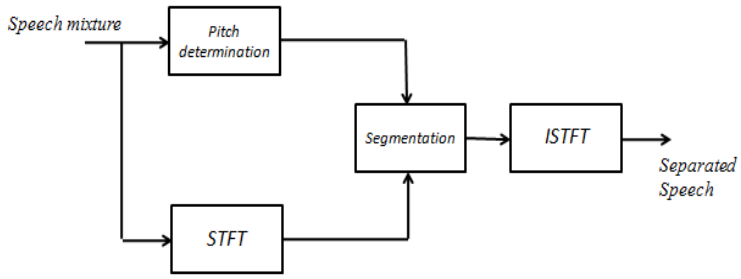

Figure 3: System Block Diagram

4.1 Pitch Determination

The pitch determination is the main process in the proposed system and is shown in Figure 3. Here Pre-whitening is the is the starting process and the channel decomposition is done based on the cochlear modelling. All the decomposed channels are combined using the periodicity detection method. The enhanced summary autocorrelation method is used to get the exact pitch ranges.

4.1.1 Pre-Whitening

The first block of pitch determination indicates the pre-whitening method. Here this method is mainly used to reduce the short term correlation of the signal and to provide a constant bit rate. The (WLP) Warped Linear Prediction [9] technique is used, and it is a whitening filter. The WLP work as ordinary linear prediction filter along that it implement critical band auditory resolution of spectral modelling instead of uniform frequency resolution. A 12-th order filter along with hamming windowing is used. It may have the same functionality of hair cell activity normalization technique.

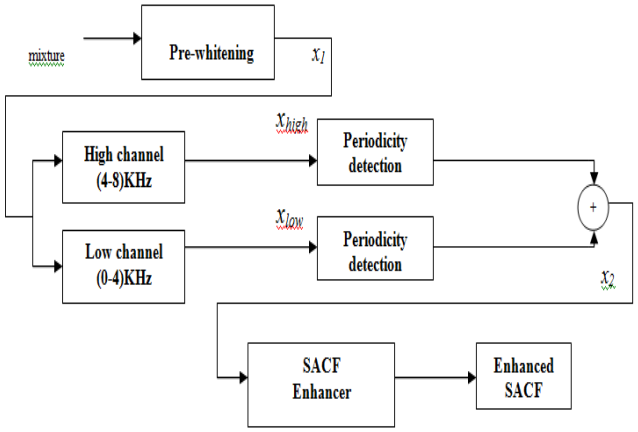

Figure 4: Pitch Detection Block

4.1.2 Cochlear Channel Modelling

Cochlea is a part of inner ear that predicts every range of frequencies from 20 Hz – 20 KHz. The response is given by the critical bands of the cochlea. The detail diagram of cochlear decomposition is shown in Fig.4. After the decomposition of the channels according to limit the frequencies, 18 channels are been resulted in the final stage. The groupings of these channels are done in the output. Here the low channel is separated from 0Hz to 4 KHz and high channel is from 4 KHz to 8 KHz. And sequent the channel may be divided according to the figure and 12 dB octave filters is used for channel separation. This method is proposed because it may have all the spectral components of signal. And it shows better result than other channel composition.

4.1.3 Periodicity Detection

The Periodicity Detection is based on generalized autocorrelation function. And these may done by Fourier Transform (FFT) Here the magnitude are been compressed of the spectral representation of the signals and then the Inverse Fourier Transform (IFFT) is done. This is shown in equation (1). Here *xlow* and *xhigh* are the two channels shown in the equation. Correspondingly we have to get the periodicity of the 18 channels as compared to cochlear model of filters as shown in Fig.4. Here k is normally taken as the value k=2. Hence for better output the value of k may vary approximately as k-1.67 ranges.

$$x_2 = \text{IDFT}\left(|\text{DFT}(x_{low})|^k\right) + \text{IDFT}\left(|\text{DFT}(x_{high})|^k\right)$$
$$= \text{IDFT}\left(|\text{DFT}(x_{low})|^k + |\text{DFT}(x_{high})|^k\right) \quad (1)$$

4.1.4 Enhancing SACF

The result of the periodicity detection is known as (SACF) Summary Auto Correlation Function. Enhancing the SACF is to find the exact peak pitch values of the speakers. These may have the steps as follows. The original signal that we get by SACF is clipped to positive values and then expanded in time by a factor of 2 and subtracted from original clipped SACF [1] function. And again the result is clipped to have the positive values. This may remove the repetitive peaks and double the time lags where the basic peak is higher than the duplicate. Hence these may be done by the time factor of 2,3,4 etc in order to remove the repetitive peaks. Hence plotting the signals we get the number of peaks based on the number of speakers. This resulting function is called Enhanced Summary Auto Correlation Function (ESACF) [2].

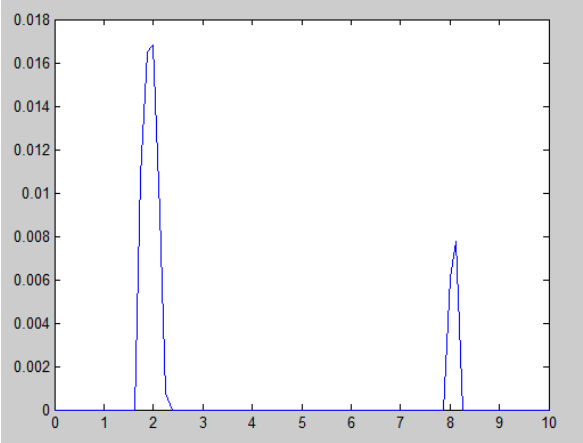

Figure 5: Time-Amplitude plot of ESACF

The above figure 5 shows the result of the ESACF. The input given is two speech signals and the figure shows the pitch of these speakers. The highest peak shows the dominant speech and the lowest peak shows the interference speech.

4.2 Short Time Fourier Transform (STFT)

Fourier transform methods are used to convert the time domain to frequency domain analysis. Here the function is to be transformed is multiplied using a window function having a period. Hence the resulting two dimensional signal is been analysed. Mostly a Hanning window, Hamming window are used. Here each window has a time period of 20 ms and an overlap of 10 ms.

4.3 Speech Segmentation

By using a pitch determination block, we get all the pitch range in each frame. Here select the most high frequency as the frequency of the dominant speaker. Normally the male speaker has the pitch of 80 Hz to 160 Hz. And the female speaker has the pitch of 180 Hz to 260 Hz. After predicting the pitch value, the speech separation are done based on these pitch value. The process is like that; we are taking -10 to +10 values of the pitch in each of the frames and separate these values from the total mixture. The Ideal Binary Masking [11] is the method used in the separation of the signals.

$$\text{Binary Mask} = \begin{cases} 0 \text{ if } f = k \times P1 + \left(\frac{\rho f_s}{N}\right) \\ 1 \qquad \text{otherwise} \end{cases} \quad (2)$$

The equation (2) represents the binary mask. Here the k represents the harmonic frequencies. Here the k varies from 1 to 3, represents the width of the mask. Here ρ represents the range of values, for same gender taking the value of -10 to +10 and for different gender it may from -15 to +15. Here this may

change approximately to get the better result.

4.4 Inverse Short Time Fourier Transform (ISTFT)

Taking the inverse of STFT is done in this method. The reconstruction of the signal is done here. Then it may be stored and reproduced. Here the time domain analysis is done using the function.

5. Results

5.1 Signal To Noise Ratio (SNR)

Table 1: Signal to Noise ratio of various mixtures and there separated mixtures.

Mixture of male & female speakers			Mixture of female speakers		
SNR	MIXTURE 1	MIXTURE 2	SNR	MIXTURE 1	MIXTURE 2
SNR ORIGINAL	-40.6080	-42.8802	SNR ORIGINAL	-41.6341	-42.8890
SNR SEPARATED	22.7733	29.8858	SNR SEPARATED	10.6661	21.8553
SNR INTERFERENCE	5.0608	12.7704	SNR INTERFERENCE	-0.7581	7.1802
Mixture of male speakers			External disturbances		
SNR	MIXTURE 1	MIXTURE 2	NOISES	SNR MIXTURE	SNR SEPARATED
SNR ORIGINAL	-38.9149	-38.6634	AIR	-9.0628	17.1675
SNR SEPARATED	-8.9401	8.3858	CAR	-10.5935	4.5774
SNR INTERFERENCE	-30.7574	-32.2329	INDUSTRY	-24.9314	-6.8109

5.2 Time Complexity Analysis

Table 2 : Time for processing

Cochlear Model Used	Time required
Mahmoodzadeh model	2.080698 sec
Proposed Model	1.97034 sec

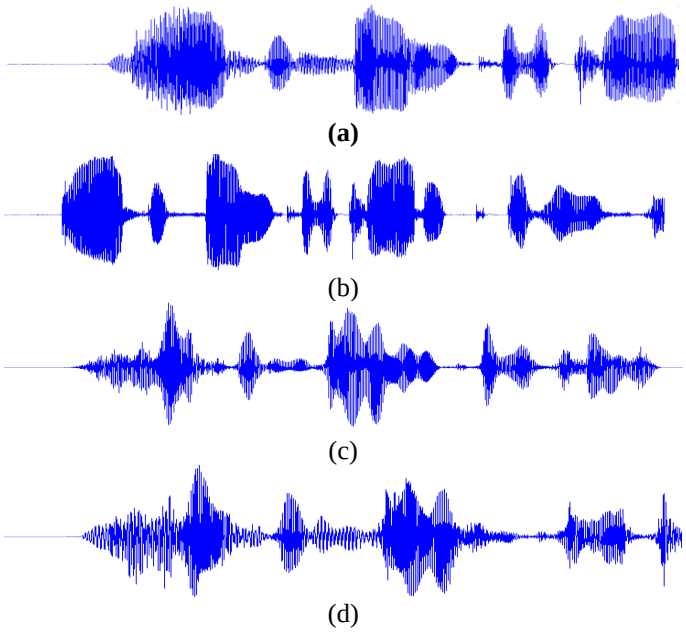

Figure 6: Time amplitude plot of signals. (a)Mixture (b) Original (c) Segregated (two channels) (d) Proposed (Cochlear channel model)

6. Conclusion

From the results it is shown that all types of channels provides intelligible speech response. The proposed cochlear model shows that it might be better comparing with the existing channel models. And the processing time may also less. The proposed system may give better results in less time and it may be less complex. So the proposed system can be easily execute comparing all existing systems.

References

[1] Lekshmi M.S, Dr.Satidevi P.S "Unsupervised speech segregation using pitch information and time frequency masking" Department of Electronics, M.G University.

[2] Tolonen T Karjalainen, "A computationally efficient multi pitch analysis model," IEEE Transactions on speech and audio processing, vol-8, No.6,November 2000.,

[3] Wang D L, Brown G J."Computational auditory scene analysis: principles, algorithms and applications," New Jersey: Wiley-IEEE Press, 2006

[4] Guoning Hu and DeLiang Wang , "Monaural Speech Segregation based on Pitch Tracking and Amplitude Modulation," IEEE Transactions on neural networks vol. 15, no. 5, September 2004.

[5] Shiju Aravindakshan, P.S. Sathidevi And R. Rajavel ., "Separation Of Dominant Speech From Simultaneous Talkers," Proceedings Of IASTED Int. Conference On Signal And Image Processing (SIP 2007) At Honolulu,USA.

[6] Van der Kouwe,A.J.W., Wang ,D.L brown,G J,‖A comparison of auditory and blind separation techniques for speech segregation‖, technical report :Osu-cisrc-6/99-trl5,department of Computer and Information Science, the ohio state university,Columbus,Ohio43210-1277,1999.

[7] D. A. Reynolds, ―Speaker identification and verification using Gaussian mixture speaker models,‖ Speech Commun., vol. 17, pp. 91–108, 1995.
[8] Z. Jin and D. L.Wang, ―HMM-based multipitch tracking for noisy and reverberant speech,‖ IEEE Trans. Audio, Speech, Lang. Process., vol.19, no. 5, pp. 1091–1102, Jul. 2011.
[9] Lanie, Karjalainen, Altosaar., "Wrapped linear prediction in speech and audio processing," ICASSP
[10] M. N. Schmidt and M. Mørup, ―Nonnegative matrix factor 2-D deconvolution for blind single channel source separation,‖ in ICA, 2005.
[11] DeLiang Wang ., "On Ideal Binary Mask As the Computational Goal of Auditory Scene - Speech Separation by Humans and Machines," Chapter 12, pp. 181-197, Kluwer Academic, Norwell MA, 2005
[12] Hartmut Traunmller and Anders Eriksson. , "The frequency range of the voice fundamental in the speech of male and female adults," Department of Linguistics, University of Stockholm 1994
[13] Guoning Hu a and DeLiang Wang,‖ Auditory Segmentation Based on Onset and Offset Analysis‖ pub/tech-report/2005.