Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 122 – 126

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Unsupervised Speech Segregation Using Pitch Information and Time Frequency Masking

Lekshmi M S[a,*],Sathidevi P S[b]

[a-b]*Department of ECE, NIT Calicut, Kerala-67360,India*

Abstract

Speech undergoes various acoustic interferences in natural environment, while many of the applications require an effective way to separate the dominant signal from the interference. In this paper, a Short-time Fourier Transform (STFT) based unsupervised method for single channel speech separation is proposed. It uses the pitch information of the dominant and interfering speakers and then generating a time frequency mask based on the pitch frequencies. Through rigorous objective and subjective evaluations, it is shown that the proposed system is capable of providing better Signal to Noise Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) compared to other related methods available in the literature.

*Keywords:*CASA; pitch; IBM.

1. Introduction

Two major problems being faced by hearing impaired persons are difficulty in understanding speech when contaminated with other speech signals and difficulty in understanding fast speech. Hence, separation of dominant speech from a mixture and its amplification will be very helpful for such persons.

Computational Auditory Scene Analysis (CASA) is an emerging field of signal processing aimed at developing computational system to simulate human auditory system. One of the main goals of CASA is speech

[*]Corresponding author. Tel.: 91-949-636-9684.
*E-mail address:*lekshmims@gmail.com

doi:10.1016/j.procs.2015.02.002

segregation. There are two approaches for speech segregation - unsupervised and model based methods. In model based method the system applies the learned knowledge of the speaker, but in the former method the system only receives the mixture signal as the input. Such systems extract the features from the mixture and these features are used as cues for segregating the speech.

In this paper, separation of dominant speech by using an unsupervised method, which is well suited for hearing aid applications, is proposed. The most important cues used in this work are the pitch frequencies of dominant and interfering speakers. Here, a computationally efficient method for the pitch estimation of the interfering speakers and separation of dominant speech from a speech mixture using the pitch information is proposed. This method exhibits superior performance in terms of signal to noise ratio when compared with the other systems available in the literature.

2. System Overview

The input speech mixture is first decomposed into its time frequency representation using STFT. Decomposed signal is then applied to the pitch determination block which determines the pitch of dominant and interfering speakers. It also identifies the gender of the speakers using the estimated pitch range [7]. After identifying the pitch of the interfering speaker, a binary mask is created and it is used for the segregation of speech (Time frequency domain). Then it is re synthesized using Inverse STFT.

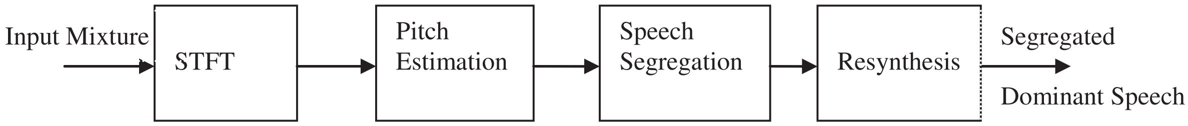

Fig 1: Basic block diagram of the proposed system

2.1 PitchEstimation

For the pitch estimation, an autocorrelation method [2] is adopted here. The input signal is separated into two channels, below and above 1 kHz. For performing channel separation we have implemented filters with 12dB per octave attenuation in the stop band. The periodicity detection is based on “generalized autocorrelation,” i.e., the computation consists of a discrete Fourier transform (DFT), magnitude compression of the spectral representation, and an inverse transform (IDFT). The signal x_2 corresponds to the summary autocorrelation function (SACF) and is obtained as

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

The value of k should be 2 for obtaining autocorrelation, but experimentally k=1.67 gives better peak values representing pitch.

The autocorrelation output from each channel is summed to get the SACF. The peaks in the SACF curve produced at the output of the model are good indicators of potential pitch periods in the signal. SACF is further enhanced by clipping the SACF to its positive values and it is up sampled by a factor of two, the up sampled signal is subtracted from the original clipped one and the resulting signal is again clipped to its positive values. Time lag corresponding to the peak value of the enhanced SACF (ESACF) gives the pitch of the dominant speaker.

Using the above pitch analysis method, the pitch of each frame P(f) is identified , where ‘f’ represent the frame number. From among these pitch frequencies most frequently occurring value is considered as the dominant pitch (P_d).

For identifying the pitch of the interfering speaker, the pitch values are sorted according to their frequency of occurrences in frames. The dominant pitch P_d is compared with subsequent frequently occurring pitch values by computing the difference between the two. The frequently occurring pitch value with difference more than 10 is considered as the pitch of the interfering speaker (P_I).

After determining the pitch of dominant and interfering speakers, the gender of the speakers are identified : if the pitch of the speaker is in between 80 and 160 then it is considered as a male speaker and if the pitch is in between 160 and 255 then it is considered as a female speaker.

2.2 Speech segregation and re-synthesis

For segmenting the mixture signal, a binary mask is generated to eliminate the unwanted TF units. Basic idea is to eliminate the interfering pitch frequency, its nearby frequencies and its harmonics.

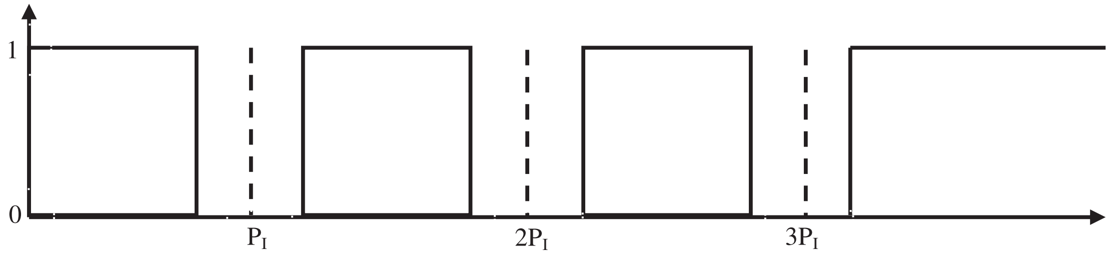

Fig 2: Schematic representation of binary mask of each frame.

$$Binary\ Mask = \begin{cases} 0 & iff = k * P_I + \frac{\rho f_s}{N} \\ 1 & Otherwise \end{cases} \qquad (2)$$

Binary mask is created in such a way as to eliminate frequencies in the range of interfering pitch frequencies and harmonics. Equation (2) represents the binary mask, where k represents the order of the harmonics (here k varies from 1 to 3), ρ represents the width of the mask (if the speakers are of same gender ρ value is from -10 to 10 otherwise it is from -15 to 15).

The binary mask of each frame is then multiplied with a cosine window given by

$$coswin(i) = \frac{1+\cos(2\pi(i-1)/wlen-\pi)}{2} \qquad (3)$$

Mask of the entire TF unit can be expressed as

$$mask(j,i) = BM(j,i) * coswin(i) \qquad (4)$$

Speech segregation is done by multiplying x(j,i) with mask(j,i), where x(j,i) is the STFT of mixture speech

$$y(j,i) = x(j,i) * mask(j,i) \qquad (5)$$

Re-synthesis of the segregated signal is performed by Inverse STFT. In the proposed system 1024 point STFT with a hamming window is implemented.

3. Results And Discussion

We have computed SNR and PESQ to evaluate the performance of the proposed system and compared with those of a closely related method [1]. In that method authors used modulation frequency representation for pitch determination and soft mask method for speech segregation. For evaluating the proposed method, we have taken recorded speech samples of male and female speakers having sampling frequency 8 KHz and they are mixed linearly by keeping one of them as dominant. The system identified the gender of the speaker with an accuracy of 93%. Power spectral density plots of the clean, segregated signal using method in [1]and the segregated signal using proposed method are provided in figure 3 to demonstrate the performance. Proposed method is implemented in Matlab 7.1.

3.1 SNR

We have arbitrarily taken 5 speech samples from male-male mixture, male-female mixture and female-

female mixture for testing the system and the performance is shown in table 1. SNR is computed using equation (6) where x(n) is clean signal and $\hat{x}(n)$is the separated signal.

$$SNR = 10 * Log_{10}\left\{\frac{x^2(n)}{[x(n)-\hat{x}(n)]^2}\right\} \quad \text{dB} \qquad (6)$$

Table 1: SNR of segregated dominant speech

	SNRof mixture	SNR of segregated speech using Ref[1] (dB)	SNR of segregated speech using Proposed system(dB)
Mixture of male speaker with male speaker	-6.56	-0.377	3.36
Mixture of female speaker with female speaker	-7.61	-6.06	2.55
Mixture of male speaker with female speaker	-6.79	-2.64	2.96

Table 2: PESQ of segregated dominant speech

	PESQ of mixture	PESQ of segregated speech using Ref [1]	PESQ of segregated speech using Proposed system
Mixture of male speaker with male speaker	1.93	2.17	2.27
Mixture of female speaker with female speaker	2.25	2.28	2.30
Mixture of male speaker with female speaker	1.84	2.05	2.27

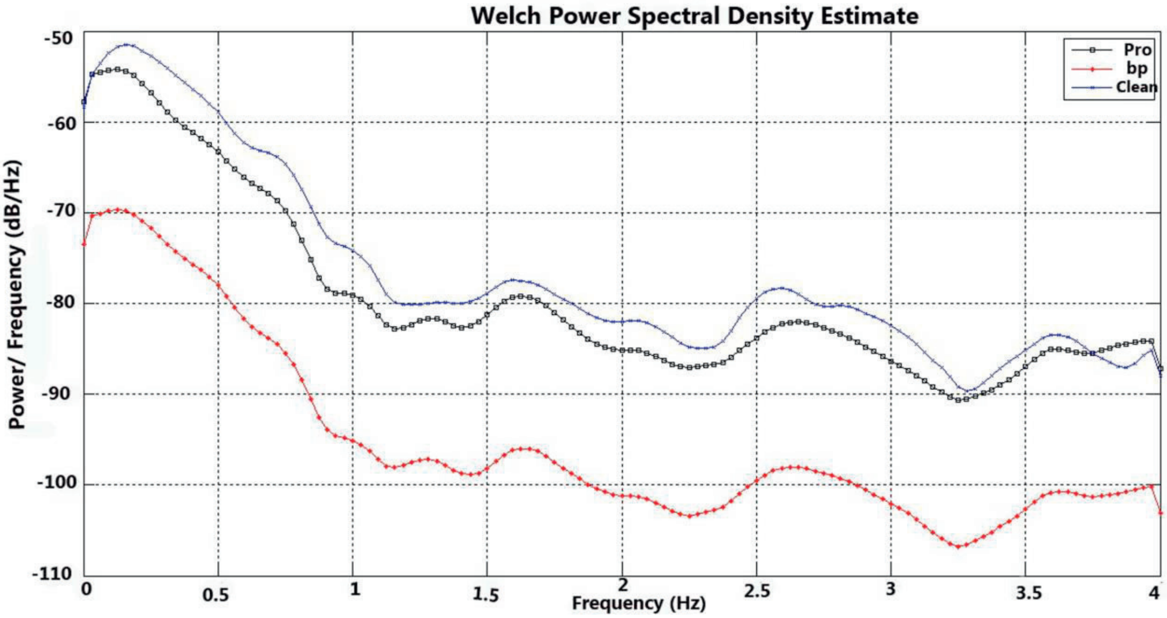

Fig 3: Power spectral density plots of clean speech (blue), separated speech using [1] (red) and separated speech using proposed system (black)

3.2 PESQ

The Perceptual Evaluation of Speech Quality (PESQ) is an international standard for estimating the Mean

Opinion Score (MOS) from both the clean speech signal and its degraded speech signal. PESQ was officially standardized by the International Telecommunication Union. It gives a score ranging from 0 to 5.

4. Conclusion

In this paper, an unsupervised speech segregation method for the separation of dominant speech from a speech mixture is proposed. Here, pitch frequencies of the dominant and interfering speakers are first determined and then binary masks are created by using this pitch information. The experimental results show that the proposed method yields a better performance compared to the related work [1] in terms of SNR and PESQ.

References

1. 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, *EURASIP Journal on Advances in Signal Processing*, 2012.
2. Tolonen T Karjalainen, A computationally efficient multi pitch analysis model, *IEEE Transactions on speech and audio processing*, November 2000.
3. Hu, Y. and Loizou, P. ,Evaluation of objective measures for speech enhancement, Proceedings of INTERSPEECH-2006, Philadelphia, PA,
4. DeLiang Wang , Guy J. Brown ,CASA BOOK principles ,algorithms and Applications, IEEE press, 2006
5. Guoning Hu and DeLiang Wang Monaural Speech Segregation based on Pitch Tracking and Amplitude Modulation*, IEEE Transactions on neural networks*, September 2004.
6. DeLiangWang , On Ideal Binary Mask As the Computational Goal of Auditory Scene - Speech Separation by Humans and Machines, p. 181-197, Kluwer Academic, Norwell MA, 2005
7. HartmutTraunmüller 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.