IPASJ International Journal of Electronics & Communication (IIJEC)
Web Site: http://www.ipasj.org/IIJEC/IIJEC.htm
A Publisher for Research Motivation ..... Email:editoriijec@ipasj.org
Volume 5, Issue 11, November 2017 ISSN 2321-5984

Speech Enhancement Using Hilbert Spectrum and Wavelet Packet Based Soft-Thresholding

Eastwood, Clint

University of California, Davis

ABSTRACT

A method of and a system for speech enhancement consists of Hilbert spectrum and wavelet packet analysis is studied. We implement ISA to separate speech and interfering signals from single mixture and wavelet packet based softthresholding algorithm to enhance the quality of target speech. The mixed signal is projected onto time-frequency (TF) space using empirical mode decomposition (EMD) based Hilbert spectrum (HS). Then a finite set of independent basis vectors are derived from the TF space by applying principal component analysis (PCA) and independent component analysis (ICA) sequentially. The vectors are clustered using hierarchical clustering to represent the independent subspaces corresponding to the component sources in the mixture. However, the speech quality of the separation algorithm is not enough and contains some residual noises. Therefore, in the next stage, the target speech is enhanced using wavelet packet decomposition (WPD) method where the speech activity is monitored by updating noise or unwanted signals statistics. The mode mixing issue of traditional EMD is addressed and resolved using ensemble EMD. The proposed algorithm is also tested using short-time Fourier transform (STFT) based spectrogram method. The simulation results show a noticeable performance in the field of audio source separation and speech enhancement.

1.INTRODUCTION

The problem of separating different sound sources can be classified as a denoising or enhancement problem, where the "signal" is the important part of the audio stream, and the "noise" is everything else. Although this is simple task for human auditory system, the automated audio source separation for speech enhancement can be considered as one of the most challenging topics in current research. Audio source separation or speech enhancement has many applications including robust automatic speech recognition, music transcription, surveillance applications, remixing of studio recording etc. Speech quality may significantly deteriorate in the presence of interfering noise signals. The modern communications systems, such as cellular phones, employ some speech enhancement procedure at the preprocessing stage, prior to further processing [1]. One approach to separate the mixed audio signals is microphone array processing [2]. The array processing requires hug computation and inefficient to be used in real world applications. Hence, present research trend is to reduce the number of microphones used in recording of the intended acoustical environment. Several noise reduction schemes have been developed which try to suppress the signal components corresponding to noise and enhance the target component. This technique corresponds to the use of only one microphone. For instance, in the application of spectral noise suppression schemes [3, 4, 5] to speech enhancement it is assumed that the signal of interest is the speech with its typical speech pauses while the noise signal is regarded as stationary and uninterrupted. Therefore, it is possible to estimate the noise spectrum during speech pauses and subsequently subtract it from the spectrum of the noise contaminated speech segments in order to obtain the.

enhanced speech signal. Computational Auditory Scene Analysis (CASA) is one of the first methods that tried to decrypt the human auditory system in order to perform an automatic audio source separation system [6]. A recent advancement of single mixture audio separation is the independent subspace analysis (ISA) method [7, 8]. The study [8] describes a single stage source separation using EMD and ICA. The method proposed KLD based clustering algorithm to group the independent basis vectors and experimental results show a good source separation performance. The implementation of ISA is the extension of basic ICA by decomposing an audio mixture into independent source subspaces. Westner [7] implemented ISA method to decompose a mixture spectrogram into independent source subspaces and inverting them to yield source separation. They employed short-time Fourier transformation (STFT) to produce the time-frequency (TF) representation (spectrogram) of the mixed signal and derived a set of frequency-independent basis vectors to represent the source subspaces. The STFT employed in TF representation includes a certain amount of cross-spectral energy during the overlap of the window between the successive time frames. Two major limitations of STFT degrade the disjoint orthogonality of the audio sources in TF domain. Another study by Ghanbariet all in [9] presents a speech enhancement algorithm based on adaptive thresholding wavelet packet

decomposition. A new VAD is designed in wavelet domain and the method shows high performance in speech enhancement. The proposed method of speech enhancement consists of two stages processing. In the first stage a source separation algorithm is implemented using Hilbert spectrum where ISA is used to separate audio sources from a single mixture and in the second stage, a adaptive denoising algorithm is implemented based on wavelet packet decomposition which further enhance the target speech quality. The Hilbert spectrum (HS) is employed for the TF representation of the audio signals. The HS does not include noticeable amount of cross-spectral energy terms. It is able to represent the instantaneous spectral information of any time series without windowing. The mode mixing issue of traditional EMD [10] is addressed and resolved using ensemble EMD and Hilbert transformation are employed together to derive HS. The decomposition of the mixed signal in spectral domain is obtained by the spectral projection between the mixture and each IMF component. This vector space is used to derive a set of independent basis vectors by applying PCA and ICA. The hierarchical clustering algorithm is used to group the basis vectors into the given number of component sources. A further removal of background noise is obtained in the second stage by wavelet packet decomposition (WPD)[9]. The enhancement process consists of a simple voice activity detection (VAD) followed by noise estimation on the basis of calculated subband SNR. From these values an adaptive soft thresholding parameter is derived to update the wavelet coefficients. This technique works well in cascade with the source separation method. The waveforms, ISNR as well as OSSR values show that this new multistage technique highly improves the performance of speech enhancement results. Regarding the organization of this paper, the basics of EMD and HS are described in section 2, the separation and enhancement algorithm are described in section 3 and 4, the experimental results are presented in section 5, and finally Section 6 contains conclusion of the study.

2. EMD BASED HILBERT SPECTRUM (HS)

The principle of the EMD technique is to decompose any signal x(t) into a set intrinsic mode functions Cm(t) (IMFs). Each IMF satisfies two basic conditions: in the whole data set the number of extrema and the number of zero crossing must be same or differ at most by one, (ii) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero.

where M is the number of IMFs and rM(t) is the final residue. Due to the presence of noise, traditional EMD is survived by mode mixing problem. Mode mixing problem is defined as an IMF that includes oscillations of dramatically disparate scales or a component of similar scale residing in different IMFs [11]. This issue is resolved by introducing ensemble EMD algorithm.

3. SUBSPACE DECOMPOSITION OF HILBERT SPECTRUM

The single mixture blind source separation BSS technique decomposes the TF space of the mixture as the sum of independent source subspaces. The ensemble empirical mode decomposition (EMD) and Hilbert transformation are employed together to derive HS. The Hilbert Spectrum of the mixture signal is constructed by properly arranging the frequency responses of the individual IMF along time and frequency axes with preferred number of frequency bins.

3.1. Algorithm for Source Separation

The block diagram of the overall enhancement technique is shown in figure 4. In this work, only the mixture of two audio sources are taken into account and hence k=2. One source is the speech signal and the other corresponds to the interfering signal, which can be any noise source.

3.2. Clustering of Independent Basis Vectors

Once the spectral independent basis vectors are obtained, the basis vectors are then grouped into the number of sources. The proposed hierarchical clustering algorithm for finding the clusters around the mixing vectors is tested. We follow bottom-up (agglomerative) strategy that the starting point is the single samples, considering them as clusters that contain only one object. Clusters are then combined, so that the number of clusters decreases while the average number of objects per cluster increases.

4. SPEECH ENHANCEMENT USING WAVELET PACKET DECOMPOSITION (WPD)

The wavelet packet transform (WPT) is a generalization of the decomposition process that offers a better performance compared to the ordinary wavelet methods [9]. In the wavelet analysis, a signal is split into an approximation and a detail. The approximation is then itself split into a second-level approximation and detail and the process is repeated.

5. EXPERIMENTAL RESULTS

The efficiency of the proposed enhancement technique is tested in two steps, 1) separate the signals from the mixture of two audio sources and 2) enhance the quality of the target source. Both mix1 (speech with telephone ring sound) and mix2 (speech with flute sound) mixtures are used in the experiments. The two recorded signals with normalized amplitudes are added to make 0dB SNRs. The speech signal of each mixture comprises utterances of several words spoken by the same speaker. The audio signals are sampled at 16 kHz sampling rate and 16-bit amplitude resolution. The mixed signal is divided into blocks of 0.25s by using Tukey window with 50% overlapping. The average value of the running short-term relative energy between the original and separated signals is used to measure the separation efficiency is termed as the original-to-separated-signal ratio (OSSR) and defined mathematically as.

The OSSR values show the similarity between two signals. If the two signals are exactly equal, the OSSR value will be 0, that is a smaller deviation of OSSR from 0 indicates a higher degree of separation. Henceforth, the male speech is denoted as Signal1 and instrument signals as Signal2. Table 1 shows the average OSSR of each signal for both mixtures. The separation efficiency is compared between the Hilbert-based method and the Fourier-based one; it shows that the performance is higher for the Hilbert-based method.

To study the quantitative analysis of enhancement performance in two stages, we have employed an improvement of signal to noise ratio (ISNR) measure. The ISNR [ISNR (dB)=SNRin-SNRout] represents the degree of enhancement of the target signal when it is degraded by interfering noise. Here SNRinand SNRout represent the input and output SNRs, respectively. Table 2 shows the ISNR of target signal for three mixtures using Fourier and Hilbert based methods in two stages. The higher value of ISNR indicates better quality signal and it is observed that Hilbert based method is better than Fourier based method (in stage 1) and a further improvement is obtained after WPD based soft thresholding method (in stage 2). It is observed that the separated speech from stage-1 contains some residual noise which is successfully reduced in stage-2.

6. CONCLUSION

We have presented a two stage method for speech enhancement. It describes the effectiveness of Hilbert spectrum (HS) in time-frequency (TF) representation of audio signals. The efficiency of HS has been compared with short-time Fourier transform (STFT) as a method of TF representation with the consideration of disjoint orthogonality of audio signals and the experimental results show that Hilbert spectrum performs better than STFT based spectrogram for TF representation. Usually, the higher resolution in TF representation demonstrates the signal inmore detail and hence improves the separation performance. A set of independent basis vectors are derived from HS by applying principal component analysis (PCA) and ICA in sequence. In both algorithms, hierarchal clustering is employed to group the independent bases to derive the source subspaces. The experiments show that the use of Hilbert spectrum in time-frequency representation fits better in subspace decomposition than Fourier-based method does and it obviously increases the separation efficiency. The distortion due to residual noise in the target speech is handled at the second stage using wavelet packet based softthresholding method which made considerable enhancement. The approach uses a wavelet signal processing strategy and controls the threshold values based on estimated subband SNRs to remove noise components that exists after the separation algorithm. The simulation results show a noticeable performance in the field of audio source separation and speech enhancement. A further improvement is expected by implementing the whole process in wavelet packet domain instead of EMD. The improvement of the robustness of the separation process will be addressed in the future work.

REFERENCES

[1] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, “Blind Source Separation Combining Independent Component Analysis and Beamforming.” EURASIP Journal on Applied Signal Processing, vol. 11, pp. 1135-1146, 2003.

[2] J. M. Valin, J. Rouat, and F. Michaud, “Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter,” Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2004.

[3] Y. Ephraim, and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Trans. on Acoustic, Speech and Signals Processing, vol. 32, pp. 1109-1121, 1984.

[4] O. Cappe, “Estimation of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” IEEE Trans. on Acoustic, Speech and Signals Processing, vol. 2, pp. 345-349, 1994.

[5] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. on Acoustic, Speech and Signals Processing, vol. 27, pp. 113-120, 1979.
[6] G. J. Brown, and M. Cooke,"Computational auditory scene analysis," Computer Speech Language, vol. 8(4), pp. 297-336, 1994.
[7] M. A.Casey, and A. Westner, "Separation of mixed audio sources by independent subspace analysis," Proc. of International Computer Music Conference, pp. 154-161, 2000.
[8] M. K. I. Molla, and K. Hirose, "Single mixture audio source separation by subspace decomposition of Hilbert spectrum," IEEE transactions on audio, speech and language processing, vol. 15(3), pp. 893-900, 2007.
[9] Y. Ghanbari, and M. R. K. Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets", Speech Communications, Elsevier, vol. 48, pp. 927-940, 2006.
[10] N. E. Huang, Z.Shen, S. R Long, et al. "The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis," Proc. Roy. Soc. London A, vol. 454, pp. 903-995, 1998.