www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3, Issue 10 October, 2014 Page No. 8983-8987

WPST Based Single Channel Speech Separation

Abhilash Surendran[1], Lekshmi M S[2]

[1]M Tech, Applied Electronics,
Ilahia College of Engineering and Technology, MG University
Kochi, Kerala, India
abhilashsurendran@hotmail.com

[2]Asst. Professor, Electronics and Communication Engineering,
Ilahia College of Engineering and Technology, MG University
Kochi, Kerala, India
lekshmims@gmail.com

Abstract: ***The process of Single channel speech separation is done to efficiently separate the required speech signals from a mixture. In this research paper we used WPST (Wavelet Packet Based Sub-band Transform) to offer a multi resolution property of wavelet transform to increase the efficiency by reducing the number of coefficients required in each sub-band vector to replace the previously used SPWT (Sub-band perceptually weighted transformation).The new approach improves the separation quality and it results lowest error bound in terms of objective measurements such as Perceptual Evaluation of Speech Quality (PESQ)),segmented SNR in comparison with SPWT based features.***

Keywords: Single channel speech separation (SCSS), Wavelet Packet Based Sub-band Transform (WPST), Vector quantization (VQ), Objective measurements.

1. Introduction

The single channel speech separation (SCSS) one of challenging scenarios in the case of audio processing and telecommunication field. The separation of audio signal is essential in the case of automatic speech recognition process (ASR). There are many kinds of noises that may interfere the speech signal such as speech babble, background noise, colored noise and white noise, among which the competing speech is the main interference to the speech signal of interest and the removal of such noise is most challenging process due to the high correlation between the temporal structures of target speech and the speakers which mask it. Which results poor separation quality.

In the common life of a human the interference due to the competing speech is a normal one and the humans have the ability to recognize and separate out the required speech. But while doing the same task by a machine, like computer it becomes very difficult task. Recently many researches are going on for the separation of two speech signals received from one communication channel which is called as single channel speech separation (SCSS).

There are many kind of applications for this single channel speech separation and it can be used as a pre-processing stage for certain systems such as Speech coding, hearing aids and automatic speaker recognition systems. It improves the robustness of all the above processes because it effectively separates out the required signals

The single channel speech separation processes are generally of two main classifications that are source driven and model driven. In the source driven methods the required speech signals of interest are extracted from the mixed signal without a prior knowledge about the underlying speakers. The computational auditory scene analysis (CASA) is the most known source driven approaches used now a days. It performs speech separation by extracting psychoacoustic cues from the given mixed signals. It has dis-advantages such as it is highly affected by the poor accuracy of the multi-pitch tracking algorithm estimates of the speech signals obtained from the mixed signal. And the outputs have poor perceptual quality due to crosstalk problem

The next type of SCSS technique is model driven in which the speech separation is carried out with the help of pre-defined speaker models in the form of codebooks. In this paper the model driven technique is followed which contains mainly three processes such as 1) Feature selection, 2) Modelling, and 3). Estimation. Accordingly based on the feature selection, Model and estimation many types of model-based single channel speech separation techniques have been proposed. The model based separation scenarios completely depends on prior knowledge about the underlying speakers called speaker models. The techniques such as Vector quantization (VQ), Gaussian Mixture model (GMM), Hidden Markov model (HMM) used to get the restrictive constraints for preparing speaker models.

In this paper the new transform called WPST (Wavelet Packet Based Sub-band Transform) is detailed that was utilized for VQ-based SCSS which have larger separation quality and good separation performance even at low SSR's. It have multi resolution property and can be associated with the perceptual model of human ear

2. Related Works

In the single channel speech separation (SCSS) process as discussed before many types of transformations are used in order to obtain the feature vectors of underlying speakers in a mixed form of speech, in previous works the authors followed to use the code vectors obtained from the Short term Fourier Transform (STFT). And vector quantization designs to cope up with poor signal quality. Later the Sub-band perceptually weighted transformation (SPWT) [1] is used to improve the separation efficiency.

In order to obtain the new feature parameters, the each STFT magnitude spectrum code vectors are normalized to its maximum value then the logarithm of normalized vectors have been taken. This reduced the dynamic range of code vectors hence improved the quality. And a signal distortion (SD) measurement is taken as a performance index. Since the SPWT depends on STFT parameters it assumes the signal to be stationary for a fixed frame period and this provides only a fixed time frequency resolution. And the SPWT does not considers the critical bands of human ear because in order to consider the critical bands it requires high frequency resolution, resulting low time resolution. Which will result long time for code book preparation.

Hence it requires a new transform which supports multi resolution, which matches perceptual model of human ear. As a solution we can make use of wavelets to solve this problem.

3. Wavelet Packet Based Speech Separation

The wavelet transform has an excellent property that it can be used for analyzing a signal in different time-frequency resolutions. If we take the packet transform it enables the signal to be analyzed in various sub-bands available, out of which the proper sub-bands can be chosen for our particular needs. This analysis based on sub-bands is now commonly used for audio signal processing and generating perceptual model of audio signals.

The human ear analyses an audio signals in various sub-bands called the critical bands. The critical bands have various sub-bands within the frequency of audibility limit (20 Hz to 20 KHz). And this variable time frequency analysis requires less number of coefficients compared to previous SPWT features.

3.1 Wavelet packet based sub-band transformation (WPST)

Considering the critical bands of human ear and the advantages of using the wavelet packets in sub-band analysis of audio signals, a Wavelet packet tree structure is proposed in order that it closely matches those critical bands, here the high frequency components are analyzed using a narrow window and the low frequency components are analyzed using a wider window. Hence the transform vectors from the proposed WPT tree structure can give a good performance in the case of model based SCSS. . This new vector obtained from wavelet transformation can be called as Wavelet packet based sub-band transformation (WPST). Since out of various wavelets available the Daubechies Wavelet Type 4 (Db4) is used here since its sub-bands closely matches the critical bands. And the Db4 wavelets are commonly used for audio signal analysis.

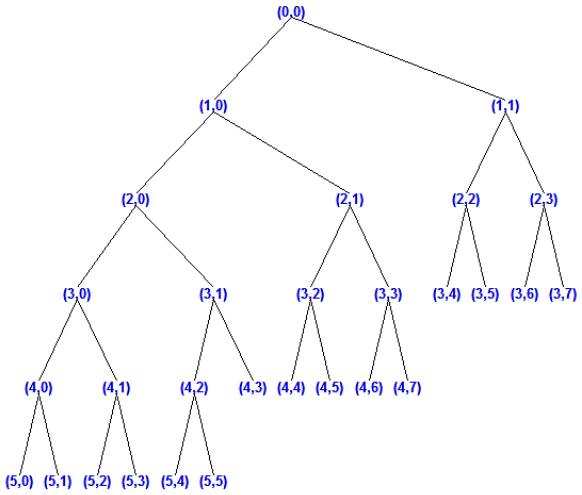

Fig 3.1: Proposed WPT tree structure

In order to obtain the new feature parameters, first the Db4 wavelet packet tree is taken such that its sub-bands closely matches the critical bands of human ear. Then each sub-band coefficients are concatenated to obtain a single vector (S_j). Then these sub-band coefficients are normalized by dividing the energy in each sub-band. This results a parameter ranging [0, 1]. This normalized format results better classification accuracy.

3.1 WPST Based separation scenario

In this technique the wavelet packet transform is used to get the sub-bands which are very closer to the critical bands of human ear. In order to obtain the new feature vector 15 sub-band vectors are chosen from the wavelet tree that are (5,0),(5,1)…..(5,5), (4,0),(4,1)…..(4,5), (3,0),(3,1)…..(3,5). And all these vectors are concatenated to obtain the single vector S_j.

The main idea behind is to create a code-book for the required speaker to be separated from the monaural speech signal Similar to the initial stage of separation in SPWT based work [1] here also the initial stage is code book creation. In order to create the code books the 10 sentences for each speaker from the TIMIT database were taken. Since the signals have a frequency 16 KHz, it is down sampled to 8 KHz. The code book contains WPST vectors obtained from the wavelet tree structure. Before creating the code books the vectors are normalized. However the usage of wavelet transform avoids the problems due to non-stationarity of the signal. So here a frame of 128ms is used to prepare WPST vectors this enables us to use the characteristics of speech extending beyond 32 ms .The frame size of 128 ms gives a vector of size 1024. The proposed WPT tree structure is applied to these frames and hence WPST vectors are obtained by concatenating the sub-bands of tree and normalizing it.

In-order to prepare the code-book LBG algorithm used in [1] SPWT is used. The codebook size is chosen to be as 1024. The performance of codebooks are evaluated by applying the test speech to the codebook of specific speaker. First the WPST is taken for the test speech and after that vector quantization is applied using the codebook of the speaker.

For the separation process speech mixture is required, the mixed speech is formed by applying an acoustic transfer function to the individual speech signals and adding them in time domain ($S_1(n)$,$S_2(n)$...). In the separation or extraction

process the $S_1(n)$ is separated by comparing the WPST vectors of S1 with the WPST vectors of mixed signals. And taking closely matching vectors from code book of speaker 1. To find the close matches the Euclidian distance function is used.

Instead of using 256 coefficients (32 ms window), selecting wider window (for eg.128 ms) is followed here. For the sub-bands below 750 HZ the band width chosen is 100 Hz, for sub-bands from 750 Hz to 2000 Hz a bandwidth of 200 Hz is used, and so on 15 sub-bands are used for the analysis. For each 128 ms duration we get 128 coefficients much smaller compared to 1024 length used in previous SPWT based method.

4. Results

The measurement of speech quality is done for obtaining the evaluation of performance of WPST based system. For that an objective measurement is carried out. It is generally calculated between the original speech signal and the distorted speech signal using some mathematical analysis. It is very simple process that it does not require any human listeners and less time consuming. The commonly used objective measurements are segmented SNR, Weighted spectral slope distance measures (WSS), Perceptual Evaluation of Speech Quality (PESQ). Where the SNR is calculated in short frames and averaged hence it is called as segmented SNR

$$seg\ SNR = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Lm}^{Lm+L-1} x^2(n)}{\sum_{n=Lm}^{Lm+L-1}\{x(n)-\breve{x}(n)\}^2} \quad (1)$$

L corresponds frame length, and M represents the number of signal frames (N = ML). The normal frame length is chosen to be 15-20 ms The WSS is a direct spectral distance measure it is based on the comparison between the smooth spectra and the distorted spectra of speech signals. It is obtained by filter bank analysis. And is defined as given below.

$$d_{wss} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)(S_c(j,m)-S_d(j,m))^2}{\sum_{j=1}^{K} W(j,m)} \quad (2)$$

Here the K is number of bands, M is the total frames, and Sc (j, m) and S_d (j, m) are spectral slopes of the j^{th} band in the m^{th} frame for respectively clean and distorted speech. The PESQ measure is the international standard for calculating the Mean Opinion Score (MOS). It is an officially standardized method used by ITU (International Telecommunication Union) it gives a score ranging from -0.5 to 4.5.

The another objective measurement is Overall Quality it is obtained by linearly combining the PESQ, LLR, WSS measures

$$c_{ovl} = 1.594 + 0.805\ PESQ - 0.512\ LLR - 0.0007\ WSS \quad (3)$$

Where the LLR is the log-likelihood ratio obtained from LPC analysis of original and the separated speech.

	Separated using SPWT	Separated using WPST
2 male speakers	3.1327	3.8559
3 male speakers	2.7969	4.0891
1 Male and 1 female	3.3507	4.4494
1 male and 2 female speakers	2.2354	3.8488

Table 4.1: Overall quality for separation of male speaker

	Separated using SPWT	Separated using WPST
2 female speakers	3.3574	4.2756
3 female speakers	2.9921	4.2694
1 female and 1 male	3.0366	4.2566
1 female and 2 male speakers	2.6318	3.6649

Table 4.2: overall quality for separation of female speaker

	Separated using SPWT	Separated using WPST
2 male speakers	5.317	23.572
3 male speakers	4.0358	21.217
1 male and 1 female	6.251	24.463
1 male and 2 female speakers	1.9341	14.3203

Table 4.3: SNR_{Seg} measured for separation of male speaker

	Separated using SPWT	Separated using WPST
2 female speakers	4.495	24.3239
3 female speakers	2.765228	17.2673
1 female and 1 male	3.044010	18.41857
1 female and 2 male speakers	1.5521	15.2011

Table 4.4: SNR_{Seg} measured for separation of female speaker

	Separated using SPWT	Separated using WPST
2 male speakers	2.325	2.844
3 male speakers	2.145	3.103
1 male and 1 female	2.514	3.549
1 male and 2 female speakers	1.706	2.905

Table 4.5: PESQ measured for separation of male speaker

	Separated using SPWT	Separated using WPST
2 female speakers	2.729	3.453
3 female speakers	2.450	3.442
1 female and 1 male	2.477	3.416
1 female and 2 male speakers	2.171	2.764

Table 4.6: PESQ measured for separation of female speaker

From the results observed from various measurements given above in the table clearly we can identify that the WPST based method is more advantageous than the previously used SPWT based method. The proposed system improves the overall performance and PESQ values. There is a large improvement in the case of segmented SNR. The power spectral density plots of various combinations of speakers in the mixture is given below. From the PSD plot also we can clearly identify the improvement of quality while using the new feature parameter.

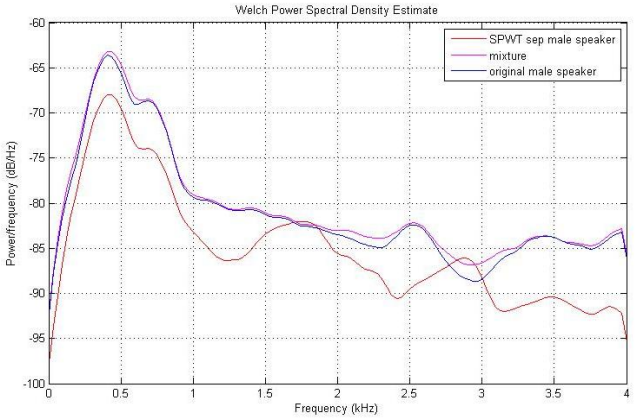

Fig 4.1: Separation of male speaker from a mix with one female speaker speakers (SPWT method)

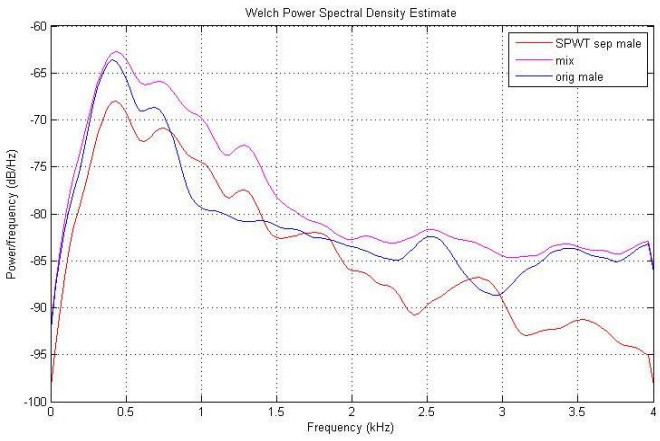

Fig 4.2: Separation of male speaker from a mix with two female speakers (SPWT method)

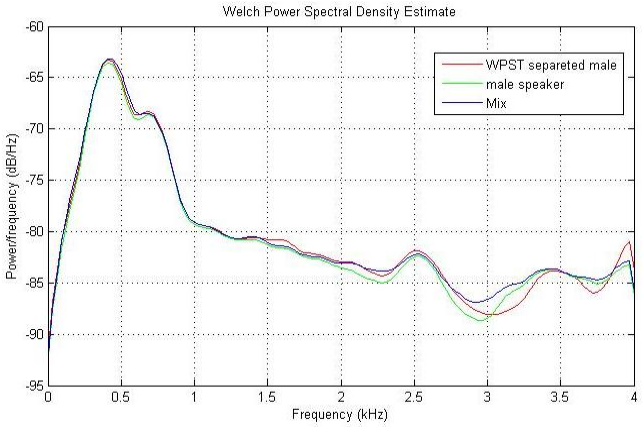

Fig 4.3: Separation of male speaker from a mix with one female speaker speakers (WBST method)

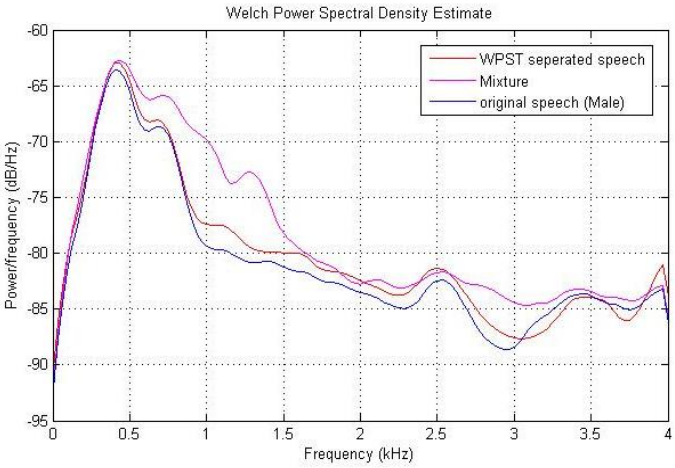

Fig 4.4: Separation of male speaker from a mix with two female speakers (WBST method)

5. Conclusion

The separation of speech signal from a single microphone recorded mixture is a challenging scenario in this paper a VQ based approach for single channel speech separation is presented. In the previous works a SPWT (based on STFT features) transformation is used to obtain the feature vector for separation. Due to the limited resolution of this features, it is found difficult to approximate this with the perceptual model of human ear, so we gone for a new transformation supporting multi resolution analysis called Wavelet packet based sub-band transformation (WPST). And this new feature vectors improved the quality of SCSS system.

References

[1] Mowlaee, P., Sayadiyan, A., Evaluating single-channel speech separation performance in transform-domain., Journal of Zhejiang University-SCIENCE C (Computers & Electronics) ISSN 1869-1951 (Print); ISSN 1869-196X (Online), 2010.

[2] Mowlaee, P., Sayadiyan, A., Performance Evaluation for Transform Domain Model-Based Single-Channel Speech

Separation., 7th ACS/IEEE Int. Conf. on Computer Systems and Applications, 2009, pp.935-942.
[3] Jensen, J., Heusdens, R., Jensen, S.H., A Perceptual Subspace Method for Sinusoidal Speech and Audio Modelling., IEEE Int. Conf. on Acoustics, Speech, and Signal Processing,2003, pp.401-404.
[4] Kondoz, A.M., Evans, B.G., Hybrid Transform Coder for Low Bit Rate Speech Coding., Proc. European Conf. on Speech Technology, 1987, pp.105-108.
[5] R. Sarikaya and H. L. Hansen, High resolution speech feature parameterization for monophone-based stressed speech recognition. IEEE Signal Process. Lett, vol. 7, no. 7, July 2000pp. 182-185.
[6] Mangesh S. Deshpande and Raghunath S. Holambe, Speaker Identification Using Admissible Wavelet Packet Based Decomposition., International Journal of Information and Communication Engineering 6:1, 2010.
[7] Mowlaee, P., Sayadiyan, A., Model Based Monaural Sound Separation by Split-VQ of Sinusoidal Parameters., 16th European Signal Processing Conf., p.1-5, 2008. 55
[8] Zavarehei, E., Vaseghi, S., Qin, Y., Noisy speech enhancement using harmonic noise model and code book based post processing., IEEE Trans. Audio Speech Lang. Process., 15(4):1194-1203, 2007.
[9] Reddy, A.M., Raj, B., Soft mask methods for single channel speaker separation. IEEE Trans. Audio Speech Lang. Process., 15(6):1766-1776, 2007.
[10] T. Virtanen, ―Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,‖ IEEE Trans. Audio, Speech, and Language Process., vol. 15, no. 3, pp. 1066–1074, 2007.
[11] P. Li, Y. Guan, S. Wang, B. Xu, and W. Liu, ―Monaural speechseparation based on MAXVQ and CASA for robust speech recognition, ―Elsevier Computer Speech and Language, vol. 24, no. 1, pp. 30–44, Jan.2010
[12] P. Divenyi, Ed., ―Speech Separation by Humans and Machines‖, 1sted., New York: Springer, 2004.
[13] D. Wang and G. Brown, Computational auditory scene analysis: Principles, algorithms and applications, Wiley-IEEE Press, New York, 2006.
[14] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Performance Evaluation of Three Features for Model-Based Single Channel Speech Separation Problem," in Proc. of International Conference on Spoken Language Processing (Inter speech--ICSLP), Pittsburgh, Pennsylvania, USA, pp. 2610-2613, Sept. 17-21 , 2006.
[15] ITU-T P.862, Perceptual evaluation of speech quality(PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation P.862, 2000. 56
[16] M. Cooke, J. R. Hershey, and S. J. Rennie, ―Monaural speech separation and recognition challenge,‖ Elsevier Computer Speech and Language,vol. 24, no. 1, pp. 1–15, 2010.
[17] Gersho, A., Gray, R.M., Vector Quantization and Signal Compression. Kluwer Academic Publishers, Boston, USA, p.345-372,1992.
[18] S. Quackenbush, T. Barnwell, and Clements, Objective measures of speech quality, Englewood Cliffs, NJ: Prentice-Hall,1988.
[19] L. Thorpe and W. Yang, ―Performance of current perceptual objective speech quality measures,‖ in Proc. IEEE Speech Coding Workshop, pp. 144–146, 1999.

·