Single Channel Speech Separation in Transform Domain Combined with DWT

Bismi Muhammed[1] Lekshmi M S[2]
[1]PG Student [2]Assistant Professor
[1,2]Department of Electronics & Communication Engineering
[1,2]Ilahia College Of Engineering And Technology, Ernakulam

Abstract— In this paper wavelet transform and vector quantization are combined for single channel speech separation. In this paper a new feature parameter called Wavelet Packet Transform (WPT) is used for speech separation. The objective behind proposed WPT is to improve the separation quality by reducing the number of coefficients and to reduce the complexity. The performance of WPT is compared with quantization performance of Sub band perceptually weighted transformation (SPWT) method for single channel speech separation. As compared with SPWT, the new feature WPT has better separation performance in terms of objective measurements.

Key words: Linde-Buzo-Grey (LBG)algorithm, Single channel speech separation(SCSS), Sub-band Perceptually Weighted Transformation(SPWT),Vector quantization(VQ),Wavelet packet transform(WPT)

I. INTRODUCTION

In many speech applications the desired speech of interest is always corrupted by different type of noise. Cock -tail party problem is one such example where several speakers are talking at the same time. Separating the desired speech from these types of mixture is called single channel speech separation (SCSS) [1] where the interfering signal is also a speech signal. Single channel speech separation is a challenging research topic for years to arrive at a more accurate method. They have many applications that they are used as a preprocessing stage in hearing aids, speech coding methods where separation of speech signal is crucial for their working.

Single channel speech separation methods fall mainly into two categories: Source driven methods and model driven methods. In source driven methods the separation of a speech signal from a mixture is carried out without having knowledge about the speakers in the mixture. A common example for source driven method is computational auditory scene analysis (CASA) [2] in which the separation is carried out using perceptual acoustic cues from the speech signal. Even though CASA methods are fast and have a better separation quality they work well with signals having a uniform pitch track otherwise cross-talk problem will arise.

A method called speech fragment decoding [3] is used for speech separation which combines both the source driven and model driven algorithms. It employs a searching then grouping algorithms using sound fragments. But this method has high computational complexity and poor performance al low SNR.

In model driven SCSS methods the separation is carried out having a previous knowledge about underlying speakers. It makes use of a speaker statistical model called codebooks which stood as a base for such methods. There are two factors that have to be considered while choosing a model driven approach: (1) Statistical model created for underlying speakers and (2) the feature type used for separation. Different types of transformations are used in previous works on single channel speech separation methods in order to obtain the feature type. Many such methods use short –time Fourier transform vectors as a first step on feature type creation. Poor separation quality of STFT vectors are resolved using vector quantization approaches. Then a transformation called Sub –Band perceptually weighted transformation (SPWT) [4] is applied on these STFT feature vectors to obtain better quality.

In SPWT the new feature type is obtained by utilizing the logarithm on the normalized magnitude STFT feature vectors. This process reduces the dynamic range of feature vectors and also improves the classification accuracy. Here the performance index is a distortion measure. It was observed that SPWT provide higher synthesis and improvement on perceptual evaluation of speech quality (PESQ) compared to STFT features .However SPWT bears some drawbacks: (1) use of STFT features on SPWT results in fixed time resolution,(2) Critical bands of human auditory system is not taken into consideration,(3)Not suitable for sub-band based analysis,(4) Use of STFT vectors leads to uniformly weighted frequency bins. These drawbacks degrade the quality of separation performance .so in this paper a new separation method that utilizing wavelet packet transform is proposed. In the proposed method Vector Quantization (VQ) [1] algorithms are employed for codebook creation.

II. SEPARATION BASED ON WAVELET TRANSFORM

The wavelet transform is a signal processing tool that manages to represent the stationary and transient behavior of a signal with lesser number of coefficients. Wavelets are an ideal tool for analysis of non-stationary signals because of their property of being irregular in shape and they are compactly supported. As compared with Fourier transform wavelet transform offers better temporal resolution of the high frequency components and better frequency resolution of low frequency components. These properties of wavelet packet transform allow the signal to be analyzed in sub-bands of our needs.

A. Wavelet Packet Transform (WPT)

The important properties of wavelet packet transform on resolution enables to consider the critical bands of human auditory system while analysis. Critical bands are the band over which human ears can analyses sound signals. To enable these advantages of wavelet transform on separation process a wavelet packet tree structure is proposed here and it enables sub band analysis since it resembles critical bands. The new feature type that is obtained from wavelet transform is called as wavelet packet transforms (WPT). Because of the resemblance of Daubechies wavelet type 4(Db4) sub bands with critical bands; Db4 is used in the proposed system.

The new feature parameter is obtained by applying Db4 wavelet to the input signal .So it's sub bands resembles to that of critical bands of human auditory system. Then the coefficients from the sub-bands of our interest are combined into a common vector named as S_j.Vector S_J is then normalized to its maximum value thus acquire a good classification accuracy. Then take the logarithm of the normalized value to reduce the dynamic range. These steps are illustrated below

$$\widetilde{S}_{j=\log(1+S_{jnormal})} \qquad (1)$$

Where $\tilde{S_j}$ is the feature parameter,$S_{jnormal}$ is the normalized vector normalized to its maximum value.

B. WPT Based Code Book Creation

The main criteria for a model based technique are a codebook or speakers statistical model for each of the underlying speeches. Linde –Buzo-Grey (LBG) algorithm in VQ training [5] that is used in SPWT is applied for this. The main aim behind codebook creation is to generate representative vectors corresponding to the speech signal. It is the initial stage of separation process which is similar to that of initial stage of SPWT [4] separation scenario. For code book creation 10 sentences for each speaker from TIMIT database is used and is down sampled to 8 KHz from 16 KHz. In this separation scenario each of the speech is framed at 128 ms which gives 1024 as vector size. Each of these speech frames are combined to a single vector and then wavelet packet transform is applied on this vector. The wavelet packet transform that we are used in this scenario is 5 level WPT with Db4 wavelets. Proposed WPT structure is shown in the figure 1.

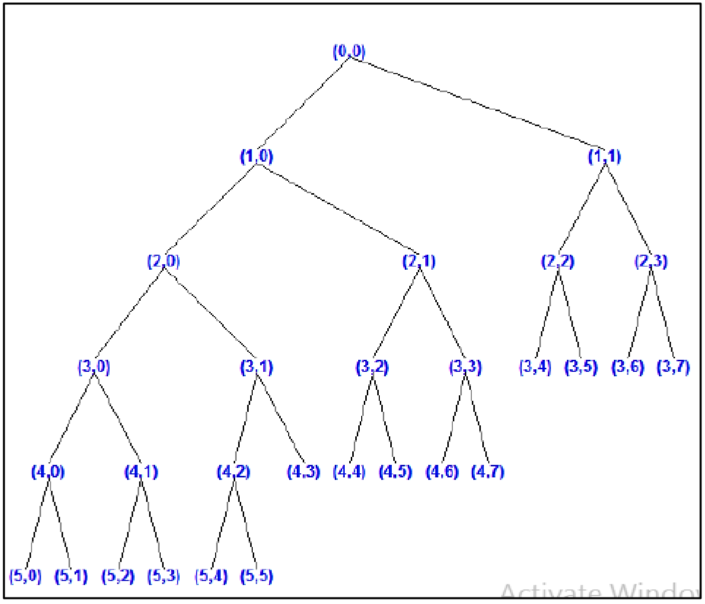

Fig. 1: proposed WPT structure

15 sub bands from the corresponding wavelet packet tree is taken for creating feature vectors. Those sub bands are (5,0),(5,1)…(5,5), (4,0),(4,1),…..(4,5) , (3,0),(3,1)….(3,5).These sub band vectors are combined into a single vector which is then normalized and take its logarithm before codebook creation. LBG algorithm on these combined vector resulted in a code book with WPT vectors obtained using wavelet packet tree. The code book size is taken as 1024. Performance evaluation of created code books is done using a test speech.

C. Separation Scenario Using WPT

The separation scenario used for analysis is shown in the figure 2. An important requirement is the code book for each of the speakers. Here M represents the codebook size which is chosen to be 1024.A mixture of speech signal is required for separation process which is obtained by mixing two speaker signals, S1 (n) and S 2(n) at certain attenuation. Euclidean distance is taken as a segregation criteria used for finding perfect match of speech signals in the mixture with the vectors in the code book. Perfect match $(\hat{S}_{i,1}(f), \hat{S}_{j,2}(f))$ is obtained by comparing WPT vectors of the mixture signal with WPT vectors of the underlying speech signals. After comparison the centroid od the optimal index for each codebook is replaced to obtain perfect match.

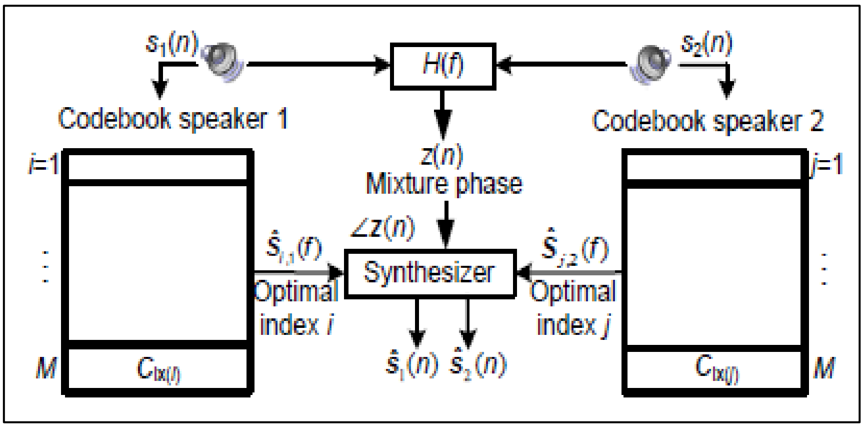

Fig. 2: Block diagram of separation scenario using WPT

During decoding process utilizing the mixture phase along with the indices of $\hat{S}_{i,1}(f), \hat{S}_{j,2}(f))$ gives $\hat{S}_1(n)$ and $\hat{S}_2(n)$ the reconstructed signal. As compared with SPWT which uses a narrow window(32 ms), a wider window (128 ms) is used with WPT which reduces the number of coefficients to 128 which is smaller than 1024 used in SPWT.

III. Simulation Results

The quality of the separated output is evaluated using objective measurements such as weighted spectral slope distance measures (WSS), Perceptual Evaluation of Speech Quality (PESQ), Segmented SNR. The corresponding evaluations are compared to those obtained using SPWT based method .The results observed from various measurements are shown in the table. Separation of speaker 1 from a two speaker mix and 3 speaker mix is shown. The power spectral density plots of separation of speaker 1 from different mixtures is also shown.

	Separated using SPWT	Separated using WPT
2 speaker mix	0.8493	0.9721
3 speaker mix	1.1002	1.7762
4 speaker mix	0.7032	0.8302

Table 1: PESQ measured for separation of speaker-1

	Separated using SPWT	*Separated using WPT*
2 speaker mix	*-1.0123*	*2.3042*
3 speaker mix	*1.2058*	*6.4537*
4 speaker mix	*1.1034*	*3.1397*

Table 2: SNR seg measured for separation of speaker1

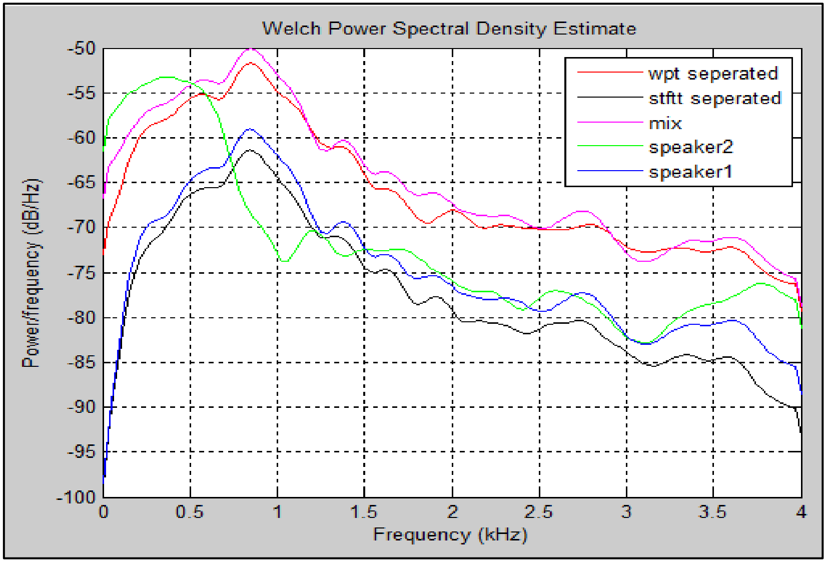

Fig. 3: separation of speaker 1 from 2 speaker mix

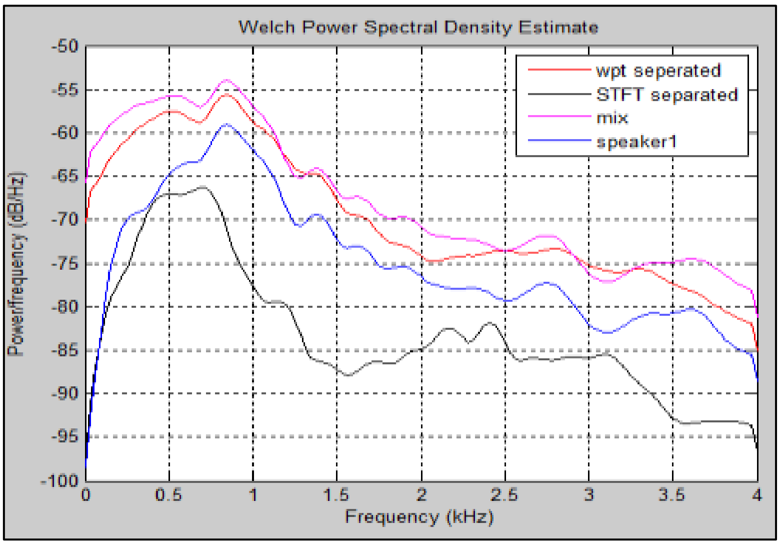

Fig. 4: separation of speaker 1 from 3 speaker mix

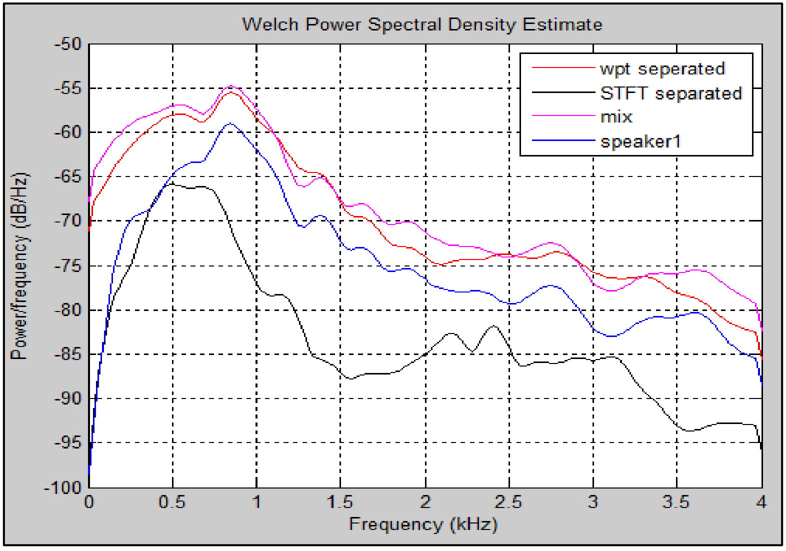

Fig. 5: separation of speaker 1 from 4 speaker mix

From the table of results obtained from various measurements it is clear that WPT based method is more superior to SPWT based method .It can be seen that the PESQ, SNRseg values have been improved as compared with SPWT based method. The power spectral density plot itself shows the improvement of WPT method over SPWT.

IV. Conclusion

In this paper a Wavelet packet transform is proposed to improve the separation quality of single channel speech separation. The objective of this method is to reduce the number of coefficients and improves separation speed. From the results of objective evaluations it was observed that WPT achieve more separation performance and higher synthesis as compared to previous methods. This new feature improves the quality of SCSS method and it also considers the critical bands of human auditory scene analysis.

References

[1] Ellis, D.P.W., Weiss, R.J., 2006. Model-Based Monaural Source Separation Using a Vector-Quantized Phase- Vocoder Representation. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.957-960. [doi:10.1109/ICASSP.2006.1661436]

[2] Hu, G., Wang, D., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans.Neur. Networks, **15**(5):1135-1150. [doi:10.1109/TNN.2004.832812]

[3] Barker, J., Cooke, M., Ellis, D., 2005. Decoding speech in the presence of other sources. Speech Commun., **45**(1):5-25.[doi:10.1016/j.specom.2004.05.002]

[4] 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.

[5] Gersho, A., Gray, R.M., 1992. Vector Quantization and Signal Compression. Kluwer Academic Publishers, Boston, USA, p.345-372.