SINGLE-CHANNEL SOURCE SEPARATION OF AUDIO SIGNALS USING BARK SCALE WAVELET PACKET DECOMPOSITION

Yevgeni Litvin and Israel Cohen

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel
e-mail: elitvin@tx.technion.ac.il

ABSTRACT

We address the problem of blind source separation from a single channel audio source using statistical model of the sources. We modify the Bark Scale aligned Wavelet Packet Decomposition, to approximately acquire shift invariance. We allow oversampling in some decomposition nodes to equalize sample rate in all terminal nodes. Statistical models are trained from samples of each source separately. The separation is performed using these models. Experimental results show improved performance compared to a competing algorithm using synthetic and real audio examples.

1. INTRODUCTION

Blind source separation (BSS) of audio signals has been an active area of research in the recent years. BSS from a single audio channel is a special case of the general BSS problem where data from only one source is available to the algorithm. The problem becomes easier if the separated audio signals belong to different signal classes that can be classified based upon prior knowledge using existing statistical learning techniques. Different attempts to solve this problem in various contexts were made, including: statistical modeling such as Gaussian Mixture Model (GMM) [1], Hidden Markov Model (HMM)[2, 3]; Computational Auditory Scene Analysis (CASA) [4]; Non-negative Matrix Factorization (NMF); sparse decomposition and others. Although many existing solutions produce satisfactory results in special cases, the general problem of single audio channel BSS remains unsolved.

Traditionally, short time Fourier transform (STFT) is used in many audio and speech processing applications. Bark-Scaled Wavelet Packet Decomposition (BS-WPD) [5] is a time-frequency signal transformation with non uniform frequency resolution. This transformation reflects the critical bands structure of the human auditory system. Mapping based complex wavelet transform (CWT) [6] is based on bijective mapping of a real signal into a complex signal domain followed by standard wavelet analysis performed on the complex signal. Among others CWT partially mitigates lack of shift invariance of wavelet analysis.

Benaroya et al. [2] proposed and analyzed blind source separation using time varying Wiener filter in the STFT domain. First the GMM models for two different signal sources are trained using training samples. Then, the separation is performed by maximizing maximum a posterior (MAP) criterion.

This work was supported by the Israel Science Foundation under Grant 1085/05 and by the European Commission under project Memories FP6-IST-035300.

In this work we propose a source separation algorithm that follows Benaroya's STFT based algorithm, but operates on non uniform WPD filter-bank. We modify the BS-WPD analysis to equalize sampling rates of different scale-bands, which enables construction of instantaneous spectral shapes that are used in training and separation stages of the separation algorithm. We also use CWT in order to achieve some level of shift invariance. The non-uniform frequency resolution of the BS-WPD filterbank, reduces the dimension of feature vectors by allocating fewer vector elements in higher frequencies. This behavior is similar to the critical bands structure of human auditory system. In a series of experiments we validate our approach using various types of wavelet families and show that the proposed approach is capable of performing the separation task.

The remainder of this paper is structured as follows. In Section 2 we shortly describe the disadvantages of classical wavelet transform applied to audio processing tasks and the CWT transform. In Section 3 we describe Bark Scaled WPD and the modification designed to equalize sampling frequencies in all sub-bands. Section 4 presents our mixing model and MAP estimators for its components. Section 5 describes training and separation stages of the algorithm. Section 6 presents our experimental results.

2. MAPPING BASED COMPLEX WAVELET TRANSFORM

In this section we describe the disadvantages of standard Discrete Wavelet Transform (DWT) and present the Mapping Based Complex Wavelet Transform (CWT), introduced in [6], that mitigates these disadvantages to some degree.

A major disadvantage of DWT that reduces its usefulness in audio signal processing applications is the lack of shift invariance. Let $x(n)$ be a time domain signal and $X_{l,n}(m) = \mathrm{DWT}\{x(n)\}$ its DWT transform. Let $x_\Delta(n) = x(n-\Delta)$ be a shifted version of the time signal. The DWT coefficients of $x_\Delta(n)$ change significantly compared to $X_{l,n}(m)$. The reason for this behavior lies in the downsampling performed on the dilated signals by the DWT. A short survey of techniques used to mitigate lack of shift invariance may be found in [6].

Let $L^2(\mathbb{R}\to\mathbb{C})$ denote a function space of square integrable complex-valued functions on a real line and $L^2(\mathbb{R}\to\mathbb{R})$ its subspace comprised of real-valued functions. Hardy-space $H^2(\mathbb{R}\to\mathbb{C})$ is defined by

$$H^2(\mathbb{R}\to\mathbb{C}) \triangleq \left\{ f \in L^2(\mathbb{R}\to\mathbb{C}) : \mathcal{F}f(\omega) = 0 \text{ for a.e. } \omega < 0 \right\}$$

where $\mathcal{F}f(\omega)$ is a Fourier transform of $f(t)$.

In [6], a function space $L^2(\mathbb{R} \to \mathbb{R})$ is shown to be isomorphic to Hardy-space $H^2(\mathbb{R} \to \mathbb{C})$ under certain conditions. It is also shown therein, that the mapping of a function in $L^2(\mathbb{R} \to \mathbb{R})$ into Hardy-space cannot be implemented using a digital filter. Softy-space is a practical approximation of a Hardy-space. The mapping into Softy-space is done using a digital filter h^+. From now on, we denote signals in Softy-space by superscript "+".

Forward CWT transform is done by mapping time domain signal $x(n)$ into its Softy-space image $x^+(n)$ followed by standard DWT transform. The inverse CWT transform consists of Inverse Discrete Wavelet Transform (IDWT) followed by the inverse mapping from the complex valued Softy-space back to the real valued time signal.

Our algorithm benefits from approximate shift invariance property. We train the GMM model using the wavelet transform coefficients. Lack of shift invariance adds redundancy to the signal space making it larger and the amount of training data required will grow accordingly.

3. BARK-SCALED WAVELET PACKET DECOMPOSITION

In this section we present the BS-WPD and introduce a modification that has some favorable properties for the frame-by-frame classification used in our algorithm.

The BS-WPD introduced in [5] is a wavelet packet decomposition constructed in such way that center frequencies of wavelet packets are located approximately 1-Bark apart. In [5], Cohen adds two additional levels of decomposition to increase frequency resolution of sub-bands. From experimental evidence we saw that adding three additional levels of decomposition results in better performance in separation task.

We define a version of the BS-WPD transform that has equal sampling frequency in each sub-band, hence every time instance can be described by a single feature vector holding the instantaneous spectral information from all sub-bands. Unfortunately, terminal nodes of BS-WPD are located at various depths and each depth is associated with different sampling frequency. In order to align signals from all sub-bands and equalize the sampling frequency we do not decimate the lowpass and detail signal in nodes with $l > 6$. We call this transform Constant Sampling Rate BS-WPD (CSR-BS-WPD). We note that by canceling decimation in lower levels of the WPD tree we introduce a certain amount of redundancy into CSR-BS-WPD representation. Fig. 1 shows CSR-BS-WPD decomposition tree.

The CSR-BS-WPD analysis produces only 168 sub-bands, compared to 513 sub-bands of STFT analysis with approximately the same frequency resolution in low frequencies. We sacrifice frequency resolution at higher frequency range, in accordance with human auditory system which also has a coarser resolution in high frequency range. Reduction in the number of sub-bands, results in smaller dimension of data that is used in training and separation stages. Smaller data dimension has a potential to increase accuracy of the GMM estimation because of the reduced redundancy in feature vectors and to reduce computational burden.

Let $x(n)$ be a time sequence and

$$x^+(n) \quad = \quad h^+(n) * x(n) \tag{1}$$

its Softy-space image. We denote the CSR-BS-WPD transform of $x^+(n)$ as $X_{l,n}(m)$ where (l,n) are indices of terminal nodes and m is time index. Since all terminal nodes have the same sampling rate we can rearrange the elements of $X_{l,n}(m)$ into a single column vector with dimension equal to the number sub-band denoted by $\bar{X}(m) \in \mathbb{C}^M$, where $M = 168$ is the number of sub-bands.

4. MIXTURE COMPONENTS ESTIMATION

Let $s_1(n)$ and $s_2(n)$ be mixture components. We assume a mixing model without noise presence

$$x(n) \quad = \quad s_1(n) + s_2(n) \tag{2}$$

Mapping mixture signal into Softy-space we get

$$x^+(n) \quad = \quad s_1^+(n) + s_2^+(n) \tag{3}$$

and in CSR-BS-WPD domain

$$\bar{X}(m) \quad = \quad \bar{S}_1(m) + \bar{S}_2(m) \tag{4}$$

$$\bar{S}_1(m), \bar{S}_2(m), \bar{X}(m) \quad \in \quad \mathbb{C}^M$$

We use posterior mean (PM) to estimate mixture components in CSR-BS-WPD domain. Let x, s_1, s_2 be vectors of N observations of $x(n)$, $s_1(n)$ and $s_2(n)$ accordingly. It is shown in [2] that for the mixing model (2) and Gaussian processes $s_1(n)$, $s_2(n)$, the PM estimators for s_1 and s_2 are given by

$$\hat{s}_c \quad = \quad \Sigma_c (\Sigma_1 + \Sigma_2)^{-1} x,\ c \in \{1,2\} \tag{5}$$

where $\Sigma_c \in \mathbb{R}^{N \times N}$ is the covariance matrix of s_c. For stationary and approximately circular processes, Fourier transform $\mathcal{F}$ diagonilizes the covariance matrices resulting in the following estimators:

$$\hat{\mathcal{F}s_c}(f) \quad = \quad \frac{\sigma_c^2(f)}{\sigma_1^2(f) + \sigma_2^2(f)} \mathcal{F}x(f),\ c \in \{1,2\} \tag{6}$$

where σ_c^2 is a vector of eigen-values of Σ_c and f is frequency index.

The covariance matrix diagonality means lack of correlations between signals in different frequency bands. We can also interpret CSR-BS-WPD transform as a filterbank, and we can expect that as long as WPD filters have good band-pass filter characteristics, the correlation between different frequency bands will be also low and diagonal covariance matrix assumption can also be extended to the CSR-BS-WPD signals.

Simple assumption of Gaussian distribution prior does not hold for most natural signals such as speech or music. The remedy is to assume Gaussian Mixture prior densities (GMM prior) [2]. GMM model describes signal distribution as an outcome of a two stage process: first an active component k is selected out of K Gaussian distributions in the mixture; then an observation sample is obtained using the selected model parameters $\left\{\mu^{(k)}, \Sigma^{(k)}\right\}$ where $\mu^{(k)}$ and $\Sigma^{(k)}$ are the expectation value and the covariance of the k-th component. The probability of selecting k-th component is given by w_k (k-th element of probability vector w). The GMM model is defined by $\left(\left\{\mu^{(k)}\right\}_{k=1}^{K}, \left\{\Sigma^{(k)}\right\}_{k=1}^{K}, w\right)$.

We estimate mixture components using GMM priors. We introduce hidden variables $q_c(m) \in \{1,\ldots,K\}, c \in \{1,2\}$ associated with index of active GMM component at time m. Let

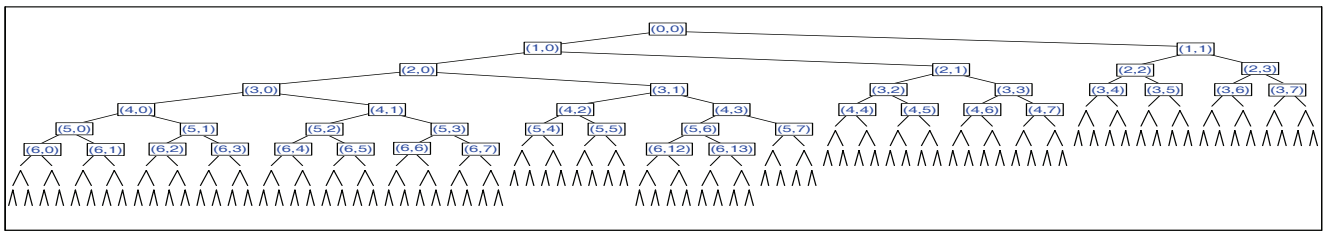

Fig. 1. CSR-BS-WPD decomposition tree. Nodes having $l > 6$ are not decimated. This way, sampling frequencies of signals in all terminal nodes will be the same. Only few of the node labels are shown due to the space limitations.

$\gamma_{j,k}(m) = p(q_1(m) = j, q_2(m) = k|x)$ be posterior probabilities of active components. $\gamma_{j,k}(m)$ is estimated from mixture observations. Conditioning (2) on active GMM component and applying to the CSR-BS-WPD signal we get PM estimator

$$\hat{\bar{S}}_1(m) = \sum_{j,k=1}^{K} \gamma_{j,k}(m)\,\Sigma_1^{(j)}\left(\Sigma_1^{(j)} + \Sigma_2^{(k)}\right)^{-1}\bar{X}(m) \quad (7)$$

The estimator for $\hat{\bar{S}}_2(m)$ is derived in the same manner.

5. TRAINING AND SEPARATION

Let L be a number of training signal time samples in CSR-BS-WPD domain. During the training stage we use signal samples of both classes $\left\{\bar{S}_1(m)\right\}_{m=1}^{L}, \left\{\bar{S}_2(m)\right\}_{m=1}^{L}$ to train two different GMM models. Both Softy-space mapping and the WPD are linear transformations, we conclude that expectation values of s, s^+ and $\bar{S}$ are zero, hence we can define a simplified GMM model that assumes zero mean of every state in the GMM

$$\Lambda_c = \left(w_c, \left\{\Sigma_c^{(k)}\right\}_{k=1}^{K}\right), w_c \in \mathbb{R}^K, \Sigma_c \in \mathbb{R}^{M\times M} \quad (8)$$

where K is the GMM model order. Following the reasoning in the previous section, we assume $\Sigma_c^{(k)}$ to be a diagonal covariance matrix.

The training of the GMM models is performed using Expectation Maximization (EM) algorithm and bootstrapped using K-Means algorithm. Expectation value of training data is assumed to be zero, it is not updated during the expectation step of the EM algorithm and set constantly to zero.

We note that the estimation of $\hat{\bar{S}}_c(m)$ is performed for every time index m. In the rest of this section we omit time index m for the clearness of notation. In order to estimate signal sources $\hat{\bar{S}}_c$ using (7) for every time instance, we first estimate posterior probability $\gamma_{j,k}$:

$$\begin{aligned}\gamma_{j,k} &\propto p\left(\bar{X}|q_1 = j, q_2 = k\right)p(q_1 = j)\,p(q_2 = k) \quad (9)\\ &= g\left(\bar{X}; \Sigma_1^{(j)} + \Sigma_2^{(k)}\right) w_1^{(j)} w_2^{(k)}\end{aligned}$$

where $g\left(\bar{X}; \Sigma\right)$ is a zero mean multivariate Gaussian probability density function. Substituting (9) into (7) and using Λ_c estimated in the training process we acquire estimators for $\hat{\bar{S}}_1$ and $\hat{\bar{S}}_2$.

	SDR_1	SIR_1	SAR_1	SDR_2	SIR_2	SAR_2
STFT	16	40	16	16	31	16
CSR-BS-WPD	20	35	20	22	40	22

Table 1. Separation performance measures for separation of synthetically generated signals for different algorithms. The measures are shown in dB.

6. EXPERIMENTAL RESULTS

In order to evaluate performance of the proposed algorithm and compare it to the STFT based algorithm [7] we use SDR, SIR and SAR distortion measures described in [8]. Now we demonstrate the effectiveness of the proposed algorithm on synthetic signals and on speech and piano music mixtures.

6.1. Synthetic Signals

First we evaluated the performance of the separation algorithm using synthetic signals. Our goal is to verify the feasibility of the separation in the CSR-BS-WPD domain. We constructed synthetic signals by dividing the entire signal time span into 400ms segments and each segment is generated by

$$x_c(m) = \begin{cases} \sum_{i=1}^{2}\cos\left(\frac{2\pi}{f_s}f_{1,i}^{(c)}n\right) & w.p.\frac{1}{2} \\ \sum_{i=1}^{2}\cos\left(\frac{2\pi}{f_s}f_{2,i}^{(c)}n\right) & w.p.\frac{1}{2}\end{cases} \quad (10)$$

where $f_{1,1}^{(1)} = 220$Hz, $f_{1,2}^{(1)} = 440$Hz, $f_{1,1}^{(2)} = 300$Hz, $f_{1,2}^{(2)} = 600$Hz. Two different signals were obtained for training and evaluation. We used GMM with two mixture components.

Table 1 shows the separation performance. High values of SIR indicate high rejection of interfering signal and relatively high values of SDR indicate low amount of distortion introduced by our separation algorithm. Both algorithms perform very well on these synthetic signals.

6.2. Natural Signals

We also evaluated the performance of the proposed algorithm on natural signals. We separated two audio source classes: speech and piano excerpts. We used training sequence of 50 seconds for model training and 11 seconds for the performance evaluation. Different speech and piano excerpts were used for training and performance evaluation. Speech signals were taken from TIMIT database and include male speakers only. Piano excerpts were taken from Chopin's preludes. All signals were sampled at 16

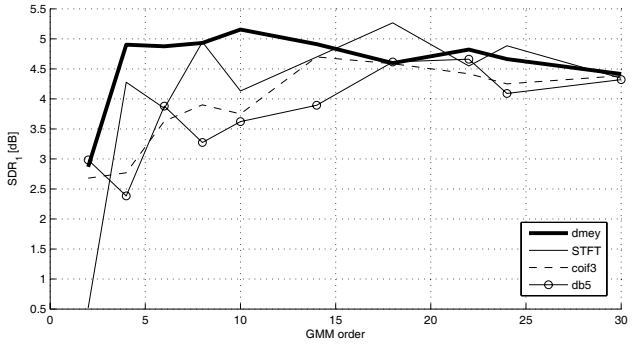

Fig. 2. Influence of the GMM model order on the signal to distortion ratio (SDR_1) of the speech signal. STFT based algorithm [7] is compared to CSR-BS-WPD based algorithm.

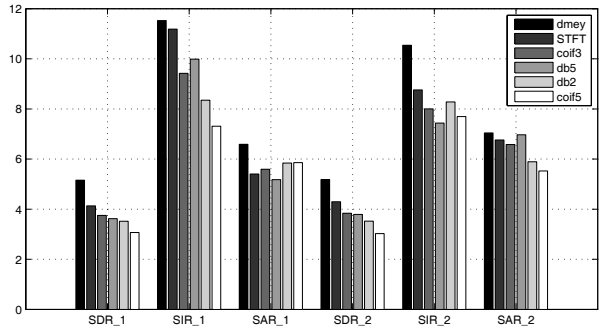

Fig. 3. Performance comparison of all wavelet families and the STFT based algorithm for GMM order of 10.

KHz. We equalized the energy of all signals before training and before mixing the signals for evaluation.

We compare the separation performance for different GMM model order and different wavelet families. Fig. 2 depicts signal to distortion ratio of speech signal (SDR_1). A higher value of SDR_1 indicates a smaller degree of speech distortion after the separation. For low orders of GMM, the CSR-BS-WPD analysis based on the discrete Meyer ($dmey$) wavelet family outperforms other tested wavelet families and STFT based algorithm. For high orders of GMM, $dmey$ based algorithm shows performance comparable to the STFT based algorithm and slightly better performance than other wavelet families. Although Fig. 2 depicts only the SDR_1 measure, other performance measures ($\mathrm{SDR}_i, \mathrm{SIR}_i, \mathrm{SAR}_i, \mathrm{LSD}_i$) showed similar behavior. Fig. 3 shows the performance of different wavelet families and the STFT based algorithm for GMM model order of 10. The $dmey$ wavelet family based algorithm shows performance superior to STFT for all objective measures compared. Other wavelet families, however show inferior separation performance to the STFT based algorithm.

In additional experiments we noticed that when $dmey$ based analysis is used, the sparseness of music and speech signals in the CSR-BS-WPD domain is highest compared to other wavelet families used in separation experiments. In [5], $dmey$ wavelet family is also used for speech enhancement and motivated by the regularity of the wavelet filter and its good frequency localization properties. Informal listening tests indicate that the CSR-BS-WPD based separation produces less "jumpy" and more pleasant signal reconstruction than the STFT based version of the algorithm.

7. CONCLUSIONS AND FUTURE WORK

Refining the frequency resolution of the critical band decomposition and introduction of a subsampling scheme in the BS-WPD decomposition tree, enables to adapt a source separation algorithm that was originally designed to work in the uniformly sampled STFT domain. Manipulation of high frequency information with coarse resolution allows us to reduce the computational burden and achieve better perceptual quality. Discrete Meyer wavelet based CSR-BS-WPD analysis yields improved separation performance compared to STFT analysis and other wavelet families tested. Future work with CSR-BS-WPD analysis may address various audio processing tasks traditionally performed in the STFT domain, which require instantaneous spectral shapes and have some relations to critical bands in human auditory system.

8. ACKNOWLEDGMENT

The authors thank Prof. Shalom Raz for teaching a graduate course on "Time-frequency methods and their applications" in the Technion, which inspired and encouraged this work.

9. REFERENCES

[1] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564–1578, July 2007.

[2] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, Jan. 2006.

[3] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems (NIPS) 13*. 2001, pp. 793–799, MIT Press.

[4] F. R. Bach and M. I. Jordan, "Blind one-microphone speech separation: A spectral learning approach," in *Advances in Neural Information Processing Systems 17*, Lawrence K. Saul, Yair Weiss, and Léon Bottou, Eds., pp. 65–72. MIT Press, Cambridge, MA, 2005.

[5] I. Cohen, "Enhancement of speech using bark-scaled wavelet packet decomposition," in *Eurospeech*, 2001, pp. 1933–1936.

[6] F.C.A. Fernandes, R.L.C. van Spaendonck, and C.S. Burrus, "A new framework for complex wavelet transforms," *IEEE Transactions on Signal Processing*, vol. 51, no. 7, pp. 1825–1837, July 2003.

[7] L. Benaroya and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," in *ICA2003*, Nara, Japan, Apr. 2003, pp. 957–961.

[8] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. 4th International Symposium on ICA and BSS (ICA2003)*, Nara, Japan, Apr. 2003, pp. 763–768.