J Sign Process Syst (2011) 65:339–350
DOI 10.1007/s11265-010-0510-9

Single-Channel Source Separation of Audio Signals Using Bark Scale Wavelet Packet Decomposition

Yevgeni Litvin · Israel Cohen

Received: 29 December 2009 / Revised: 2 May 2010 / Accepted: 20 July 2010 / Published online: 6 August 2010

Abstract We address the problem of blind source separation from a single channel audio source using a statistical model of the sources. We modify the Bark Scale aligned Wavelet Packet Decomposition, to acquire approximate-shiftability property. We allow oversampling in some decomposition nodes to equalize sampling rate in all terminal nodes. Statistical models are trained from samples of each source separately. The separation is performed using these models. The proposed psycho-acoustically motivated non-uniform filterbank structure reduces signal space dimension and simplifies training procedure of the statistical model. In our experiments we show that the proposed algorithm performs better when compared to a competing algorithm. We study the effect that different wavelet families have on the performance of the proposed signal analysis in the single-channel source separation task.

Keywords Audio source separation · BS-WPD · CSR-BS-WPD · GMM · CWT · Monaural source separation

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

Y. Litvin (✉) · I. Cohen
Department of Electrical Engineering, Technion—Israel Institute of Technology, Technion City, Haifa 32000, Israel
e-mail: selitvin@gmail.com

I. Cohen
e-mail: icohen@ee.technion.ac.il

1 Introduction

Blind source separation (BSS) is the task of recovering a set of signals from a set of observed signal mixtures. The problem of BSS is common for different signal processing tasks. It is also at the heart of numerous applications in audio signal processing. BSS algorithms that operate on audio signals are sometimes called Blind Audio Source Separation (BASS) algorithms [1].

Cherry [2] coined the ability of the human hearing system to concentrate on a single speaker in presence of interfering signals as a "cocktail party effect". Although, human audio segregation abilities are fascinating, not necessarily a full audio separation is performed in the inner ear or somewhere in the auditory cortex. It is possible that the human hearing system is only capable of recognizing semantic objects in one of several audio streams the listener is exposed to.

Different settings for the BSS task arise in different applications. In different settings the prior and the posterior information available to a source separation algorithm may differ, such as number of sources and number of observed channels; mixing model (instantaneous, echoic, convolutive, linear, non-linear); prior information on statistical properties of signals; and presence of noise.

One of the crucial factors in the definition of the BSS problem is the ratio of the number of observed channels to the number of audio sources in the mixture. If the number of observed channels is equal to the number of sources then it is called an even-determined or a determined case. In an over-determined case the number of channels is greater than the number of sources and in an under-determined case the number of channels is smaller than the number of sources. The

under-determined case is the most difficult to handle and requires stronger assumptions on the mixture component properties.

Another important factor that differentiates between BSS problem setups is the mixing model. The instantaneous mixing model implies that several instantaneous mixtures are observed, each having source components mixed in a different proportion. Echoic mixing model allows different delays for each component in each channel. The convolutive mixing model allows different linear filtering of sources at each channel. Naturally, the instantaneous mixing model is a degenerate case of the echoic mixing model and the echoic model is a degenerate case of the convolutive mixing model. The convolutive mixing model is the most appropriate in describing most of the real world scenarios, but is also the hardest to handle.

Most source separation algorithms assume that mixture components are statistically independent. Although, this is a reasonable assumption in many cases, it is not necessarily true for all applications. For example, one of the source separation applications is the separation of an individual musical instrument from a polyphonic musical excerpt. In this case, the assumption of statistical independence is inaccurate for most musical styles where several musical instruments perform parts in a certain musical key and according to a common tempo.

The blind source separation problem was first formulated in a statistical framework by Herault et al. in 1984. Comon [3] introduced the Independent Component Analysis (ICA) in 1994 and numerous theoretical and practical works followed. A basic ICA algorithm assumes even-determined BSS case and instantaneous mixing model. Under these assumptions, a demixing matrix has to be found. In order to find such matrix the ICA algorithm minimizes statistical dependency between unmixed channels. Various methods may be used in order to reduce statistical dependency, such as maximization of non-Gaussianity between channels or minimization of mutual information [4]. The search is usually done using gradient descent or fixed point algorithms. Unfortunately, most of the algorithms in the ICA family require several mixtures to be observed in order to perform the separation.

In some cases a database of audio samples is available and statistical signal models can be trained in a supervised manner before the separation process. In these cases, various techniques from statistical learning can be used. Algorithms that rely on these kind of statistical models are sometimes called Semi-Blind Source Separation Algorithms (SBSS) [5, 6].

In [7], Benaroya et al. introduced a source separation algorithm based on Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) statistical modeling of source signal classes. First GMM or HMM models are trained for each signal class using spectral shapes acquired from the Short-Time Fourier Analysis (STFT) analysis. During the separation stage, these models are used to estimate mixture components using Maximum A-posterior (MAP) or Posterior Mean (PM) estimates. The authors also showed that using more complicated HMM models does not improve the separation performance significantly when compared to the GMM model. Some extensions to that work were presented in [6]. For example, Gaussian Scaled Mixture Model (GSMM) which takes into account variations in amplitude of sounds with similar spectral shapes.

Another signal modeling technique that was found useful in single channel source separation is Auto Regressive (AR) modeling. Srinivasan et al. [8] proposed a codebook of Linear Predictive Coefficients (LPC) trained on a speech and an interfering signal. The maximum likelihood estimator is used to find the most probable pair of codebook members. Wiener filter is used later to suppress the interfering signal. In [9] LPC coefficients are treated as random variables. In these works both algorithms are described and tested for speech enhancement in non stationary noise setup. Nevertheless, they are also applicable to a source separation scenario by modeling one of the sources as the speech and the other as the noise.

Traditionally, short time Fourier transform (STFT) is used in many audio and speech processing applications. Bark-Scaled Wavelet Packet Decomposition (BS-WPD) [10] is a time-frequency signal transformation with non uniform frequency resolution. This transformation is psychoacoustically motivated and reflects the critical bands structure of the human auditory system. Mapping based Complex Wavelet Transform (CWT) [11] 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.

The algorithm presented in this paper, addresses a single-channel separation of instantaneous mixture of two audio sources. It follows Benaroya et al. [7] STFT based algorithm, but operates with a 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 to the higher frequencies. This behavior mimics 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 performs better when compared to a competing algorithm in some scenarios. Partial results of this work were presented in [12].

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 present the CWT transform. In Section 3 we introduce Bark Scaled WPD and a modification designed to equalize the sampling frequencies in all subbands. In Section 4 we formulate our mixing model and derive MAP estimators for its components. Section 5 specifies training and separation stages of the algorithm. Section 6 outlines the performance measures, and Section 7 shows some experimental results.

2 Mapping Based Complex Wavelet Transform

In this section we describe the disadvantages of the standard Discrete Wavelet Transform (DWT) and present the mapping based Complex Wavelet Transform (CWT), introduced by Fernandes et al. in [11, 13], that mitigates these disadvantages to some degree.

A major disadvantage of the 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. Different researchers proposed various methods to address the problem. A survey of techniques used to mitigate lack of shift invariance may be found in [13].

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 $\mathcal{F}f(\omega)$ denotes the Fourier transform of $f(t)$. The mapping of a signal to the Hardy-space is equivalent to finding its analytic signal.

Fernandes et al. showed that a function space $L^2(\mathbb{R}\to\mathbb{R})$ is isomorphic to Hardy-space $H^2(\mathbb{R}\to\mathbb{C})$ under a certain inner product. They also showed that the mapping of a function in $L^2(\mathbb{R}\to\mathbb{R})$ into Hardy-space cannot be implemented using a digital filter. As a remedy, they defined Softy-space and proved that it has properties similar to the Hardy-space. The mapping of a function in $L^2(\mathbb{R}\to\mathbb{R})$ into the Softy-space is done using a projection digital filter h^+ that has a passband over $[0,\pi)$ and a stopband over $[-\pi,0)$. Softy-space signals are denoted by a superscript "+" in [13] because of the attenuated negative frequencies.

We adapt this notation. Let $x(n)$ be a time sequence. Its Softy-space image is given by

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

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

$$x(n) = \mathcal{R}e\left\{ g^+(n) * x^+(n) \right\}. \tag{2}$$

where $g^+(n)$ is also a digital filter. The relation between $h^+(n)$ and $g^+(n)$ is described in [13].

Simoncelli et al. [14] defined a "shiftability" as a transform property that guarantees that transform subband energy is invariant under signal shifts. The shiftability property is weaker than shift invariance but easier to achieve. Simoncelli et al. argued that a transform is shiftable if and only if there is no aliasing any subband. It follows that the shiftability it is not achievable for any non-redundant wavelet transform, except for the practically unrealizable Shannon wavelet.

A Hardy-space signal has half the bandwidth of the corresponding real-valued function due to the lack of negative frequency components. Nyquist condition holds for the Hardy-space signal in each analysis subband and the shiftability property follows. The practical CWT transform uses an approximation of the Hardy-space, hence the CWT is approximately shiftable provided sufficiently long filters are used in the implementation as explained in [11]. We note that because of the complex transform coefficients, mapping based CWT is a redundant transform with a redundancy factor of two.

Our algorithm benefits from approximate shift invariance property. We train GMM model using the

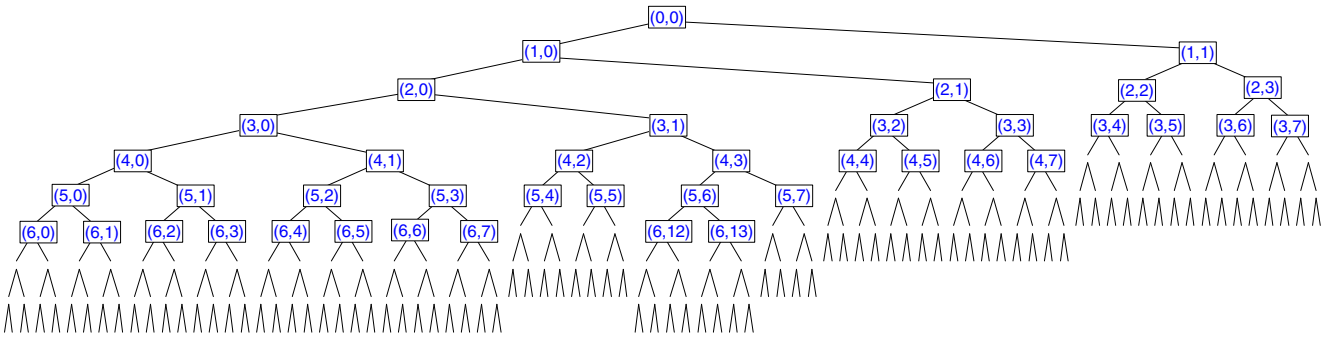


Figure 1 CSR-BS-WPD decomposition tree. Nodes having $l > 6$ are not decimated. This way, sampling frequencies of signals in all terminal nodes will be the same. Only few of the node labels are shown due to the space limitations.

wavelet transform coefficients. Lack of shift invariance makes the signal space unnecessarily large. This can make the statistical modeling of the signal more complicated in terms of the selected statistical model, computational burden or the amount of data required for training.

3 Bark-Scaled Wavelet Packet Decomposition

In this section we present the BS-WPD as defined in [10] and introduce a modification that has some favorable properties for the frame-by-frame classification and filtering used in our algorithm.

Let $E \subset \{(l,n) : 0 \le l < L, 0 \le n < 2^l\}$ be a set of terminal nodes of a WPD tree. The center frequency of a terminal node $(l,n) \in E$ is roughly given by

$$f_{l,n} = 2^{-l}\left(\mathrm{GC}^{-1}(n) + 0.5\right) f_s,$$

where $\mathrm{GC}^{-1}(\mathrm{n})$ is the inverse Gray code of n and f_s is a sampling frequency. Critical band WPD (CB-WPD) [10] filterbank structure is obtained by selecting a terminal nodes set E in a way that positions center frequencies $f_{l,n}$ approximately one Bark apart. Another constraint that must be taken into consideration is that a dyadic interval set $\{I_{l,n} : I_{l,n} = 2^{-l}n, 2^{-l}(n+1)\}$ must form a disjoint cover of [0, 1). Only in this case, the set of wavelet packet family functions will be able to span the signal space.

BS-WPD is defined as a CB-WPD with two additional levels of terminal node expansion. Due to a higher frequency resolution, the BS-WPD performed better in the task of speech enhancement. Our experiments showed that adding three additional levels of decomposition results in better performance of our source separation algorithm.

The proposed source separation algorithm needs an access to the instantaneous spectral information from all analysis subbands. We define a new version of the BS-WPD transform that has equal sampling frequency in each sub-band. Unfortunately, terminal nodes of the BS-WPD are positioned at different depths and each depth is associated with a different sampling frequency. In order to align signals from all sub-bands and equalize the sampling frequency we do not allow decimation in nodes deeper than level 6 (i.e. $l > 6$). We call the resulting 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 the CSR-BS-WPD representation. Figure 1 shows CSR-BS-WPD decomposition tree.

The CSR-BS-WPD analysis has only 168 sub-bands, compared to 513 sub-bands of STFT analysis with approximately the same frequency resolution in low frequencies. Like the human auditory system, we sacrifice frequency resolution at higher frequency range. Reducing the number of sub-bands results in smaller dimension of data used in the training and separation stages. A smaller data dimension has a potential to

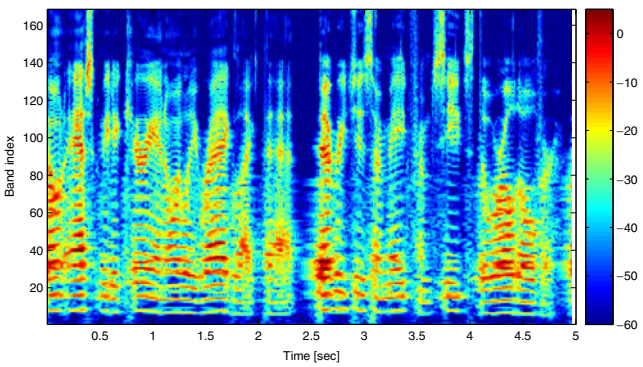


Figure 2 CSR-BS-WPD time-frequency representation of a speech signal. *Horizontal axis*.

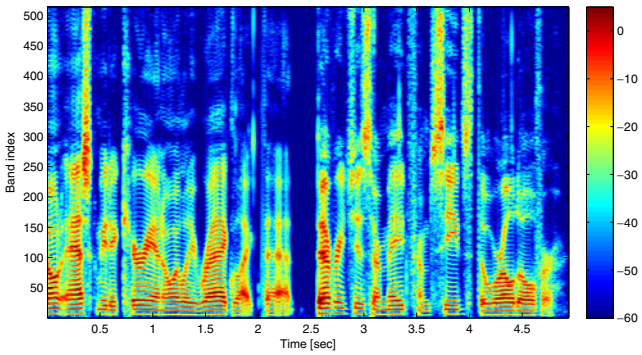


Figure 3 STFT representation of a speech signal.

increase accuracy of the GMM estimation because of less redundancy in the high frequency features, in addition to lower computational burden.

We denote by $X_{l,n}(m)$ the CSR-BS-WPD transform of $x^+(n)$ in Eq. 2, where (l, n) are the 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 complex, single column vector $\bar{X}(m) \in \mathbb{C}^M$. The dimension of $\bar{X}(m)$ is equal to the number of sub-bands $M = 168$. The CSR-BS-WPD is a reversible transform. Given CSR-BS-WPD signal $\bar{X}(m)$ we can acquire a time domain signal $x(n)$ first by inverting the WPD and then filtering $x^+(n)$ with a digital filter g^+ as given by Eq. 2.

Figures 2 and 3 show a time frequency plot of a speech signal. The intensity of color shows the coefficient magnitude in dB. Band indices are shown on a vertical axis. In Fig. 2, band indices on the vertical axis are the indices of the terminal nodes in the growing mid-band frequency order. It can be seen that less time-frequency coefficients are dedicated to the higher frequency bands in the CSR-BS-WPD analysis than in the conventional STFT analysis.

4 Mixture Components Estimation

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

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

Mapping the mixture signal into the Softy-space we get

$$x^+(n) = s_1^+(n) + s_2^+(n), \tag{4}$$

and in the CSR-BS-WPD domain

$$\bar{X}(m) = \bar{S}_1(m) + \bar{S}_2(m) \tag{5}$$

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

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 $s_2(n)$ respectively. It is shown by Benaroya et al. in [6] that for the mixing model (3) and Gaussian processes $s_1(n)$, $s_2(n)$, the PM estimators for s_1 and s_2 are given by

$$\hat{s}_c = \Sigma_c (\Sigma_1 + \Sigma_2)^{-1} x,\ c \in \{1, 2\}, \tag{6}$$

where $\Sigma_c \in \mathbb{R}^{N \times N}$ is the covariance matrix of s_c.

For a stationary and approximately circular processes, Fourier transform $\mathcal{F}$ diagonalizes the covariance matrices resulting in the following estimators:

$$\widehat{\mathcal{F}s}_c(f) = \frac{\sigma_c^2(f)}{\sigma_1^2(f) + \sigma_2^2(f)} \mathcal{F}x(f),\ c \in \{1, 2\}, \tag{7}$$

where σ_c^2 is a vector of eigen-values of Σ_c and f is frequency index. We notice that the solution coincides with is a well known Wiener filter.

Diagonal covariance matrix of the STFT expansion coefficients is often assumed in the speech processing applications [15]. Benaroya used this assumption in the application to musical signals. The covariance matrix diagonality means lack of correlations between signals

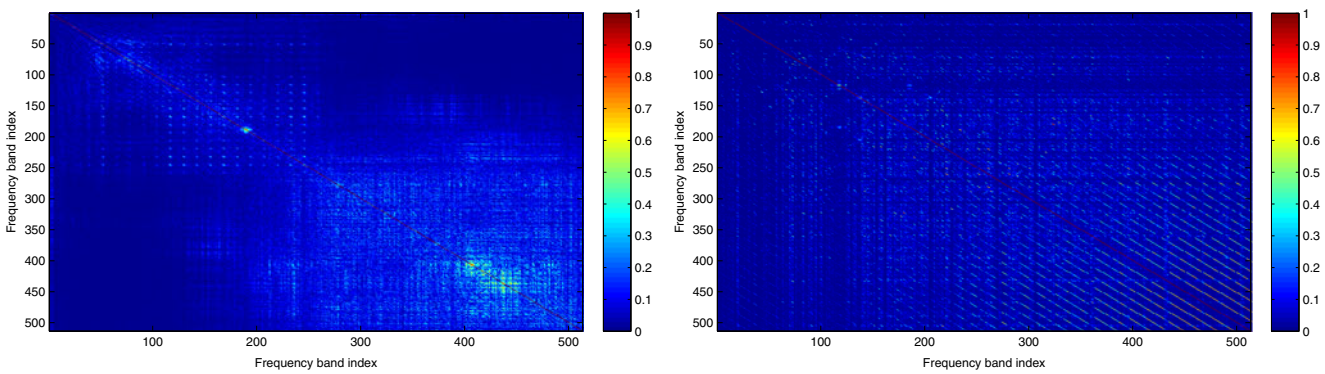


Figure 4 Correlation coefficients matrix of STFT expansion for: **a** speech signal; **b** musical signal.

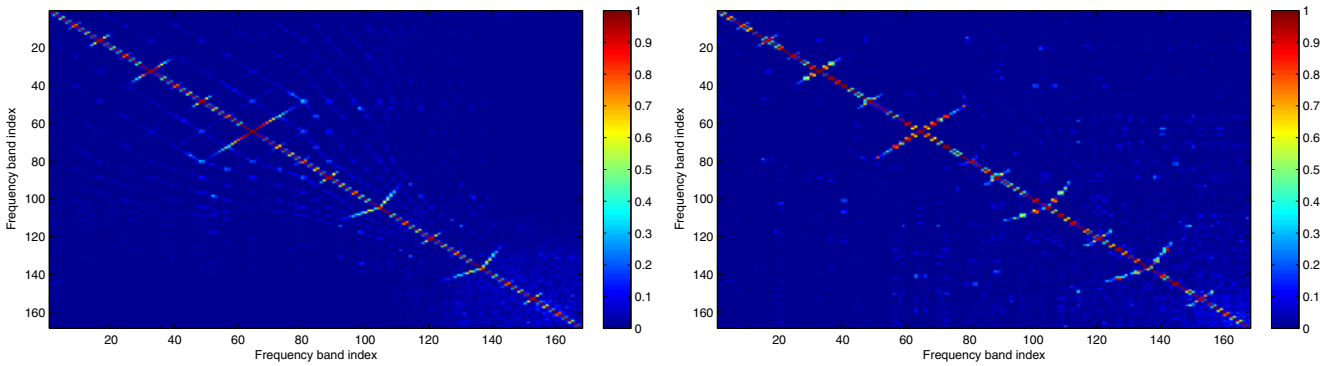


Figure 5 Correlation coefficients matrix of CSR-BS-WPD expansion using *dmey* wavelet for: **a** speech signal; **b** musical signal.

in different frequency bands. We can interpret CSR-BS-WPD transform as a filterbank. We can expect that as long as WPD filters have good band-pass filter characteristics, the correlation between different frequency bands will also be low and diagonal covariance matrix assumption can also be extended to the CSR-BS-WPD signals. Next, we verify this assumption on empirical data.

Figures 4, 5 and 6 show absolute value of correlation coefficient matrices for STFT and CSR-BS-WPD transforms. The CSR-BS-WPD transform is based on the discrete Meyer wavelet (*dmey*) and Daubechies wavelet of order 2 (*db2*). It is clear that the diagonal covariance assumption is inaccurate for all transforms, including STFT. Highest amount of correlation is present in the *db2* based CSR-BS-WPD transform, especially for the musical signal. High inter-band correlation may be explained either by correlated events occurring at different frequencies or by a leakage of energy between frequency bands due to non-ideal bandpass characteristics of analysis filters. Since same audio signals were used in Figs. 4–6, the differences in the correlation coefficients are explained by the properties of analysis filters. The analysis filter of the *db2* wavelet is only 4 samples long compared to a more than 100 coefficients of the *dmey* analysis filter. *db2* wavelet has inferior frequency localization and worse attenuation of side lobes compared to the *dmey* analysis filter, hence the correlation coefficients are higher for the *db2* based CSR-BS-WPD expansion coefficients.

Despite its weakness, especially for some kinds of wavelet families, we assume the lack of correlation between expansion coefficients since the estimation of the large number off diagonal elements of a covariance matrix is impractical.

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) [6]. 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

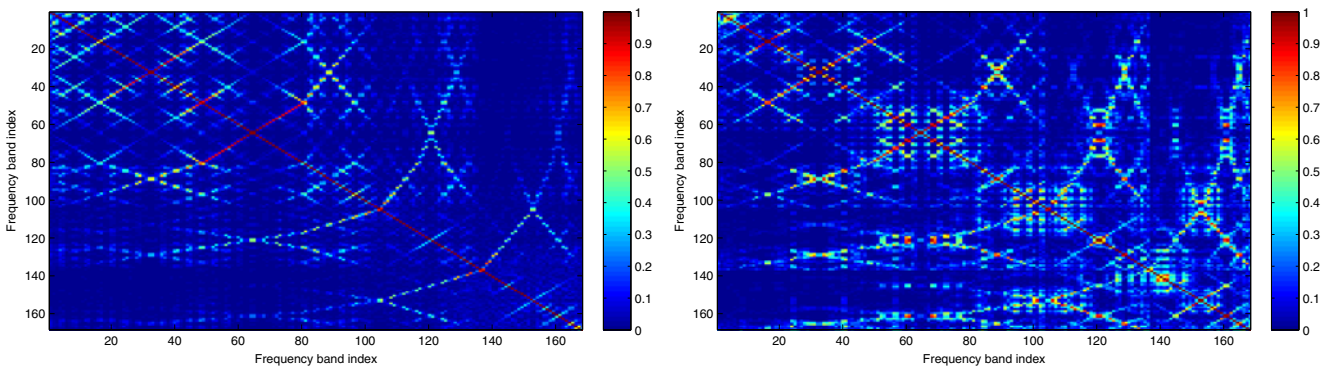


Figure 6 Correlation coefficients matrix of CSR-BS-WPD expansion using *db2* wavelet for: **a** speech signal; **b** musical signal.

observation sample is obtained using the selected model parameters $\{\mu^{(k)}, \Sigma^{(k)}\}$ where $\mu^{(k)}$ and $\Sigma^{(k)}$ are the expectation value and the covariance of the k-th component. The probability of selecting the k-th component is given by w_k (k-th element of probability vector w). The GMM model is defined by $\left(\{\mu^{(k)}\}_{k=1}^{K}, \{\Sigma^{(k)}\}_{k=1}^{K}, w\right)$.

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 $\gamma_{j,k}(m) = p(q_1(m) = j, q_2(m) = k|x)$ be the posterior probabilities of active components. Then $\gamma_{j,k}(m)$ is estimated from mixture observations. Conditioning Eq. 3 on active GMM component and applying to the CSR-BS-WPD signal we get the PM estimator

$$\hat{\bar{S}}_1(m) = \sum_{j,k=1}^{K} \gamma_{j,k}(m)\, \Sigma_1^{(j)} \left(\Sigma_1^{(j)} + \Sigma_2^{(k)}\right)^{-1} \bar{X}(m). \quad (8)$$

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

5 Training and Separation

Let L denote the number of training signal time samples in the CSR-BS-WPD domain. During the training stage we use signal samples of both classes $\{\bar{S}_1(m)\}_{m=1}^{L}, \{\bar{S}_2(m)\}_{m=1}^{L}$ to train two different GMM models. Both Softy-space mapping and the WPD are linear transformations, so the 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

$$\Lambda_c = \left(w_c, \{\Sigma_c^{(k)}\}_{k=1}^{K}\right), w_c \in \mathbb{R}^K, \Sigma_c \in \mathbb{R}^{M \times M}, \quad (9)$$

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 the Expectation Maximization (EM) algorithm [16] bootstrapped by the K-Means algorithm. Expectation value of training data is assumed to be zero and is not updated during the expectation step of the EM algorithm, i.e. is set 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 the time index m for clearness of notation. In order to estimate signal sources $\hat{\bar{S}}_c$ using Eq. 8 for every time instance, we first estimate the posterior probability $\gamma_{j,k}$:

$$\begin{aligned}\gamma_{j,k} &\propto p\left(\bar{X}|q_1 = j, q_2 = k\right) p(q_1 = j)\, p(q_2 = k) \\ &= g\left(\bar{X}; \Sigma_1^{(j)} + \Sigma_2^{(k)}\right) w_1^{(j)} w_2^{(k)}, \end{aligned} \quad (10)$$

where $g\left(\bar{X}; \Sigma\right)$ is a zero-mean multivariate Gaussian probability density function. Substituting Eq. 10 into Eq. 8 and using Λ_c estimated in the training process we obtain estimators for $\hat{\bar{S}}_1$ and $\hat{\bar{S}}_2$.

6 Evaluation Criteria

In this section, we define evaluation criteria used in experiments to evaluate the performance of the proposed algorithm.

We use common distortion measures described in [17] and BSS_EVAL toolbox [18]. Mixture components s_1, s_2 are assumed to be uncorrelated. Let $\hat{s}_c$ be an estimate of s_c. The estimator will have the following decomposition:

$$\begin{aligned} \hat{s}_c &= y_c + e_{c,\text{interf}} + e_{c,\text{artif}} \\ y_c &\triangleq \langle \hat{s}_c, s_c \rangle s_c \\ e_{c,\text{interf}} &\triangleq \langle \hat{s}_c, s_{c'} \rangle s_{c'} \\ e_{c,\text{artif}} &\triangleq \hat{s}_c - \left(y_c + \langle \hat{s}_c, s_{c'} \rangle s_{c'}\right), \end{aligned}$$

where c is the target class and c' is the interfering class. Now the following performance measures are defined:

$$\text{SDR} \triangleq 10 \log_{10} \frac{\|y_c\|^2}{\left\|e_{c,\text{interf}} + e_{c,\text{artif}}\right\|^2}$$

$$\text{SIR} \triangleq 10 \log_{10} \frac{\|y_c\|^2}{\left\|e_{c,\text{interf}}\right\|^2}$$

$$\text{SAR} \triangleq 10 \log_{10} \frac{\left\|y_c + e_{c,\text{interf}}\right\|^2}{\left\|e_{c,\text{artif}}\right\|^2}.$$

Signal to Distortion Ratio (SDR) measures the total amount of distortion introduced to the original signal, both due to the interfering signal and artifacts introduced by the algorithm. Signal to Interference Ratio (SIR) measures the amount of distortion introduced into the original signal by the interfering signal. Signal to Artifact Ratio (SAR) measures the amount of artifacts introduced into the original signal by the separation algorithm that do not originate in the interfering signal.

Usually some parameters can be chosen to tune the trade-off between interfering signal leakage (SIR) and

the distortion to the desired signal (SAR). For example it is possible to reduce SIR to $-\infty$ simply by zeroing source estimation. However, the SAR measure will become very high in this case. SDR is a combined measure for both SIR and SAR, hence it is convenient to compare the algorithm performance based on SDR.

We use a sub-index to indicate which extracted signal the measure is referring to. For example, SDR_1 refers to the signal-to-distortion ratio of the first extracted component.

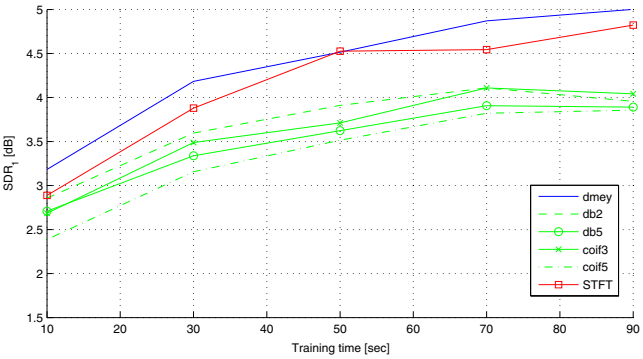


Figure 7 The signal-to-distortion (SDR) ratio of the extracted speech signal based on four test excerpts 90 s each.

7 Experimental Results

In this section, we present some experimental results. We study the algorithm performance on speech utterances mixed with different musical excerpts. We study the effect of the training excerpt length, wavelet family choice, the GMM order and speaker gender on the performance of the algorithm.

We compare the performance of the proposed algorithm to a similar algorithm which is based on the STFT signal analysis [6, 7]. In order for the comparison to the STFT-based algorithm to be fair, we select STFT analysis parameters to match time and frequency resolution of the CSR-BS-WPD at the lowest frequency band. The length of the STFT analysis window is chosen to be 1,024 and the overlap between subsequent frames is 960 samples. We use Hamming synthesis window and its bi-orthogonal vector as the analysis window. We note that the number of expansion coefficients for the STFT transform is roughly three times greater than for the CSR-BS-WPD transform.

We selected four different musical excerpts: two solo piano pieces and two wind quartet pieces. Titles of the musical excerpts and the abbreviations we use in the text are shown in Table 1. A speech signal consists of a single sentence pronounced by different male speakers. All speech utterances were taken from the TIMIT database. While TIMIT database provided speech utterances recorded in the same controlled environment, we did not have access to a similar musical signal repository. We used non-overlapping parts of the same recording for model training and for the mixture in order to overcome this problem.

Table 1 List of musical excerpts used in the experimental results.

Abbreviation	Name	Instrument
W1	L.v. Beethoven, Quartet No. 10 in Eb, Op. 74 - Poco adagio	Wind quartet
W2	L.v. Beethoven, Quartet No. 10 in Eb, Op. 74 - Adagio ma non troppo	Wind quartet
P1	J.S. Bach, Air	Piano solo
P2	J.S. Bach, French Suite No. 4	Piano solo

First, we examined the effect of the training signal length on the separation performance. We trained GMM models using 10 to 90 s of the speech and music excerpts. All signals were sampled at 16 KHz. As a preprocessing step, we normalized the energy of all signals (the normalization is important since the statistical model used is not invariant to signal amplitude. A Gaussian Scaled Mixture Model [6] can be used as a remedy). We evaluated the separation performance using another 90 s of speech and music mixture.

We used the following wavelet families in our experiments: discrete approximation of Meyer wavelet (*dmey*), Daubechies wavelets of order 2 and 5 (*db2, db5*), Coiflets of order 3 and 5 (*coif3, coif5*). Figures 7 and 8 present signal-to-distortion ratio (SDR_1, SDR_2) of the extracted speech and musical signal respectively for different algorithms. These figure present the SDR average of the recovered speech and music signals. High value of SDR indicates a small degree of speech distortion in the extracted component. Other measures

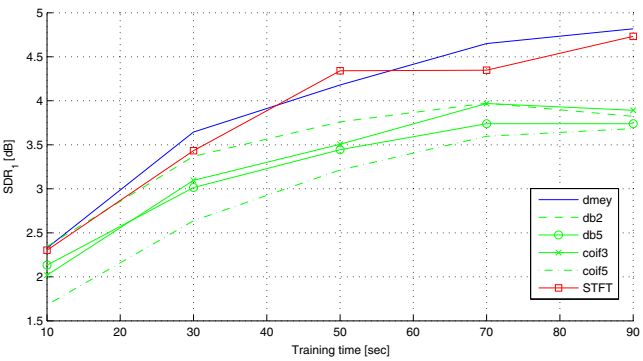
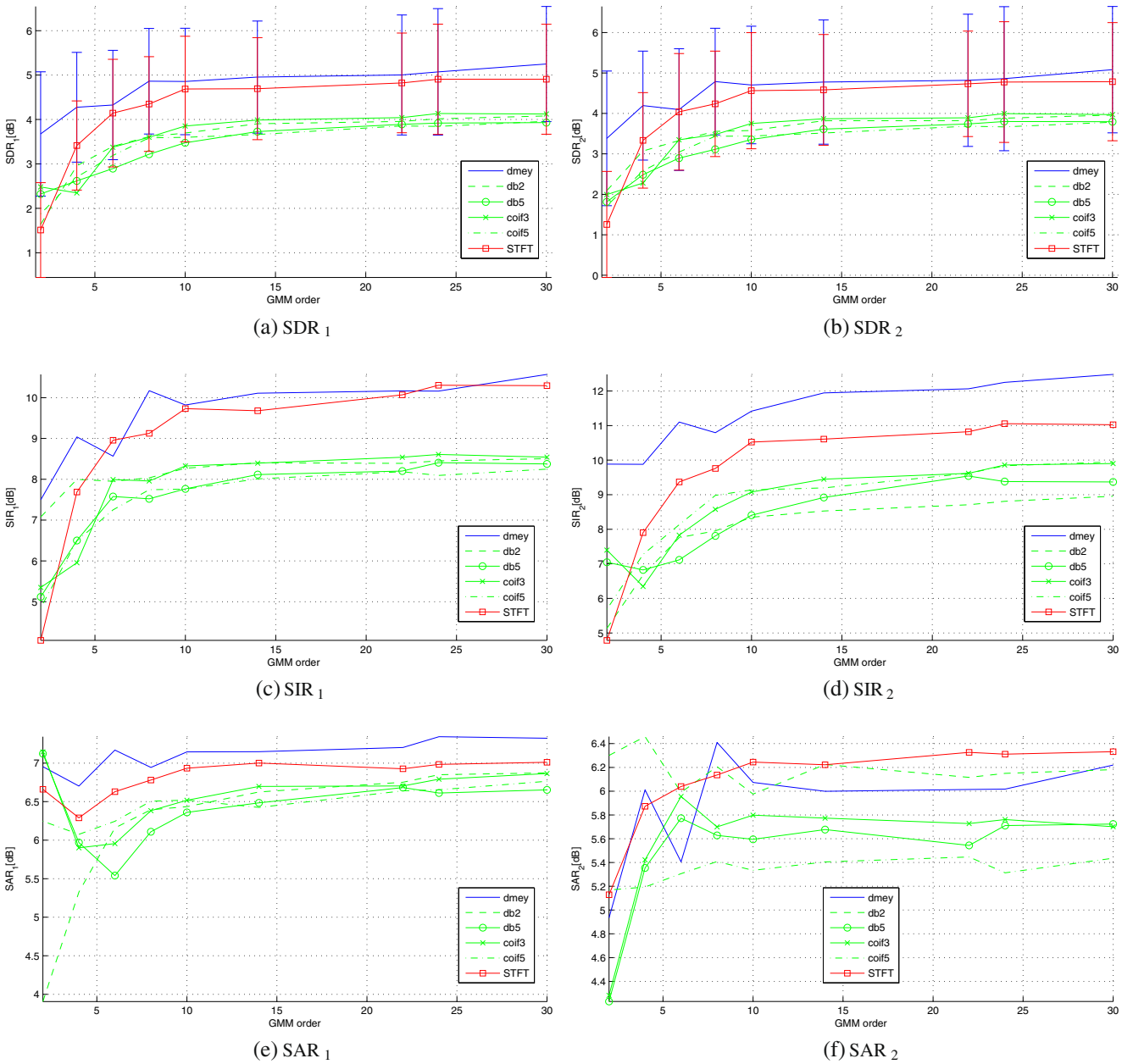


Figure 8 The signal-to-distortion (SDR) ratio of the extracted musical signal based on four test excerpts 90 s each.

exhibit similar performance trend and are not shown here. The SDR values shown in these figure are the average of four test sequences, 90 s each. The Discrete Meyer (*dmey*) based CSR-BS-WPD algorithm demonstrates superior performance compared to the signal analysis based on other wavelet families for all training signal lengths. The performance of the STFT based algorithm is comparable to the *dmey* based algorithm. When the length of the training signal is more than 70 s, the SDR is only slightly improved for the *dmey* and STFT based algorithms and remains the same for the others. In the rest of the experiments we use 90 s long training sequences.

Figure 9 presents the separation performance as a function of GMM model order ($K = 2, \ldots, 30$). Higher order GMM models are useful to describe more complex statistical behaviors at the price of higher computational complexity and larger data sets required for training. Error bars on the SDR plots show confidence intervals of 68% for STFT and *dmey* based CSR-BS-WPD algorithms. The *dmey* based CSR-BS-WPD and STFT exhibit superior performance to other



(a) SDR $_1$ (b) SDR $_2$

(c) SIR $_1$ (d) SIR $_2$

(e) SAR $_1$ (f) SAR $_2$

Figure 9 SDR, SIR and SAR averages for speech (sub-index 1) and music (sub-index 2) recovered signals. All measures are averaged over 6 min of test mixture signals.

wavelet families, under all the performance criteria. A trade-off between SIR and SAR can be seen occasionally, but the SDR measure supports our claim consistently. For a very low GMM orders such as 2 or 4, the STFT based algorithm performs significantly worse than all CSR-BS-WPD based algorithms.

Figures 10 and 11 show spectrograms of the original and extracted speech and piano components. When the trained GMM model does not have a component that describes an observed mixture component well enough, a sequence of frames may be entirely filtered out as can be seen in Figs. 10b and 11b. This kind of an artifact can be found in both STFT and CSR-BS-WPD based

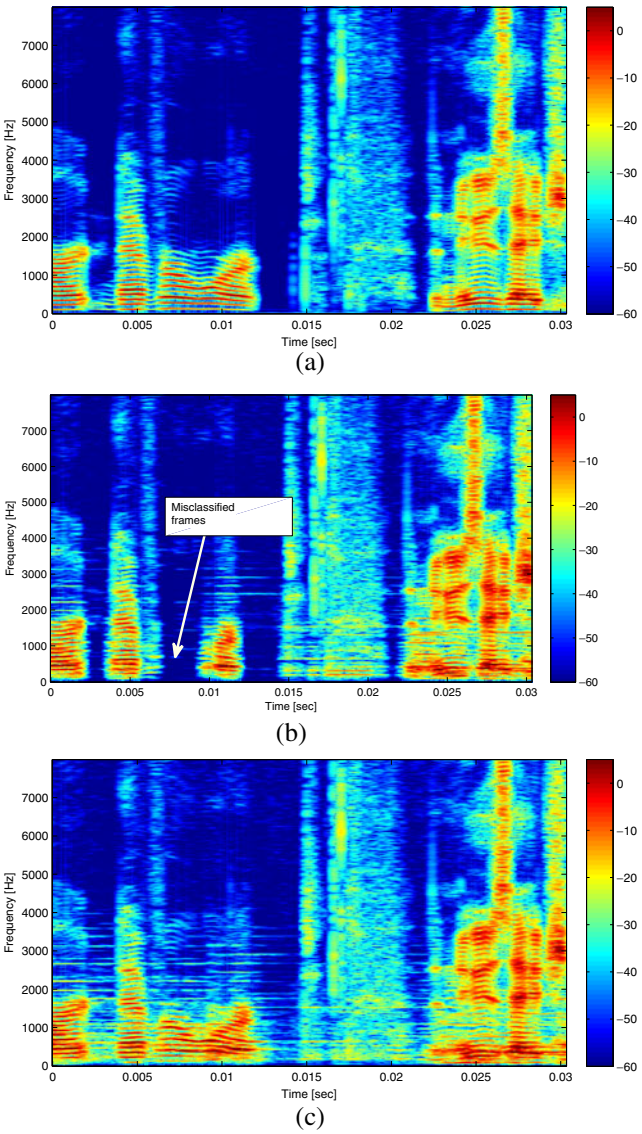


Figure 10 Spectrograms (in dB) of (**a**) speech signal used in the mixture; speech component extracted from the mixture (**b**) using STFT based algorithm; (**c**) using CSR-BS-WPD based algorithm based on the *dmey* wavelet family.

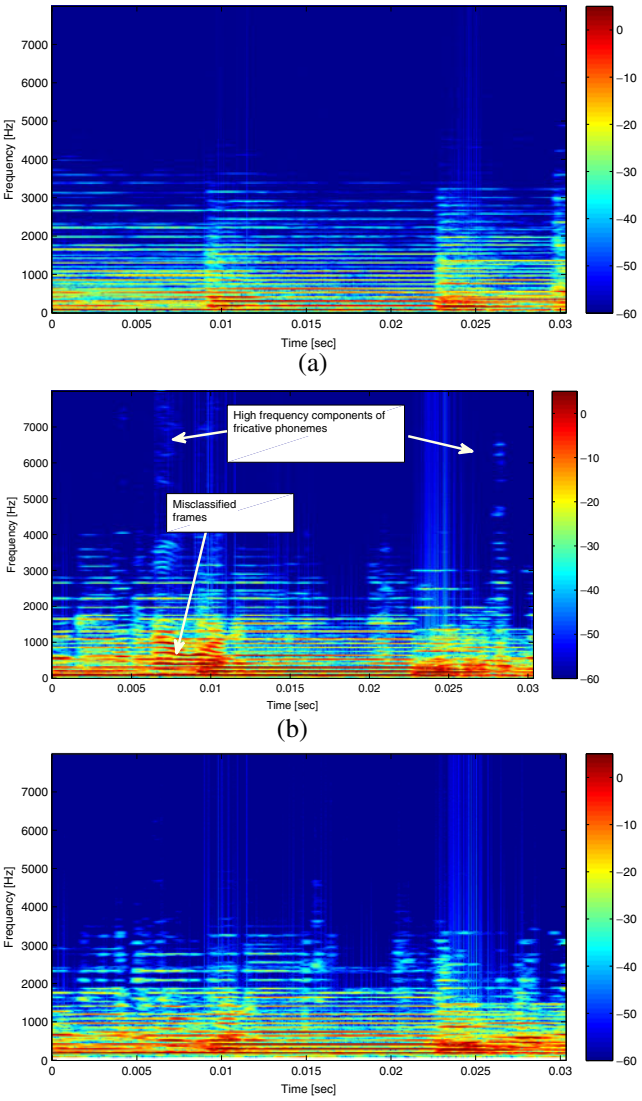


Figure 11 Spectrograms (in dB) of (**a**) piano signal (P1) used in the mixture; piano component extracted from the mixture (**b**) using STFT based algorithm; (**c**) using CSR-BS-WPD based algorithm based on the *dmey* wavelet family.

algorithms, but it is more common to the STFT based algorithm. This can be explained by a better fit of the GMM model due to lower signal dimension. Another artifact observed in Fig. 11b is high-frequency residual of fricative phonemes that leak into the extracted music. This artifact appears mostly in musical signals extracted by the STFT based algorithm.

We verified the algorithm performance on speech excerpts pronounced by male and female speakers. We trained the GMM model on 90 s of speech pronounced only by female speakers. We tested the performance on additional 90 s of a speech and music mixture. We used GMM order of $K = 30$. Table 2 presents experiment

Table 2 Signal-to-distortion ratio of the extracted speech component SDR_1 and music component SDR_2.

Wavelet family	SDR_1		SDR_2	
	Female	Male	Female	Male
coif3	5.27	4.12	5.20	3.97
coif5	5.05	3.96	4.93	3.78
db2	5.22	4.08	5.14	3.96
db5	5.12	3.93	5.05	3.80
dmey	6.19	5.25	6.01	5.08
STFT	5.58	4.91	5.50	4.79

Either male only or female only speech is used in the training and the separation. Order of the GMM model is $K = 30$.

results averaged over all four musical excerpts (P1, P2, W1, W2). The proposed algorithm and the STFT based algorithm perform better when separating female speech from music than separating male speech from music. The SDR measure is approximately 1 dB higher for female speech than for male speech for both algorithms and for all wavelet families. The SIR and SAR performance measures exhibited similar behavior.

Informal listening tests reveal that the subjective quality of the extracted musical signals are comparable for the STFT and the CSR-BS-WPD based algorithms. The extracted speech signal on the other hand, has slightly more pleasant sound when recovered by the proposed algorithm. A relatively large amount of misclassified frames, similar to those shown in Fig. 11b results in an unnatural speech sound and reduces the subjective signal quality.

Audio files with mixtures used in the experiments and the extracted signals can be downloaded from http://sipl.technion.ac.il/~elitvin/CSR-BS-WPD.

8 Conclusions and Future Work

We have described how a Bark-scaled WPD can be adapted to the source separation task. Introduction of a different subsampling scheme and approximate shiftability to the psychoacoustically motivated BS-WPD decomposition tree, enabled us to adapt a source separation algorithm that was originally designed to work in the uniformly sampled STFT domain.

We found several advantages of using the proposed analysis together with a GMM based source separation algorithm. The most significant advantage is the reduction of the dimension of the signal space at the expense of high frequency resolution. As a result, the computational burden and the amount of trained parameters of the GMM model are reduced.

We found that a choice of wavelet family used with the proposed separation algorithm is crucial. The discrete Meyer wavelet based CSR-BS-WPD analysis yielded improved separation performance compared to the STFT based analysis with very low orders of GMM model. Other wavelet families failed to produce performance comparable to the STFT based algorithm.

Cohen [10] used *dmey* wavelet family for speech enhancement. He justified the advantage of the *dmey* wavelet family by the regularity of the wavelet filter and its good frequency localization properties. In Section 4 we showed that superior frequency localization properties reduce correlation between expansion coefficients, hence improve fit to our separation model. Due to the similarity of our separation algorithm to the STFT based algorithm, the performance and pitfalls of both algorithms are comparable.

Future work with the CSR-BS-WPD analysis may address various audio processing tasks traditionally performed in the STFT domain, which require instantaneous spectral shapes and where the psychoacoustically motivated band structure of the time-frequency analysis might be desirable.

References

1. Vincent, E., Févotte, C., Benaroya, L., & Gribonval, R. (2003). A tentative typology of audio source separation tasks. In *Proc. 4th international symposium on independent component analysis and blind signal separation (ICA2003)* (pp. 715–720). Nara, Japan.
2. Cherry, C. E. (1953). Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 25*(5), 975–979.
3. Comon P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*(3), 287–314.
4. Hyvärinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. Wiley-Interscience.
5. Ozerov, A., Philippe, P., Bimbot, F., & Gribonval, R. (2007). Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs. *IEEE Transactions on Audio, Speech & Language Processing, 15*(5), 1564–1578.
6. Benaroya, L., Bimbot, F., & Gribonval, R. (2006). Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech & Language Processing, 14*(1), 191–199.
7. Benaroya, L., & Bimbot, F. (2003). Wiener based source separation with HMM/GMM using a single sensor. In *Proc. 4th international symposium on independent component analysis and blind signal separation (ICA2003)* (pp. 957–961). Nara, Japan.
8. Srinivasan, S., Samuelsson, J., & Kleijn, W. B. (2006). Codebook driven short-term predictor parameter estimation for speech enhancement. *IEEE Transactions on Audio, Speech & Language Processing, 14*(1), 163–176.
9. Srinivasan, S., Samuelsson, J., & Kleijn, W. B. (2007). Codebook-based bayesian speech enhancement for nonstationary environments. *IEEE Transactions on Audio, Speech & Language Processing, 15*(2), 441–452.

10. Cohen, I. (2001). Enhancement of speech using bark-scaled wavelet packet decomposition. In *Proc. 7th European conf. speech, communication and technology, EUROSPEECH-2001* (pp. 1933–1936). Aalborg, Denmark.
11. Fernandes, F. C .A., van Spaendonck, R. L. C., & Burrus, C. S. (2003). A new framework for complex wavelet transforms. *IEEE Transactions Signal Processing, 51*(7), 1825–1837.
12. Litvin, Y., & Cohen, I. (2009). Single-channel source separation of audio signals using bark scale wavelet packet decomposition. In *2009 IEEE international workshop on machine learning for signal processing (MLSP09)*.
13. Fernandes, F. C. A. (2002). *Directional, shift-insensitive, complex wavelet transforms with controllable redundancy*. Ph.D. thesis, Rice Univ., Houston, TX, USA.
14. Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multiscale transforms. *IEEE Transactions on Information Theory, 38*(2), 587–607.
15. Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(6), 1109–1121.
16. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1), 1–38.
17. Gribonval, R., Benaroya, L., Vincent, E., & Févotte, C. (2003). Proposals for performance measurement in source separation. In *Proc. 4th international symposium on ICA and BSS (ICA2003)* (pp. 763–768). Nara, Japan.
18. Févotte, C., Gribonval, R., & Vincent, E. (2005). BSS_EVAL toolbox user guide revision 2.0. Tech. Rep. 1706, IRISA, Rennes, France.

Yevgeni Litvin received the B.Sc. (Summa Cum Laude) and M.Sc. degrees in Electrical Engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 2007 and 2010, respectively.

From 1996 to 2003, he was a Software Engineer with Israeli Defense Forces. From 2008 to 2009, he was a Teaching Assistant and a Project Supervisor with the Signal and Image Processing Lab (SIPL), Electrical Engineering Department, the Technion. Since 2009, he has been an Algorithm Engineer with Zoran.

His research interests are statistical signal processing, speech enhancement, statistical machine learning and video processing.

Israel Cohen (M'01-SM'03) received the B.Sc. (Summa Cum Laude), M.Sc. and Ph.D. degrees in Electrical Engineering from the Technion – Israel Institute of Technology, Haifa, Israel, in 1990, 1993 and 1998, respectively.

From 1990 to 1998, he was a Research Scientist with RAFAEL Research Laboratories, Haifa, Israel Ministry of Defense. From 1998 to 2001, he was a Postdoctoral Research Associate with the Computer Science Department, Yale University, New Haven, CT. In 2001 he joined the Electrical Engineering Department of the Technion, where he is currently an Associate Professor. His research interests are statistical signal processing, analysis and modeling of acoustic signals, speech enhancement, noise estimation, microphone arrays, source localization, blind source separation, system identification and adaptive filtering. He served as Guest Editor of a special issue of the *EURASIP Journal on Advances in Signal Processing* on Advances in Multimicrophone Speech Processing and a special issue of the *EURASIP Speech Communication Journal* on Speech Enhancement. He is a coeditor of the Multichannel Speech Processing section of the *Springer Handbook of Speech Processing* (Springer, 2007), a coauthor of *Noise Reduction in Speech Processing* (Springer, 2009), and a cochair of the 2010 International Workshop on Acoustic Echo and Noise Control.

Dr. Cohen received in 2005 and 2006 the Technion Excellent Lecturer awards, and in 2009 the Muriel and David Jacknow award for Excellence in Teaching. He served as Associate Editor of the IEEE Transactions on Audio, Speech, and Language Processing and IEEE Signal Processing Letters.