Signal Processing 90 (2010) 3147–3163

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Monaural speech/music source separation using discrete energy separation algorithm[☆]

Yevgeni Litvin[a], Israel Cohen[a,*], Dan Chazan[b]

[a] *Department of Electrical Engineering, Technion - Israel Institute of Technology, Technion City, Haifa 32000, Israel*
[b] *IBM Research Laboratory in Haifa, Israel*

ARTICLE INFO

Article history:
Received 15 August 2009
Received in revised form
14 May 2010
Accepted 17 May 2010
Available online 26 May 2010

Keywords:
Single-channel source separation
Instantaneous frequency
Time–frequency masking
Frequency modulation
DESA

ABSTRACT

In this paper, we address the problem of monaural source separation of a mixed signal containing speech and music components. We use Discrete Energy Separation Algorithm (DESA) to estimate frequency-modulating (FM) signal energy. The FM signal energy is used to design a time-varying filter in the time–frequency domain for rejecting the interfering signal. The FM signal energy was chosen due to its good ability to differentiate between speech and music signals using localized information both in time and frequency. We present experimental results which demonstrate the advantages and limitations of the proposed method using synthetic data and real audio signals.

1. Introduction

Blind source separation (BSS) of audio signals has been an active area of research in recent years. BSS from a single audio channel is a special case of general BSS problem where data from only one sensor is available to the algorithm. This problem is generally manageable when the separated audio signals belong to different signal classes, which are distinguishable based on prior knowledge.

Different attempts to solve this problem in various contexts were made, including: statistical modeling, such as Gaussian Mixture Model (GMM) [1], or Hidden Markov Model (HMM) [2,3]; computational auditory scene analysis (CASA) [4]; non-negative matrix factorization (NMF) [5]; sparse decomposition [6] and others. Single-audio-channel BSS is an under-determined problem with arbitrary many solutions, so some prior knowledge is required to perform the separation. In order to fill this gap, various methods use either perceptual principles or statistical models. Although many existing solutions produce satisfactory results in special cases, the general problem of single-audio-channel BSS remains unsolved.

Teager and Teager [7,8] studied airflow and fluid dynamics of human speech apparatus, and described several nonlinear phenomena as well as their sources. Later, Kaiser [9,10] formulated the Teager energy operator (TEO). In [10–12] the TEO was used to derive a discrete energy separation algorithm (DESA) that separates a signal into its amplitude (AM) and frequency modulating (FM) components. Applications of the AM–FM decomposition of audio signals include formant tracking [13], extraction of speech recognition and speaker recognition features [14–17], speech coding [18], and analysis and re-synthesis of musical instruments sound [19].

Sinusoidal modeling was previously used for BSS by Virtanen and Klapuri [20]. Their approach requires peak tracking in the spectral domain to establish sinusoidal trajectories followed by grouping of detected trajectories into different audio streams. Although, our approach can also be viewed as a kind of sinusoidal modeling, it does

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

[*] Corresponding author. Tel.: +972 4 8294731; fax: +972 4 8295757.
E-mail addresses: elitvin@tx.technion.ac.il (Y. Litvin), icohen@ee.technion.ac.il (I. Cohen).

0165-1684/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2010.05.020

not require peak tracking or grouping, which may improve the robustness of the separation algorithm. Yilmaz et al. [21] defined approximate W-disjoint orthogonality (W-DO) as an approximate "disjointness" of several signals in the short-time Fourier transform (STFT) domain. They introduced a quantitative W-DO measure and provided evidence of the high level of the W-DO for two speech signals.

Analysis of the AM component of the subband AM–FM decomposition was previously used for monaural source separation by Atlas and Janssen in [22]. The authors filtered a subband amplitude modulating signal using a filter that was learned from training data. The segregated components were reconstructed by combining the filtered AM signal with the original carrier.

Disch and Edler [23] used adaptive subband AM–FM signal decomposition. The signal is first decomposed into a set of analytical bandpass signals. The center frequencies of bandpass filtering are chosen adaptively and are aligned with local centers of gravity. The AM–FM decomposition is performed at each subband and AM and FM components are processed separately. The application demonstrated in [23] is polyphonic key mode change, but monaural source separation also seems possible using the proposed framework.

Recently we proposed a source separation algorithm that separates piano play signals and speech signals from a single channel [24]. The algorithm used subband FM analysis to distinguish between different signal classes. Preliminary experiments showed promising results. In this work, we propose a source separation algorithm that segregates audio sources from a single channel. Different signal classes may posses different statistical properties of subband FM components. The proposed algorithm uses these differences to separate sources. Like [22], our algorithm uses AM–FM analysis, but we rely on the properties of the FM and not the AM signal to differentiate between audio signal classes. First we filter the input signal by a STFT filterbank. Then we use the DESA to estimate a frequency modulating signal in each of the filterbank outputs and the energy of the frequency modulating signal (EFMS). In the training stage a statistical model of the EFMS values is learned for each signal class. In the separation stage, time–frequency bins in the STFT domain are classified into one of the target signal classes using EFMS values. The interfering signal is suppressed by zeroing time–frequency bins attributed to the interfering signal. Finally, we reconstruct the separated component by inverting the STFT.

The signal classes are assumed to differ in subband frequency-modulating component statistics, hence our algorithm is not capable of separating two signals from the same class (e.g. a mixture of two speech excerpts). We show good performance in separating speech from several types of music, assuming that training signals from both classes are available to the algorithm for model training. Only a 2-class problem is addressed in this work. We investigate the impact of music type on the quality of the separation and study the reasons for differences in performance.

Not every musical instrument can be perfectly separated from speech using the proposed method. Percussive musical instruments cannot be separated from speech due to the non-harmonic structure of their sound. Musical instruments with strong non-harmonic components are separated poorly. For this reason, in some applications, where the music has mostly harmonic components, the proposed algorithm may be used as a stand-alone method. In others, where music has a large amount of non-harmonic components, such as music from percussive instruments, the proposed method may be combined with existing techniques of monaural source separation to improve the overall performance. The proposed method can also be used in some scenarios of speech enhancement where the noise has harmonic nature.

The main contribution of this paper is to show that by using the subband FM component statistics, well localized time–frequency regions can be reliably assigned to the correct source. We present experimental results that demonstrate feasibility of our approach both on synthetic signals, real speech, and musical signals. For simplicity, our study is constrained to experiments on instantaneous mixtures without noise presence. The performance is compared to a competitive source separation algorithm.

The remainder of this paper is structured as follows. In Section 2, we describe the TEO and the DESA used for the AM–FM analysis. In Section 3, we present some real audio signal examples and explain why the proposed method should perform well in the separation task. Section 4 describes a simple training procedure used to learn EFMS features and a Bayesian approach used for the creation of an STFT domain binary mask. Section 5 presents standard separation performance evaluation criteria which are later used in Section 6 to evaluate the performance.

2. Discrete energy separation algorithm

In this section, we introduce mathematical notations and define AM–FM analysis using TEO (DESA [10]). Let $x_c(t)$ be a continuous time signal and $x(n)$=$x_c(nT)$ be its sampled version with sampling period of T. We assume the following signal model

$$x(n) = a(n)\cos\left(\Omega_c n + \sum_{i=0}^{n} q(i)\frac{1}{T} + \theta\right) \tag{1}$$

where n is a discrete time index, Ω_c is an angular frequency of a carrier, θ is some constant phase value, and $a(n)$ and $q(n)$ are the amplitude and frequency modulating signals, respectively.

A discrete version of TEO uses the function $\Psi[x(n)]$, which is defined as follows:

$$\Psi[x(n)] = x^2(n) - x(n-1)x(n+1) \tag{2}$$

The instantaneous frequency of a continuous signal is defined by $\Omega_i \triangleq d/dt\angle x(t)$. $\Psi[x(n)]$ is used for estimating the instantaneous frequency $\hat{\Omega}_i(n)$ and the instantaneous amplitude $\hat{a}(n)$:

$$\hat{\Omega}_i(n) \approx \frac{1}{2}\arccos\left(1 - \frac{\Psi[x(n+1) - x(n-1)]}{2\Psi[x(n)]}\right) \tag{3}$$

$$\approx \Omega_c + q(n) \tag{4}$$

$$|\hat{a}(n)| \approx \frac{2\Psi[x(n)]}{\sqrt{\Psi[x(n+1)-x(n-1)]}} \tag{5}$$

These approximations are valid if the following conditions hold:

$$\Omega_a \ll \Omega_c \quad \text{and} \quad \kappa \ll 1 \tag{6}$$

$$\Omega_f \ll \Omega_c \quad \text{and} \quad \frac{\sup\{q(n)\}}{\Omega_c} \ll 1 \tag{7}$$

where Ω_a and Ω_f are the highest non-zero angular frequencies of $a(n)$ and $q(n)$, respectively, and κ is an AM modulation index ($a(n)$ assumed to be positive). In [10] this version of DESA algorithm is called DESA-2.

3. Motivation for analysis in frequency modulation domain

In this section, we demonstrate frequency modulation analysis on some examples of speech and piano signals. We define the energy of the frequency modulating signal (EFMS) and show that EFMS of speech and piano signals can be used as a local time–frequency discriminating factor and used for rejecting the interfering source. These examples will motivate the formulation of our algorithm.

Harmonic signals, such as vowels in speech or musical notes played by a harmonic musical instrument, contain harmonic partials, which are sine signal components located at integer multiples of the fundamental frequency. Partials of voiced phonemes in speech signals have a stronger frequency modulating component than partials of piano signals. Unvoiced phonemes, such as plosive and fricative phonemes, do not contain harmonic partials. An AM–FM decomposition of unvoiced phoneme subbands produces a noisy FM component with stronger frequency modulating component than the AM–FM decomposition of voiced phonemes. To define an algorithm that exploits this property we need to formulate a quantitative measure for this phenomenon. Let $x(n)$ denote a time signal. We assume $x(n)$ is an harmonic signal with one or more harmonic partials. We treat each partial as a separate carrier. Most of the AM–FM demodulation algorithms, including DESA, cannot deal with multiple carriers in the analyzed signal. To apply the analysis we note that each of the signals produced by filtering the analyzed signal with a narrow band filterbank likely contains a single AM–FM modulated carrier. In our work we use STFT filterbank.

Let $X_k(m)$ be the STFT transform of $x(n)$, where k and m are frequency and time indices. The STFT transform is given by

$$X_k(m) = \sum_{n=-\infty}^{\infty} w(mM-n)x(n)e^{-j(2\pi/N)kn} \tag{8}$$

where N and M define the frequency and time resolution of the transform, and $w(n)$ is the analysis window with support of N samples and angular bandwidth of b_w. Eq. (8) can be rewritten in a filter like form as

$$X_k(m) = e^{-j(2\pi/N)mM}(x * w_a)(mM) \tag{9}$$

where $w_a(n)$ is an analytic bandpass filter generated by shifting $w(n)$ in frequency by $2\pi k/N$ radians.

The time series $X_k(m)$ indexed by m, can be treated as a time domain bandpass version of the analytic signal of $x(n)$ with bandpass center frequency shifted to zero. We assume that only a single partial is present in $X_k(m)$. This allows us to use AM–FM decomposition algorithm. In the AM–FM decomposition, each harmonic partial will act as a carrier. Instantaneous deviations from the carrier frequency (caused by intonation in speech and speech production nonlinearities) will appear as a frequency-modulating signal.

3.1. EFMS calculation

Assume the AM–FM model for the l-th harmonic partial

$$x_l(n) = a(n)\cos\left(\Omega_c n + \sum_{i=0}^{n} q(i)\frac{1}{T} + \theta\right) \tag{10}$$

Let b_x be the angular bandwidth of $x_l(n)$.

Assume that almost all the energy of $x_l(n)$ resides in the k-th subband of the STFT filterbank. This results in the approximation

$$x_a(n) \approx x_l(n) * w_a(n) \tag{11}$$

where $x_a(n)$ is an analytic signal of $x_l(n)$. Eq. (11) holds only approximately since the theoretical bandwidth of a frequency modulated signal is infinite. We can also write

$$\Omega_c + \frac{b_x}{2} < \frac{2\pi}{N}k + \frac{b_w}{2} \cap \Omega_c - \frac{b_x}{2} > \frac{2\pi}{N}k - \frac{b_w}{2} \tag{12}$$

$$\left|\frac{2\pi}{N}k - \Omega_c\right| < \frac{b_w - b_x}{2} \tag{13}$$

After modulating $x_a(n)$ by a complex exponent $e^{-j2\pi m/N}$ and decimation by a factor of M, the output of the STFT filterbank (9) is given by

$$X_k(m) \approx a(mM)\exp j\left(\tilde{\Omega}_c mM + \sum_{i=0}^{mM} q(i)\frac{1}{T} + \theta\right) \tag{14}$$

where

$$\tilde{\Omega}_c = \Omega_c - \frac{2\pi}{N}k \tag{15}$$

and from (13) we have $|\tilde{\Omega}_c| < b_w/2$. The angular bandwidths of $a(n)$ and $q(n)$ grow by a factor of M, and $\tilde{\Omega}_c$ is close to zero. Therefore, the DESA assumptions $\Omega_f \ll \tilde{\Omega}_c, \Omega_a \ll \tilde{\Omega}_c$ no longer hold. The remedy is to modulate the filterbank output to some intermediate frequency Ω_{if} by multiplying $X_k(m)$ by $e^{j\Omega_{\text{if}} m}$.

We choose $\Omega_{\text{if}} = \pi/3$ (shift $X_k(m)$ by $\pi/3$ (rad/s)) i.e. we set a new carrier frequency to be in the lower third of the frequency axis so as to minimize the risk of aliasing (although $\Omega_{\text{if}} = \pi/2$ would be a more appropriate choice, the experimental results showed better performance when $\Omega_{\text{if}} = \pi/3$ is used). DESA operates on the real valued signals, we use only the in-phase component of the modulated filterbank output

$$\tilde{X}_k(n) = \Re(X_k(n)e^{j\Omega_{\text{if}} n}) \tag{16}$$

To avoid aliasing during modulation and in-phase component extraction the following conditions must hold:

$$\Omega_{\text{if}} > \frac{b_x}{2}M \tag{17}$$

$$\Omega_{\text{if}} < \pi - \frac{b_x}{2}M \tag{18}$$

Both conditions can be satisfied by choosing a sufficiently small M. It can be shown that if $\Omega_{\text{if}} = \alpha\pi$, $\alpha \in [0,1]$ then the value of M must satisfy $M \le \min\{\alpha N, (1-\alpha)N\}$.

Fig. 1 shows an example of the processing steps. A synthetic harmonic signal with 10 harmonic partials is used in this example. The first partial is an FM modulated signal. The FM modulating signal is a sinusoid having an amplitude of 2π and a frequency of 10 Hz. The Fourier transform of the signal is shown in Fig. 1(a). Most of the energy of the first partial is located in the 21-st band. The Fourier transform of $X_{21}(m)$ is shown in Fig. 1(b). $X_{21}(m)$ is a complex signal, hence the magnitude values of the Fourier transform are not symmetric. Fig. 1(c) shows the Fourier transform of $\tilde{X}_{21}(n)$. $\tilde{X}_{21}(n)$ is a real valued signal modulated to the intermediate frequency. The dashed line shows regions of the spectrum originally filtered out by w_a.

DESA estimator can now be used to find the FM component of $\tilde{X}_k(n)$

$$\hat{\Omega}_{i,k}(n) \approx \frac{1}{2}\arccos\left(1-\frac{\Psi[\tilde{X}_k(n+1)-\tilde{X}_k(n-1)]}{2\Psi[\tilde{X}_k(n)]}\right) \tag{19}$$

The instantaneous frequency $\hat{\Omega}_i$ also includes a constant $\tilde{\Omega}_c$ term. To remove it we filter $\hat{\Omega}_i$ with a high-pass filter h_q which results in an estimate of $q(n)$. Note that Ω_c is not necessarily constant in time, but we assume that it changes slowly compared to $q(n)$,

$$\hat{q}(n) \approx (\hat{\Omega}_i * h_q)(n) \tag{20}$$

$$\approx ((\tilde{\Omega}_c + \Omega_{\text{if}} + q(n)) * h_q)(n) \tag{21}$$

Define the EFMS by

$$\hat{E}_k(m) \triangleq (u * \hat{q}_k^2)(m) \tag{22}$$

where $u(n)$ is an N_u points Hamming window designed to reduce the variance of the energy estimator $\hat{q}_k^2(m)$. In the rest of the paper we denote the EFMS of a time signal $x(n)$ by $\hat{E}\{x\}_k(m)$ and omit x and the indices k and m when the meaning is clear from the context.

Fig. 2 shows a speech fragment containing the utterance "don't ask me to carry". The upper pane shows the 50 lower frequency bands of the STFT filterbank output. First six harmonic partials are visible. We manually pick the 16-th frequency band which contains the second harmonic partial for some period of time. The second pane shows amplitude envelope $\hat{a}_{16}(m)$ of the selected frequency band estimated by the DESA. There are several amplitude peaks corresponding to voiced phonemes. The third pane shows the $\hat{\Omega}_{i,16}$ estimate. The lowest pane shows EFMS $\hat{E}_{16}(m)$ values. In

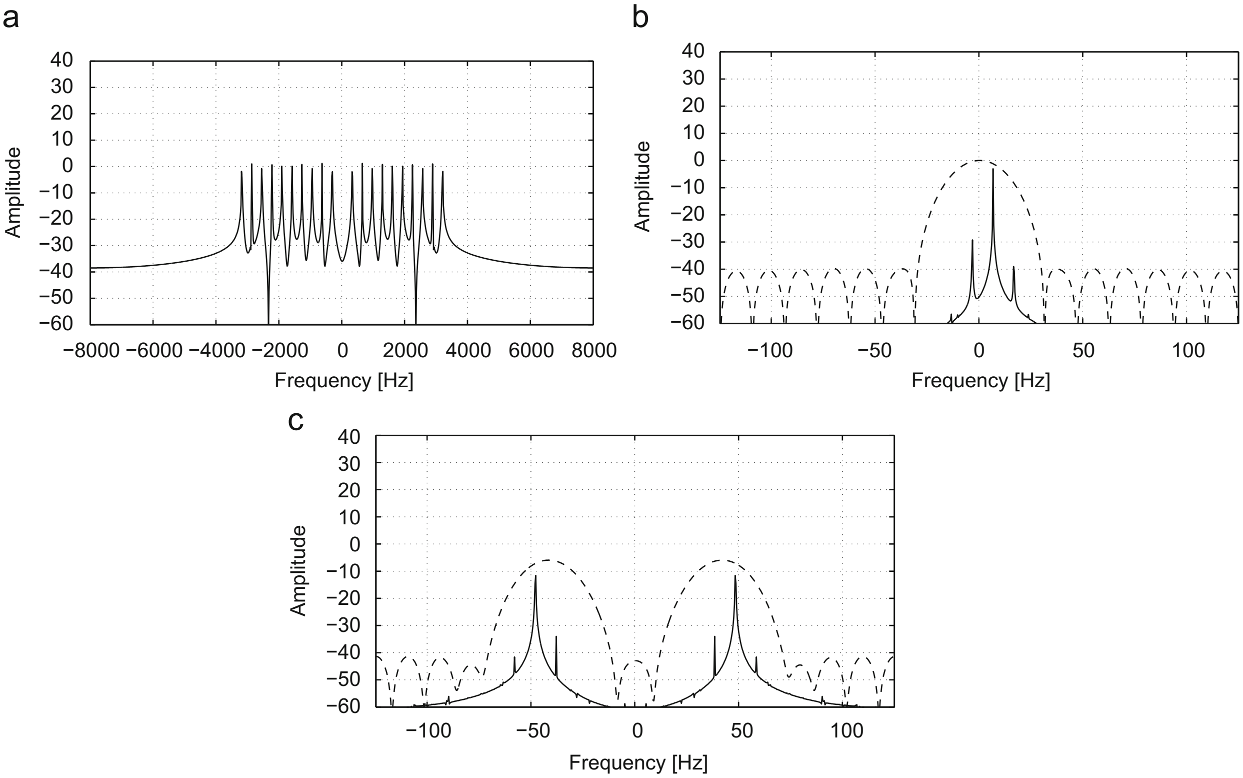

Fig. 1. Input signal preprocessing for the DESA. A dashed line shows which portions of the spectrum were originally filtered out by $w_a(n)$. (a) Input signal that contains 10 carriers; (b) frequency domain representation of the signal at the STFT filterbank output ($X_k(m)$); (c) STFT filterbank output modulated to the intermediate frequency ($\tilde{X}_k(n)$).

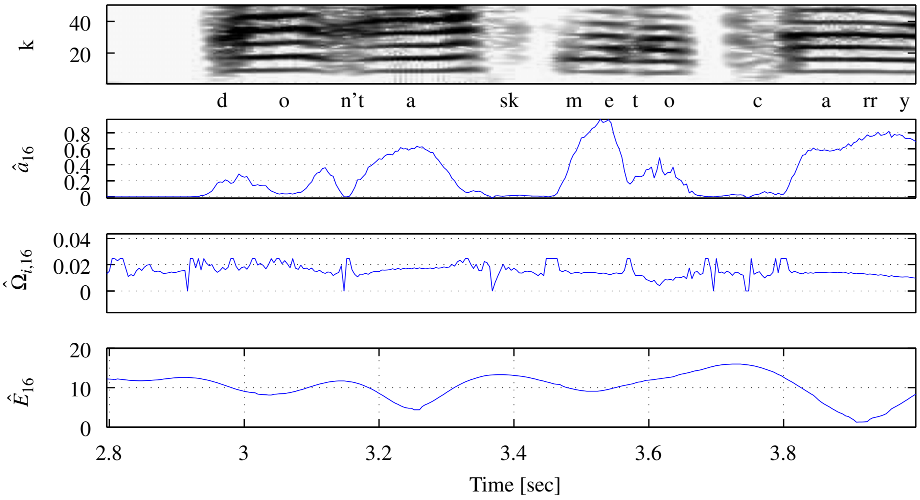

Fig. 2. Upper pane shows the spectrogram (50 lower frequency bands) of the "don't ask me to carry" utterance. Vertical axis labels show frequency band numbers. Second pane shows the estimated AM component of the 16-th frequency band ($\hat{a}_{16}$). Third pane shows the instantaneous frequency estimation $\hat{\Omega}_{i,16}$ of the 16-th frequency band. Lower pane shows the EFMS ($\hat{E}_{16}(n)$).

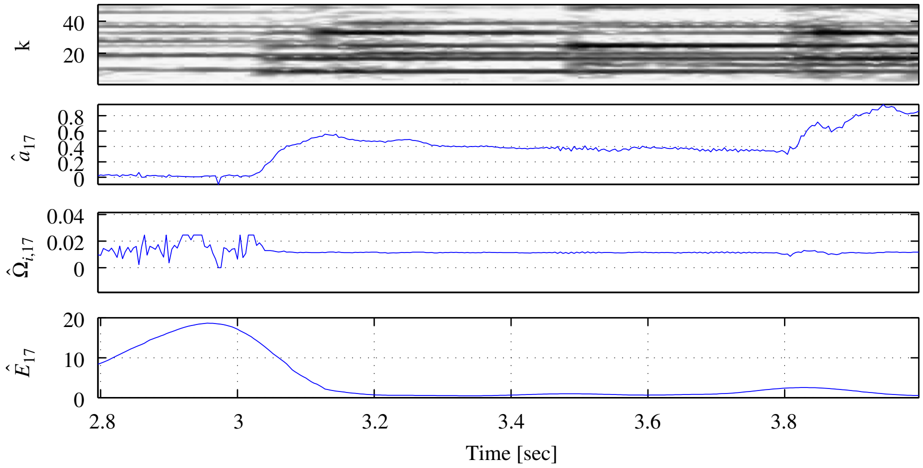

Fig. 3. Upper pane shows the spectrogram (50 lower frequency bands) of the piano play sample. Vertical axis labels show frequency band numbers. Second pane shows the estimated AM component of the 17-th frequency band ($\hat{a}_{17}$). Third pane shows the instantaneous frequency estimation $\hat{\Omega}_{i,17}$ of the 17-th frequency band. Lower pane shows the EFMS ($\hat{E}_{17}(n)$).

the voiced parts of the speech fragment the energy of the FM component is low. Unvoiced phonemes are not described well by the AM–FM model. The DESA estimate of the instantaneous frequency has high variance at these time–frequency locations. As a result, the values of EFMS at the location of unvoiced phonemes are high. This observation is consistent with our claim that the EFMS of a speech signal is higher than the EFMS of a piano play.

The piano play fragment depicted in Fig. 3 contains several piano notes. As in the previous case, we manually pick a frequency band that contains a single harmonic partial. We take the 17-th band and perform the same analysis. We observe that $\hat{E}_{17}(m)$ values are low while the note is being played, hence we have an evidence that a piano produces audio signals with low EFMS. From the examination of Figs. 2 and 3 we conclude that in order to discriminate between speech and piano signals using amplitude modulating envelope it is necessary to define some non-trivial model that would describe different onset and decay behaviors of these signals. On the other hand, two signal classes can be easily distinguished by examining a one-dimensional value of the EFMS.

The width of the STFT bands as well as the DESA assumptions (6) and (7) imply some limitations on the bandwidth of the AM–FM signal components. Despite these limitations, our algorithm is not too sensitive to violations of these conditions. Our basic assumption is that a speech signal likely has an FM component with higher bandwidth and energy than music signal. If the bandwidth of the FM signal is too high for (7) to hold, the DESA produces an inaccurate estimate of the FM component. This estimate has characteristics of high energy noise, hence the time–frequency regions containing a misbehaving signal will be classified as speech. This is consistent with the desired outcome.

In the next example, we apply EFMS analysis to synthetic signals: an harmonic signal (x_1) and white noise with unit variance (x_2). The harmonic signal has fundamental frequency f_0=250 Hz and N_p=30 partials. Let p denote the index of a partial. The carrier frequency and the amplitude of the frequency modulating signal of p-th partial are $f_0 \cdot p$ and $A_0 \cdot p$. Both grow linearly with the index of the partial, like in speech or music signals. The frequency f_{FM} of the FM component is fixed f_{FM}=10 Hz:

$$x_1(n) = \sum_{p=1}^{N_p} x_{1,p}(n) \tag{23}$$

$$x_{1,p}(n) = \cos\left(2\pi f_0 pn + \sum_{i=0}^{n} q_p(n)\frac{1}{T}\right) \tag{24}$$

$$q_p(n) = 2\pi A_0 p \cos(2\pi f_{FM} n) \tag{25}$$

Fig. 4 shows the distribution of the EFMS values for every value of frequency (only values of EFMS that are located at time–frequency bins that have high energy participate in this analysis. The exact method for selecting these frequency bins is described in Section 4.1.). The amplitude of the FM signal grows linearly with the index of the partial. In the case of a sinusoidal signal, the square root of signal energy is proportional to its amplitude. The dashed line in Fig. 4 shows theoretically predicted values of $\sqrt{\hat{E}}$. It is given by $(A_0/\sqrt{2}f_0)f$. Actual values of $\sqrt{\hat{E}}$ are located in the vicinity of the theoretically predicted values, but not exactly on it. There are several reasons for the mismatch:

- Bandpass filtering of a frequency modulated signal alters its sidebands. This results in distortion of the FM modulating signal. This is especially true for high frequency partials: their bandwidth is relatively high

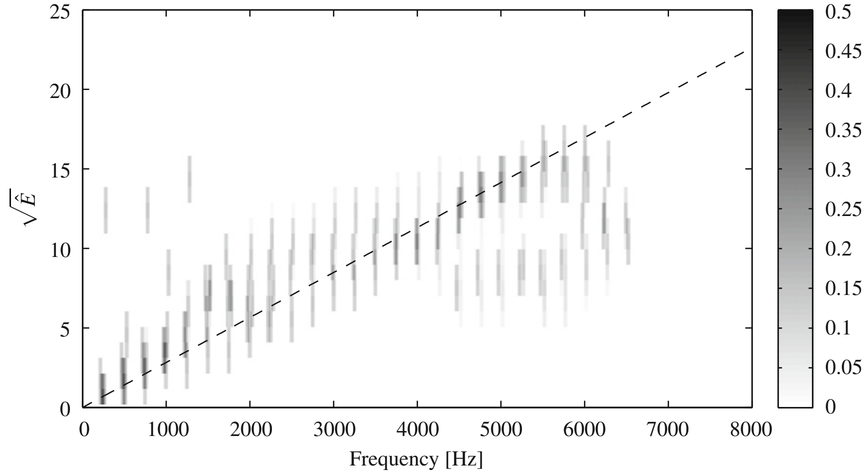

Fig. 4. Distribution of the EFMS values for the synthetic signal (x_1) having 30 partials with linearly increasing amplitude of frequency modulating component. A dashed line shows theoretically predicted values.

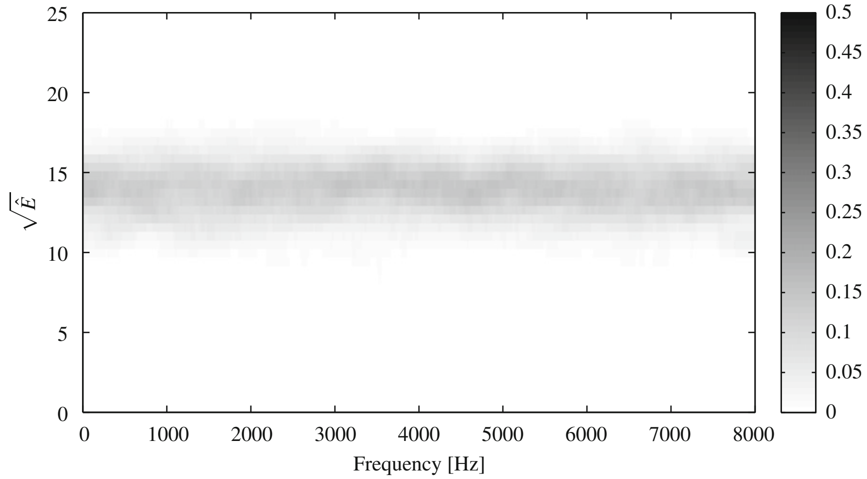

Fig. 5. Distribution of the EFMS values for white noise (x_2).

due to the high amplitude of the modulating signal.

- Partials that ''leak'' to neighboring bands have low SNR levels and result in EFMS estimates similar to EFMS of white noise.

White noise signal is not described well by the AM–FM model. Hence, the EFMS values in all frequency bands are distributed randomly around some constant value as can be seen in Fig. 5.

Figs. 6 and 7 show the EFMS distributions for speech and piano signals, respectively. The EFMS analysis of speech resembles white noise for frequency greater than 500 Hz. Smaller values of EFMS are present under 500 Hz but nevertheless they are generally higher than EFMS values of a piano play. The piano play has relatively low values of EFMS that increase gradually with frequency. Harmonic signal model predicts linear growth of $\sqrt{\hat{E}}$. Fig. 7 shows rough resemblance to a linear growth.

The training part of the separation algorithm presented in the next section, yields an estimate of an empirical probability density function (pdf) independent of the frequency band. This is a simplified training procedure since we saw in Figs. 6 and 7 that the EFMS distribution depends on the frequency. Despite that, we found that this approach is sufficient for our purposes.

Figs. 8–12 show an empirical distribution of $\sqrt{\hat{E}}$. Each figure compares the empirical pdf of speech and music of different types. More specifically, wind quartet, piano, piano–brass duet, guitar and orchestra. Smaller overlap between music and speech pdf means smaller theoretical classification error. We observe that the overlapping area is smaller for wind quartet, piano and guitar (Figs. 8–10) compared to piano–brass duet and orchestra (Figs. 11 and 12). This observation allows us to predict that the former mixtures can be separated better than the later. We confirm this prediction by the experimental results presented in Section 6. A more comprehensive discussion regarding the advantages and shortcomings of the proposed feature in context of different types of music is given in Sections 6 and 7.

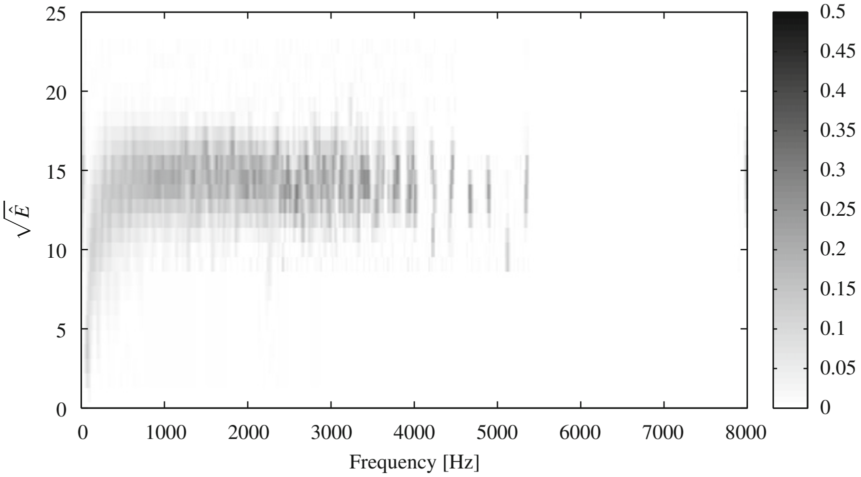

Fig. 6. Distribution of the EFMS values for a speech signal.

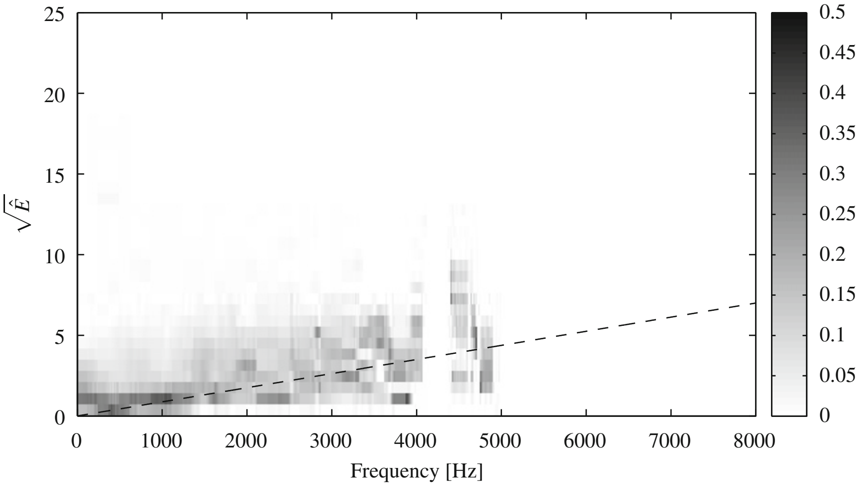

Fig. 7. Distribution of the EFMS values for a piano play (x_2).

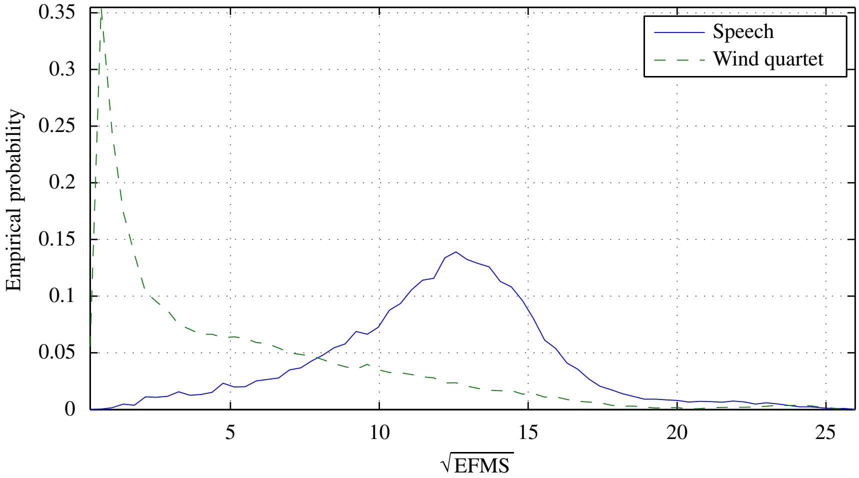

Fig. 8. Empirical probability density function of a musical excerpt played by a wind quartet and speech.

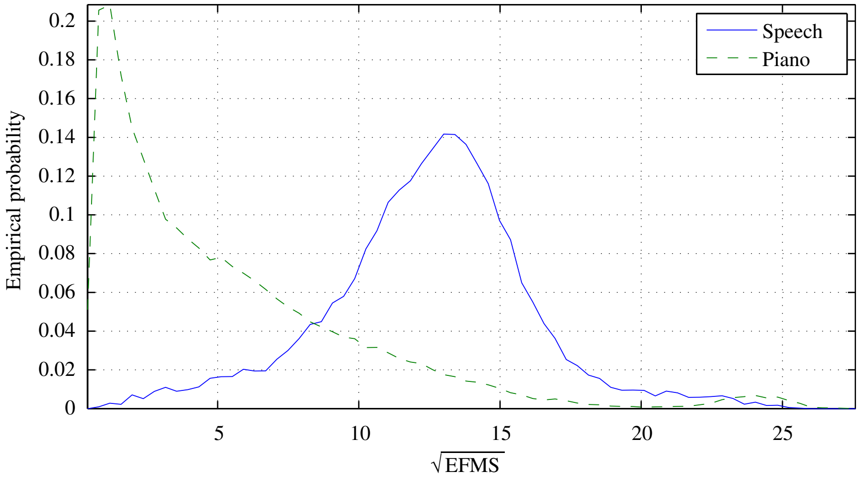

Fig. 9. Empirical probability density function of a musical excerpt played by a piano and speech.

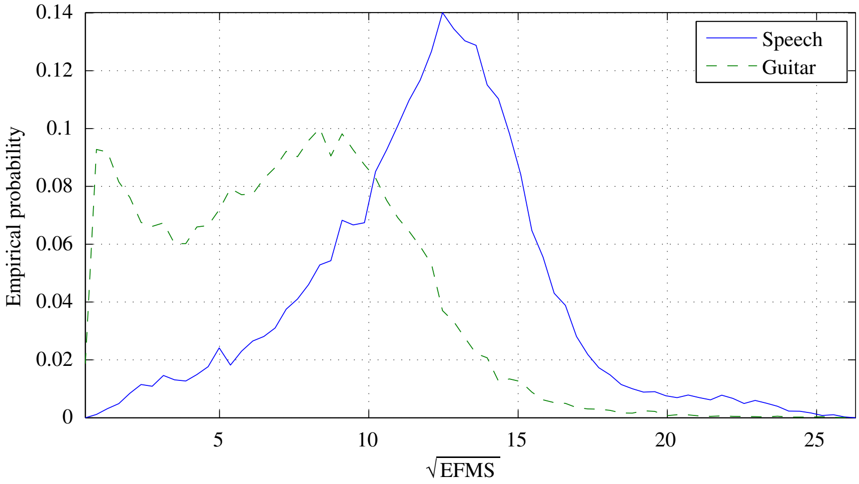

Fig. 10. Empirical probability density function of a musical excerpt played by a guitar and speech.

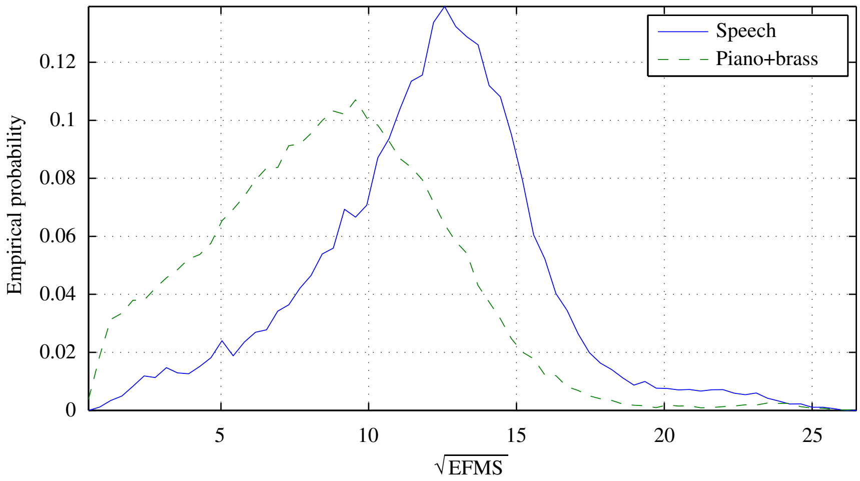

Fig. 11. Empirical probability density function of a musical excerpt played by a piano–brass duet and speech.

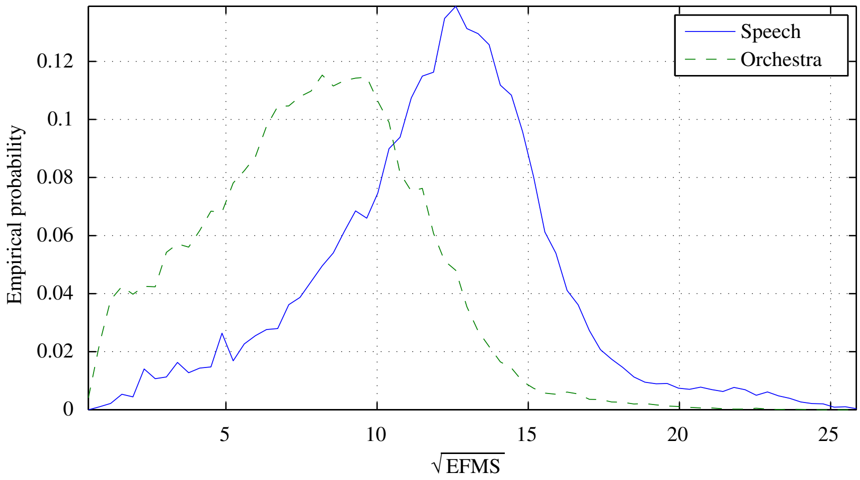

Fig. 12. Empirical probability density function of a musical excerpt played by an orchestra and speech.

4. Source separation procedure

In this section, we define a mixture model and the accompanying notations. We describe the model training and separation procedures.

Let $s_1(n)$ and $s_2(n)$ be time domain signals that belong to different signal classes. Let $x(n)$ be a mixture of $s_1(n)$ and $s_2(n)$

$$x(n) = s_1(n) + s_2(n)$$

As in the previous sections, we denote the STFT by capital letters, e.g. STFT of $s_c(n)$ is denoted by $S_{c,k}(m)$, where $c \in \{1,2\}$ denotes the signal class index. In the training stage we find the empirical probability density function for $\hat{E}\{s_1\}$ and $\hat{E}\{s_2\}$. In the separation stage we use the empirical pdf to define a minimum risk decision rule for classification of the STFT time–frequency bins based on $\hat{E}\{x\}$.

4.1. Training

The empirical pdf for class $c(\hat{\Pr}_c(\hat{E}))$ is estimated using a normalized histogram of

$$\{\hat{E}\{s_c\}_k(m) | (k,m) \in S_c\}$$

where S_c is a set of time–frequency bin indices where the energy is high compared to the neighboring bins. Let $M_{k,m}$ be the median of energy values in the time–frequency vicinity of (k,m) bin

$$M\{S_c\}_{k,m} = \text{median}\{|S_{c,i}(j)| \,|\, |i-k| \le d, |j-m| \le d\} \tag{26}$$

where d defines the vicinity. Define by δ_E a threshold for energy values, S_c is given by

$$S_c \triangleq \left\{ (k,m) \middle| 20\log_{10} \frac{|S_{c,k}(m)|}{M\{S_c\}_{k,m}} \ge \delta_E \right\} \tag{27}$$

Fig. 9 shows empirical pdfs of $\sqrt{\hat{E}}$ for speech and piano play signals. Large non-overlapping areas indicate that a separation of these signals using only $\hat{E}\{x\}$ values may be possible.

4.2. Separation

In this section, we explain our use of Bayes' minimum-cost with reject option decision rule to construct optimal STFT binary masks, the filtering process and the recovery of demixed source signal estimates.

Denote by $\gamma_k(m) = \hat{E}\{x\}_k(m)$ the estimated value of EFMS. Let $H_k^{(c)}(m)$ be an hypothesis that the signal from source c is present in (k,m) time–frequency bin. We will omit indices (k,m) for brevity where possible. Let $\mathcal{R}_1$ and $\mathcal{R}_2$ be the classification decision regions for different classes and $\mathcal{R}_r$ a rejection region, i.e. we prefer not to assign the sample into either class [25]. Let λ_{ij} be a penalty for assigning a sample γ to class i when in fact the sample belongs to class j. We define a loss function

$$L = \int_{\mathcal{R}_1} \lambda_{12} p(H^{(2)}|\gamma)p(\gamma)\,d\gamma + \int_{\mathcal{R}_2} \lambda_{21} p(H^{(1)}|\gamma)p(\gamma)\,d\gamma + \int_{\mathcal{R}_r} \lambda_r p(\gamma)\,d\gamma$$

To minimize this loss function, the decision regions should satisfy

$$\gamma \in \mathcal{R}_i \iff \begin{cases} \lambda_{ij} p(H^{(j)}|\gamma)p(\gamma) < \lambda_{ji} p(H^{(i)}|\gamma)p(\gamma) \\ \lambda_{ij} p(H^{(j)}|\gamma)p(\gamma) < \lambda_r p(\gamma) \end{cases}$$

$$\gamma \in \mathcal{R}_r \iff \lambda_r p(\gamma) \le \lambda_{ij} p(H^{(j)}|\gamma)p(\gamma) \quad i,j \in \{1,2\}; \;\; i \ne j.$$

We define $\eta \triangleq p(\gamma|H^{(1)})p(H^{(1)})/p(\gamma|H^{(2)})p(H^{(2)})$. Likelihood values $p(\gamma|H^{(1)})$ and $p(\gamma|H^{(2)})$ are estimated using the empirical probability density function obtained during the training. The $p(H^{(1)})$ and $p(H^{(2)})$ reflect prior belief of either class to be present in a time–frequency bin. Here we assume constant values of $\frac{1}{2}$ for $p(H^{(1)})$ and $p(H^{(2)})$. We note that these values affect the operation point of the algorithm. For example, a larger value of $p(H^{(1)})$ results in smaller distortion of the first-class signal at the expense of lower rejection rate of the interfering signal. Values of $p(H^{(1)})$ and $p(H^{(2)})$ can be estimated more accurately for any signal class by analyzing the sparsity of the signal representation in the STFT domain.

Using Bayes' formula, the decision rule can be rewritten as

$$\gamma \in \mathcal{R}_r \iff \begin{cases} \dfrac{\lambda_r}{\lambda_{12}} \le \dfrac{1}{1+\eta} \\ \dfrac{\lambda_r}{\lambda_{21}} \le \dfrac{1}{1+1/\eta} \end{cases} \tag{28}$$

$$\gamma \in \mathcal{R}_1 \iff \begin{cases} \dfrac{\lambda_{12}}{\lambda_{21}} < \eta \\ \dfrac{\lambda_r}{\lambda_{12}} > \dfrac{1}{1+\eta} \end{cases} \tag{29}$$

$$\gamma \in \mathcal{R}_2 \iff \begin{cases} \dfrac{\lambda_{12}}{\lambda_{21}} > \eta \\ \dfrac{\lambda_r}{\lambda_{21}} > \dfrac{1}{1+1/\eta} \end{cases} \tag{30}$$

We can tune the algorithm by changing values of $\lambda_{12},\lambda_{21},\lambda_r$. To decrease the number of class 1 time–frequency bins that are classified falsely as class 2, we may increase λ_{12}. This will result in higher penalty for this kind of mistake on the one hand (less false alarm errors), but on the other hand more time–frequency bins that truly belong to class 1 will now be classified as class 2 (more misdetect errors). If we decrease λ_r, more time–frequency bins will be rejected, i.e. not assigned to any of the signal classes. This will increase the number of time–frequency bins that cannot be classified reliably and will decrease the number of time–frequency bins of the interfering signal in both audio sources simultaneously. In our application, actual values of λ_{12}, λ_{21}, and λ_r are tuned manually.

We design a binary mask in the STFT domain by assigning each time–frequency bin to one of the signal classes based on (28)–(30). Time–frequency bins that are assigned to the interfering source or rejected are zeroed and those assigned to the desired signal are set to 1. For the binary mask to be effective, we assume that approximate W-disjoint orthogonality [21] holds. We verify this assumption in Section 6. Binary masks are defined by

$$M_k^{(c)}(m) = \begin{cases} 1, & \gamma_k(m) \in \mathcal{R}_c \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

$$c \in \{1,2\} \tag{32}$$

The interfering source is removed by multiplying the STFT of the mixture by $M^{(c)}$

$$\hat{X}_k^{(c)}(m) = M_k^{(c)}(m) X_k(m) \tag{33}$$

Inverse STFT gives a time domain estimate of the demixed source:

$$\hat{x}^{(c)}(n) = \mathrm{ISTFT}\{\hat{X}_k^{(c)}(m)\} \tag{34}$$

5. Evaluation criteria

In this section, we define the evaluation criteria used later on to evaluate the performance of the proposed algorithm and compare it to other separation algorithms. We use common distortion measures as described in [26] and BSS_EVAL toolbox [27]. Mixture components s_1 and s_2 are assumed uncorrelated. Let $\hat{s}_c$ be an estimate of s_c. The estimator will have the following decomposition:

$$\hat{s}_c = y_c + e_{c,\mathrm{interf}} + e_{c,\mathrm{artif}} \tag{35}$$

$$y_c := \langle \hat{s}_c, s_c \rangle s_c \tag{36}$$

$$e_{c,\mathrm{interf}} := \langle \hat{s}_c, s_{c'} \rangle s_{c'} \tag{37}$$

$$e_{c,\mathrm{artif}} := \hat{s}_c - (y_c + \langle \hat{s}_c, s_{c'} \rangle s_{c'}) \tag{38}$$

where c is the target class and $c^{'}$ is the interfering class. Now the following criteria are defined:

$$\mathrm{SDR} := 10\log_{10} \frac{\|y_c\|^2}{\|e_{c,\mathrm{interf}} + e_{c,\mathrm{artif}}\|^2} \tag{39}$$

$$\mathrm{SIR} := 10\log_{10}\frac{\|y_c\|^2}{\|e_{c,\mathrm{interf}}\|^2} \tag{40}$$

$$\mathrm{SAR} := 10\log_{10}\frac{\|y_c+e_{c,\mathrm{interf}}\|^2}{\|e_{c,\mathrm{artif}}\|^2} \tag{41}$$

SDR measures the total amount of distortion introduced to the original signal, both due to the interfering signal and artifacts introduced by the algorithm. SIR measures the amount of distortion introduced to the original signal by the interfering signal. SAR measures the amount of artifacts introduced to the original signal by the separation algorithm that do not originate in the interfering signal.

An additional measure employed in some experiments is the log spectral distance (LSD)

$$\mathrm{LSD}(X,Y) := \sqrt{\sum_{k=1}^{K}\sum_{m=1}^{N}\left(20\log_{10}\frac{|X_k(m)|}{|Y_k(m)|}\right)^2}$$

where $X_k(m)$ and $Y_k(m)$ are signals in the STFT domain.

6. Experimental results

In this section, we present experimental results. First, we verify the feasibility of source separation on synthetic signals. Then, we separate real audio recordings of speech and musical excerpts. We compare the performance of the proposed algorithm to an existing source separation algorithm. We study the effect of particular speakers and different musical instruments on the performance of the algorithm.

6.1. Synthetic signals separation

First we verify the ability of the proposed algorithm to segregate between signals that differ in their frequency modulation extent. We choose synthetic signals whose properties are similar to voiced phonemes and piano play (i.e. several frequency modulated partials). More specifically:

$$s_c(n) = \sum_{l=0}^{N_h}\cos\left(l\cdot 2\pi f_c^{(c)}n/f_s + \sum_{m=0}^{n} q_l^{(c)}(m)\frac{1}{T}\right) \tag{42}$$

$$q_l^{(c)}(n) = l\cdot d^{(c)}\cos(2\pi f_m^{(c)}n/f_s) \tag{43}$$

where $c\in\{1,2\}$ is the class index, N_h is the number of harmonic partials, f_c, f_s and f_m are carrier, sampling and modulation frequencies, respectively, and d is the modulating signal amplitude. We choose N_h=6, $f_c^{(1)}$=400 Hz, $f_m^{(1)}$=10 Hz, $d^{(1)}$=20, $f_c^{(2)}$=500 Hz, $f_m^{(2)}$=10 Hz, $d^{(1)}$=1. Note that $d^{(1)} \gg d^{(2)}$ as assumed by our model for speech and piano. We normalize the variance of s_c to 1. Fig. 13 shows spectrograms of synthetic signals used in this experiment, and Fig. 14 shows source signal estimates recovered from the mixture.

We perform an additional experiment that shows that our algorithm is capable of separating white noise from weakly frequency-modulated signal. These signals have properties similar to unvoiced phonemes and piano play. Fig. 15(a) shows the spectrogram of the mixture, and Fig. 15(b) shows the estimated harmonic component of the mixture. No white noise residua are visible in the harmonic component estimate. Table 1 shows separation performance results of experiments performed on synthetic signals. A small artifact is visible at Fig. 14(b). We notice that some energy of the second partial (800 Hz) of s_1 has "leaked" to $\hat{s}_2$. We attribute this leakage to a simplified learning of EFMS distribution at different frequency bands. The amplitude of the frequency modulating signals of each harmonic partial grows linearly with the amplitude of the frequency modulating signal of the fundamental partial as shown in Fig. 4. This behavior is ignored by the training of signal model. By visually examining the spectrograms of the signal estimates and noticing high values of objective measures we conclude that our method is capable of segregating two audio signals that have similar properties to real speech and real piano play signals.

6.2. Separation of a speech and piano play mixture

Now we describe the simulation and the informal listening test results of the proposed algorithm and compare its performance to a Gaussian Mixture Model (GMM) monaural separation algorithm [2]. We use 60 s of speech (either male or female) taken from TIMIT database sampled at 16 KHz and Chopin's

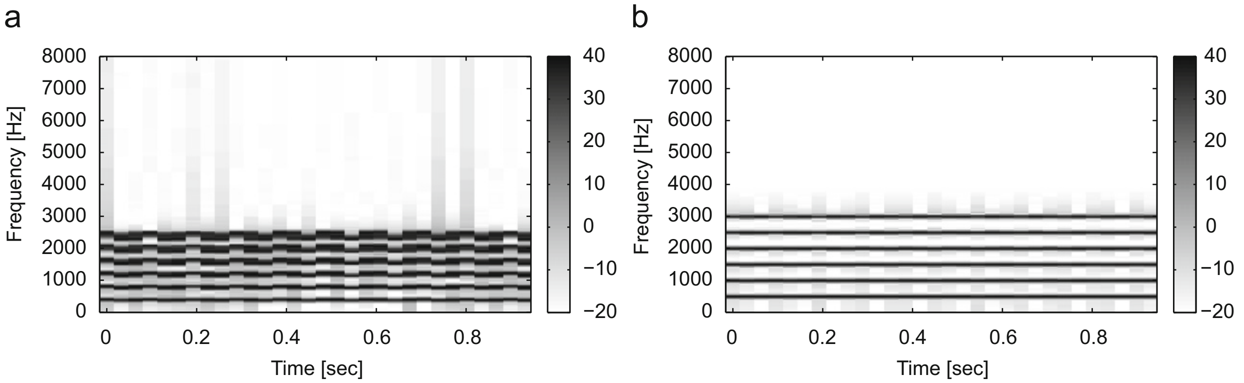

Fig. 13. Spectrogram (in dB) of synthetic signals used for testing: (a) strongly frequency modulated signal; (b) weakly frequency modulated signal.

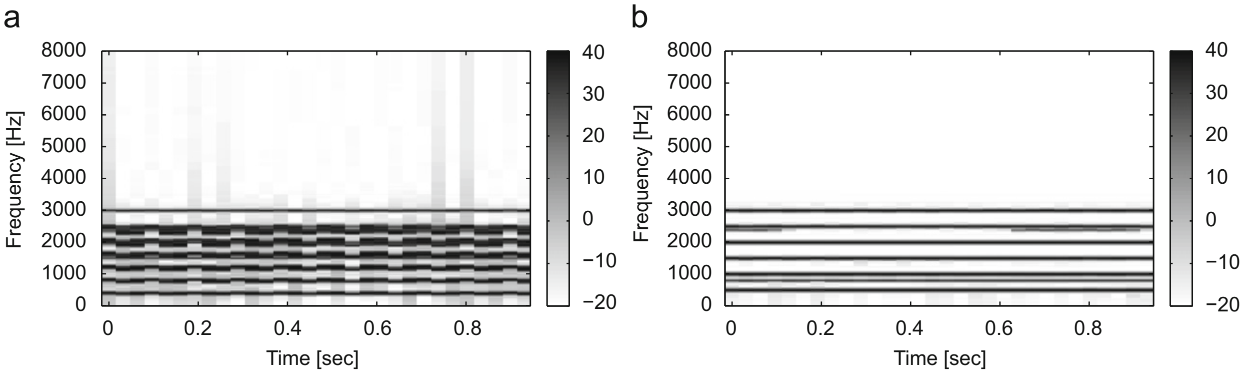

Fig. 14. Spectrograms (in dB) of (a) an estimate of strongly frequency modulated signal; (b) weakly frequency modulated signal. Both signals are recovered from a 0 dB mixture.

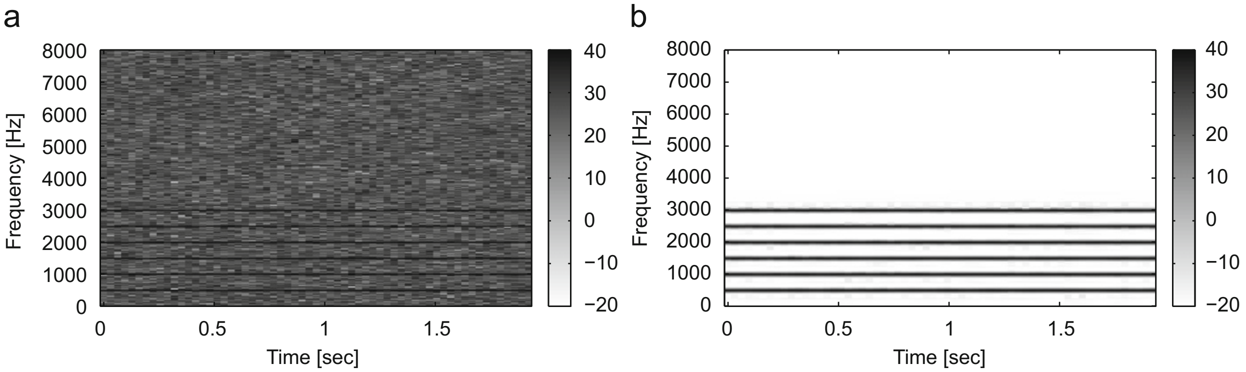

Fig. 15. Spectrograms (in dB) of (a) a mix of white noise and weakly frequency-modulated signal; (b) estimate of weakly frequency modulated signal.

Table 1
Synthetic signal separation.

	SDR$_1$	SIR$_1$	SAR$_1$	SDR$_2$	SIR$_2$	SAR$_2$
(a)	10.4	15.3	12.3	10.8	24.5	11.1
(b)	9.8	10.4	18.6	15.2	35.8	15.2

(a) Two frequency modulated Signals; (b) noise and frequency modulated signal.

Table 2
Separation performance analysis.

	SDR$_1$	SIR$_1$	SAR$_1$	LSD$_1$	SDR$_2$	SIR$_2$	SAR$_2$	LSD$_2$
Oracle mask	18.9	42.6	18.9	0.73	17.9	47.2	18.0	0.8
EFMS female	6.0	11.5	7.7	1.9	5.8	20.6	6.0	1.6
EFMS male	5.7	11.8	7.3	2.4	5.5	17.3	5.9	1.6
GMM	2.4	9.3	3.8	2.9	2.6	7.9	4.8	2.5

prelude for piano Opus 28 No. 6 for GMM training. We use 1024 points STFT, Hamming synthesis window, 50% overlap and 12 components GMM. The parameters used for the proposed algorithm were: N=1024, M=64, N_u=121, $\delta_E = 15\,\text{dB}$, $\lambda_{12} = \lambda_{21} = 1$, $\lambda_r = \infty$. The high-pass filter used for the removal of Ω_c component was a 122 taps FIR filter with stop angular frequency of 0.01π.

The W-DO value [21] for the pair of signals used in our experiment is 0.94, which according to [21] guarantees perceptually perfect separation when the following binary masks are used in the STFT domain:

$$\tilde{M}_k^{(1)}(m) = \begin{cases} 1, & \dfrac{|S_{1,k}(m)|}{|S_{2,k}(m)|} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{44}$$

$$\tilde{M}_k^{(2)}(m) = \begin{cases} 1, & \dfrac{|S_{1,k}(m)|}{|S_{2,k}(m)|} \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{45}$$

In [28] these masks were coined "oracle" masks. The performance of the oracle mask in source separation induces an upper bound on the performance of our algorithm. We present the results of source separation using these masks together with other separation results.

We calculated SDR and LSD measures using 30–120 s long test sequences. For all verified test signal length, the calculated measure values fell into a 0.2 dB range, hence the sensitivity of these measures to the test sequence length is relatively low. We evaluated the performance of algorithms using 45 s of speech and piano signals. We used speech and music excerpts different from the ones used for training. Chopin's prelude for piano Opus 28 No. 7 was used as a test musical excerpt.

The results are shown in Table 2. A 0 dB mixture of test signals was used in all experiments. When the performance was evaluated using a non-separated mixture, the SDR and SIR measures are approximately zero (± 0.3 dB) and the SAR measure was greater than 260 dB. We notice that the separation quality of the mixture that contained female speech is slightly higher than male speech. This can be explained by the absence of low frequency pitch tracks that are falsely estimated as music components. We elaborate on this issue in the following paragraphs when we discuss the residual signal.

Figs. 16 and 17 show spectrograms of speech and piano play signals used in the mixture together with the signals recovered by the GMM based algorithm and by the proposed algorithm. Smaller amounts of interfering signals can be seen in signals recovered by the proposed method compared to the GMM based algorithm. The spectrogram of piano play signal reveals that the non-harmonic components of piano note onsets are missing. The reason is that piano strings excited by a strike of a felt covered hammer produce a strong non-harmonic component near the note onset. Only harmonic components of piano play are detected by our algorithm and the rest of the signal leaks into the estimated speech component.

Informal listening to the signals separated by the proposed algorithm reveals that despite visible artifacts in the extracted speech and piano signals, the proposed method produces perceptually more plausible sound than the GMM based algorithm. We note that a mixture separated using oracle masks perceptually has almost perfect perceptual quality.

To find out which part of the speech signal leaks into the piano channel, we applied our algorithm to a clean speech signal (instead of speech–piano mixture, i.e. $x(n)=s_1(n)$). Perfect separation algorithm would estimate $\hat{s}_2(n)=0$. The spectrogram of the actual $\hat{s}_2$ signal is presented in Fig. 16(d). The leaking speech parts are harmonic in their nature, located mostly in low frequencies and have constant pitch over relatively long periods of time (0.5–1 s). These low frequency harmonic partials can also be seen in Fig. 17(c). The low-frequency leaking components may be the reason for a slightly better separation performance for female over male speakers. A certain amount of musical noise is also present. Applying the algorithm to a clean piano play signal and examining $\hat{s}_1$, Fig. 17(d) reveals that most of the leaking signal results from the piano hammer strikes. This conclusion was confirmed by informal listening.

To achieve the most perceptually plausible separation results we tuned λ_{12}, λ_{21}, and λ_r manually. We chose $\lambda_{12}=4, \lambda_{21}=1$, and $\lambda_r=0.4$. We measured $\mathrm{SDR}_1=-0.1$, $\mathrm{SIR}_1=18.6$, $\mathrm{SAR}_1=0.0$, $\mathrm{SDR}_2=5.5$, $\mathrm{SIR}_2=21.5$, $\mathrm{SAR}_2=5.6$. Although some measures show deteriorated performance, the extracted speech is intelligible and contains a very low amount of audible interfering signal. Piano audible quality remains similar to the previous experiment with slightly higher rate of interfering signal rejection.

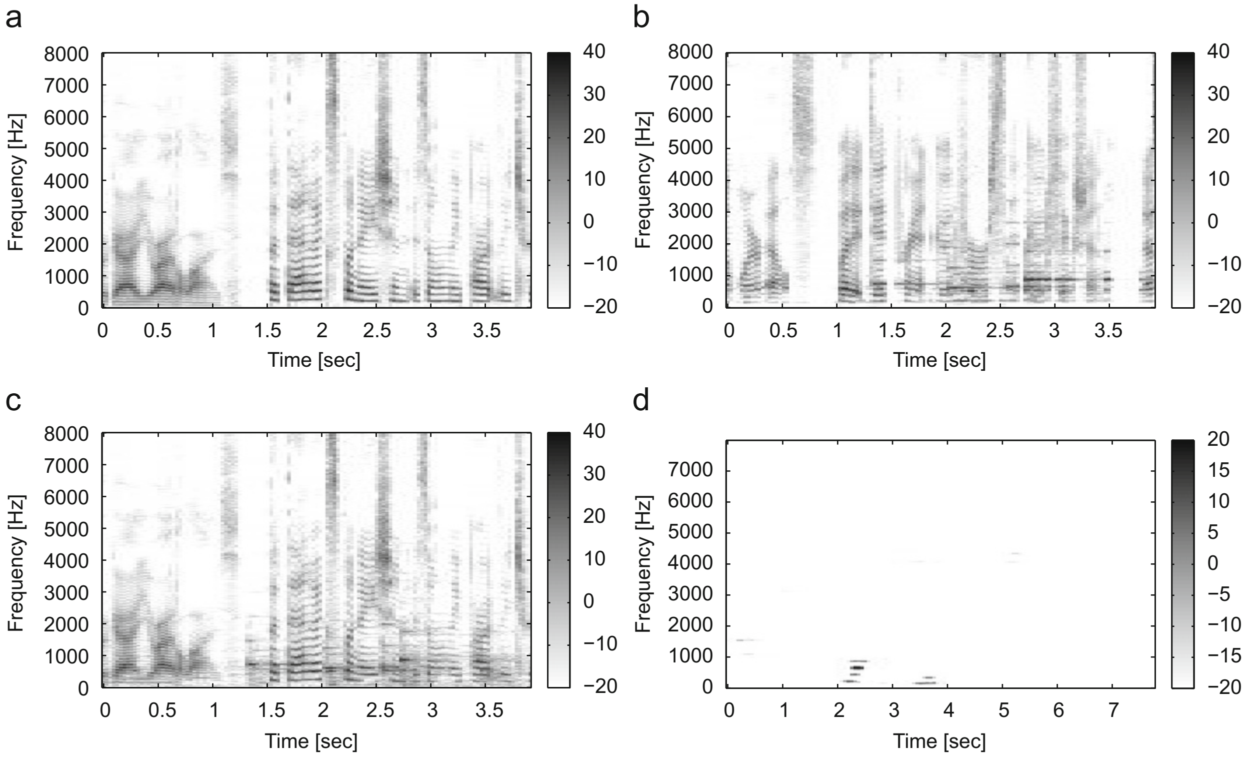

Fig. 16. Spectrograms (in dB) of the (a) clean, (b) GMM based algorithm recovered, (c) the proposed algorithm recovered speech signals, and (d) residual speech signal after applying the algorithm to clean speech signal.

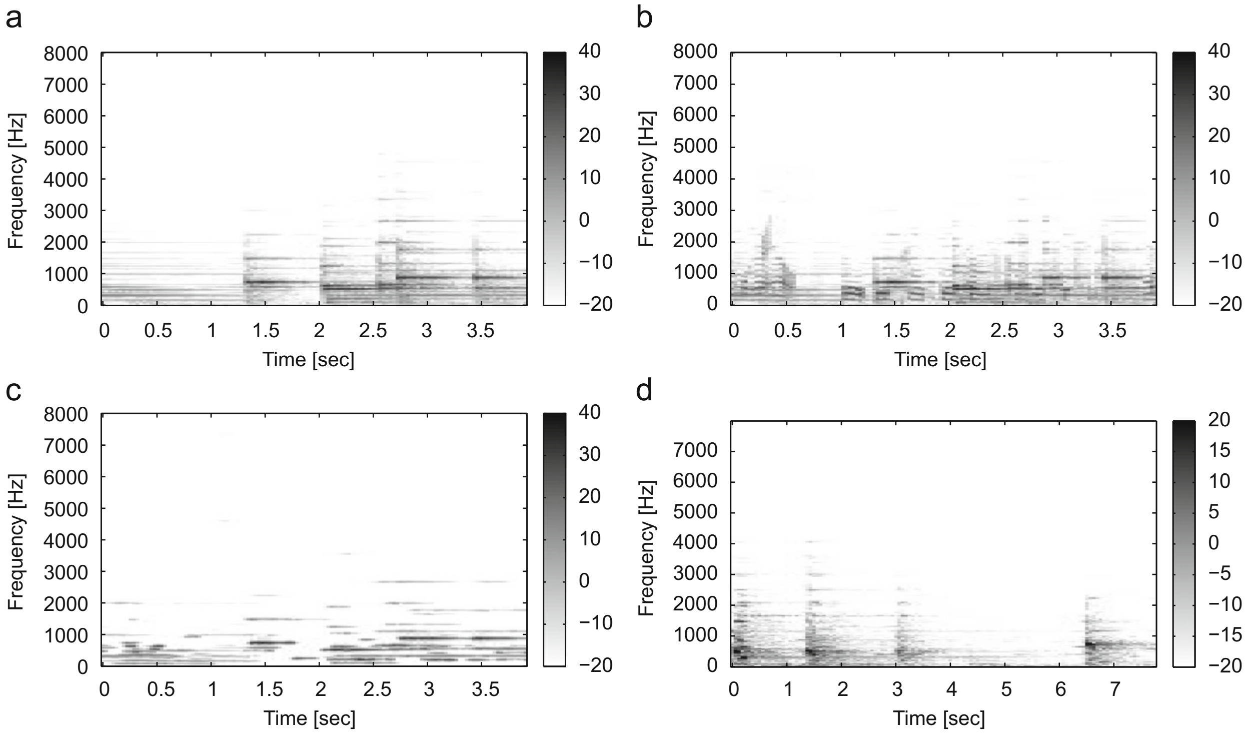

Fig. 17. Spectrograms (in dB) of the (a) clean, (b) GMM based algorithm recovered, (c) the proposed algorithm recovered piano signals, and (d) residual piano play signal after applying the algorithm to clean piano signal.

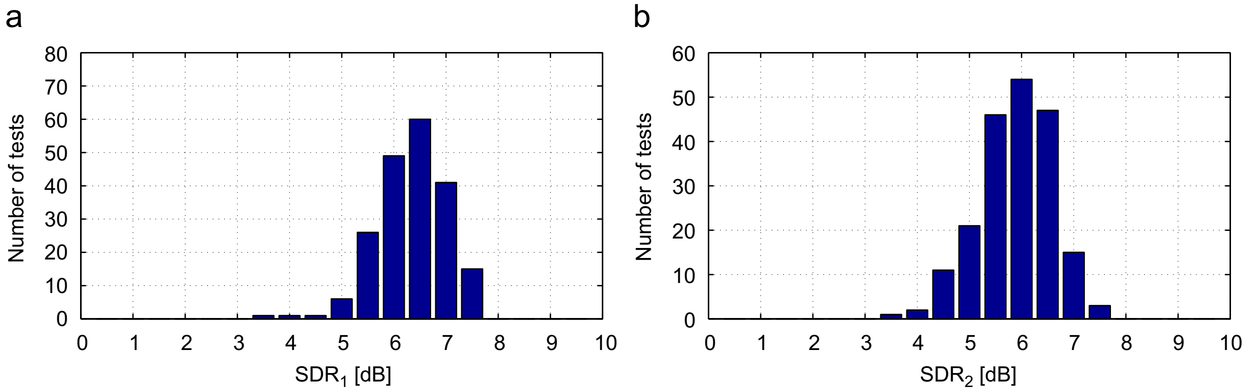

Fig. 18. Histograms of SDR$_1$ and SDR$_2$ measures calculated for 200 speech/piano separation experiments performed on different speakers.

6.3. *Dependency on a particular speaker*

We studied the dependency of a particular speaker on the quality of the separated signal. Two separate male and female speech models were trained using 60 s of TIMIT speech. We generated 200 speech excerpts containing concatenated sentences pronounced by 200 TIMIT speakers, different from those used for the model training. Each one of the 200 speech excerpts was mixed with a piano play excerpt (Chopin prelude Opus 28 No. 7) and the sources were extracted using our method. The values of SDR$_1$ and SDR$_2$ were calculated. These values are depicted in the histograms in Fig. 18. Eighty-five percent of all SDR$_1$ and SDR$_2$ values are found within a 2 dB range, and 98% of all SDR$_1$ and SDR$_2$ values are found within a 3 dB range. This means that the dependency of our algorithm on a particular speaker is low. By listening to the speech excerpts that were poorly separated, we noticed that they were pronounced by a monotone reading voice. This kind of speech is uncommon in regular conversation and has almost constant pitch which results in low values of EFMS and time–frequency bin misclassification.

6.4. Dependency on the type of music

To demonstrate the applicability of the proposed method to different musical instruments we examine separation performance on the following musical excerpts: wind quartet, slow piano, fast piano, orchestra, piano and brass duet, and fast playing guitar. The training of the algorithm is performed on the first minute of the musical piece and the separation performance is evaluated on the second minute of the same piece. The same piano model (based on slow piano play) was used for separation of speech from slow and fast piano play. Although we showed that a higher quality of separation can be achieved on female speech compared to the male speech, we note that the algorithm has similar performance when trained on a single or mixed gender speech. In this experiment, mixed male/female speech was used for training and testing of the algorithm.

Table 3 shows the W-DO together with the SDR of the separated signals produced by our algorithm. The best separation is achieved in the case of wind quartet excerpt.

Despite a relatively high objective measure achieved in the separation of speech from guitar play, the perceptual quality of musical signal is rather low. Although, the melody can be easily recognized and speech residua are low, a pluck sound, characteristic to a guitar play is missing from the extracted musical signal. This makes it hard to identify that the musical piece is played by a guitar.

The separation performance of our algorithm on a fast piano piece is inferior compared to a slow piano piece. Long time segments with a constant pitch improve EFMS estimation while a naive approach to smoothing in Eq. (22) blurs the onsets and offsets of notes and affects mostly fast musical pieces.

The fast tempo of the piano/brass duet is similar to the fast piano piece, except for the additional monophonic brass instrument. In this case the separation performance is similar to the fast playing piano. This is not surprising, since both musical pieces are similar in their tempo and the STFT representation sparsity.

The worst separation was observed on a mixture of speech and orchestra. Percussion instruments were present in the orchestra. These instruments have only non-harmonic components hence our algorithm confuses them with speech. Indeed, when listening to the estimated mixture components, all percussion instruments are audible in the speech estimate.

We conclude that the separation performance deteriorates when music present in the mixture has a large amount of non-harmonic components. The musical excerpts that were most poorly separated by the algorithm had also lowest W-DO values. Despite this fact, we cannot conclude that low W-DO values were the reason for poor separation quality. The differences among tested cases in the W-DO measure are not very large. According to [21] a perfect separation would be guaranteed if appropriate masks would have been used. Besides, in our experiments musical excerpts with lower W-DO values also had a greater amount of non-harmonic components.

Summarizing the experimental results we conclude that for the speech/piano play scenario (Table 2), the proposed algorithm exhibits superior separation performance both in terms of objective measures and subjective listening tests when compared to the GMM based algorithm. Relatively a short training signal (tens of seconds) is sufficient for the algorithm. The quality of the extracted female speech is superior to the quality of the extracted male speech. The algorithm is not sensitive to a particular speaker; however, the performance slightly deteriorates when the speech is pronounced by a monotone reading voice, not common to regular speech.

We studied differences in separation performance for various types of music and musical instruments. Musical instruments that have less non-harmonic components are better separated by the proposed algorithm. Slow musical excerpts are better separated than the fast ones. Percussive instruments cannot be separated from speech by our method. Failure to distinguish non-harmonic signal components in music and speech produces audible artifacts in the extracted signals.

Audio files used for training and performance evaluation as well as separation results can be downloaded from http://sipl.technion.ac.il/∼elitvin/EFMS/.

7. Conclusions

We have presented and evaluated a novel technique for single-channel source separation based on the energy of frequency modulating signal. The proposed method requires a relatively simple training and produces separation results that are superior to a more complicated GMM based method, when compared in the speech/piano play separation scenario. We demonstrated that the FM based instantaneous features are well localized in time and frequency, and carry sufficient information to allow signal classification and separation.

There are two key properties that would guaranty a good separation. First, there must be a significant difference in the energy of subband FM components of two signal classes. In other words, it is important that one of the signals would have more rapidly changing pitch tracks. It is also important that at most one of the signals would contain non-harmonic components. Second, signals

Table 3
Analysis of separation performance for different musical instruments and styles.

	W-DO	SDR_1	SDR_2
Wind quartet[a]	0.95	9.4	9.3
Piano slow[b]	0.95	5.6	5.3
Guitar[c]	0.93	4.85	4.75
Piano fast[d]	0.92	4.4	4.1
Piano/brass duet[e]	0.93	4.0	4.2
Orchestra[f]	0.92	2.2	2.1

[a] Beethoven: Quartet No. 10 in Eb, Op. 74.
[b] Chopin: prelude Opus 28 No. 7.
[c] Francisco Tarrega: Romance De Juegos Prohobidos.
[d] Chopin: prelude Opus 28 No. 1.
[e] Loes: This Little Light of Mine.
[f] Berlioz: Apotheose from symphonic funebre et triomphale.

must be W-disjoint orthogonal in order to guaranty good separation quality using binary masks. Although these restrictions are rather strict, and not any pair of signal classes can be separated by the proposed algorithm, we have shown that for speech and some types of music, the separation is feasible.

There are several pitfalls in the proposed method. The EFMS pdf is estimated independently of the frequency band index. More accurate statistical modeling for each frequency band or a group of neighboring frequency bands can solve the problem of signal leakage at lower frequencies as described in Section 6.

The algorithm classifies signals based only on the energy of the subband FM component and it is blind to different FM signal nuances that are beyond their energy level. Detailed subband FM signal analysis, such as spectral decomposition or time varying statistics, may provide a better classification feature. It may extend the variety of audio classes suitable for separation by our method, improve the accuracy of speech/music separation, and allow addressing a multiclass separation problem.

Non-harmonic components present in some types of music are impossible to separate using our method. Additional information must be employed by the algorithm to enable separation of non-harmonic signals. It might be useful to incorporate other features used in Music Information Retrieval community, for example the GMM based algorithm proposed by Benaroya et al. [29].

The computational complexity of the separation procedure is relatively low. The most computationally intensive processing stages are the standard FFT and the DESA, both evaluated one time per each frame. The computational complexity of the training procedure is slightly higher due to calculation of the median. The algorithmic delay is defined by half the length of the EFMS smoothing window. In our experimental setup the resulting delay value was 242 ms. A real-time implementation seems to be possible, although the delay may be too high in some applications.

The plots of empirical pdf values of the EFMS for speech and music suggest that the EFMS can serve as a good classification feature. In experiments that are not reported in this paper, we observed that in the task of frame-by-frame classification of piano and speech signals the EFMS feature showed superior performance compared to the well known MFCC feature.

Despite the training signals availability requirement, our method is applicable to various real life applications such as audio tracks remastering or speech enhancement in the presence of music. The proposed algorithm can also operate in a semi-supervised manner as part of audio editing software. The properties of subband frequency modulating signals may provide additional information that may be useful in other audio processing applications, such as speech enhancement, audio coding or audio classification.

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments and helpful suggestions.

References

[1] A. Ozerov, P. Philippe, F. Bimbot, 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 15 (5) (July 2007) 1564–1578.

[2] L. Benaroya, F. Bimbot, Wiener based source separation with HMM/GMM using a single sensor, in: ICA 2003 Nara, Japan, April 2003, pp. 957–961.

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

[4] F.R. Bach, M. I. Jordan, Blind one-microphone speech separation: a spectral learning approach, in: NIPS, Vancouver, 2004.

[5] M. Helén, T. Virtanen, Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine, in: Proceedings of the 13th European Signal Processing Conference (EUSIPCO 2005), Turkey, 2005.

[6] T. Virtanen, Sound source separation using sparse coding with temporal continuity objective, in: International Computer Music Conference, ICMC, 2003.

[7] H.M. Teager, S.M. Teager, A phenomenological model for vowel production in the vocal tract, in: R.G. Daniloff (Ed.), Speech Science: Recent Advances, College-Hill Press, San Diego, CA, 1985, pp. 73–109 (Chapter 3).

[8] H.M. Teager, S.M. Teager, Evidence for nonlinear sound production mechanisms in the vocal tract, in: W.J. Hardcastle, A. Marchal (Eds.), Speech Production and Speech Modeling, vol. 55, Kluwer Academic, Boston, 1989, pp. 241–261.

[9] J. Kaiser, On a simple algorithm to calculate the 'energy' of a signal, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, April 1990, pp. 381–384.

[10] P. Maragos, J. Kaiser, T. Quatieri, Energy separation in signal modulations with application to speech analysis, IEEE Transactions on Signal Processing 41 (10) (October 1993) 3024–3051.

[11] P. Maragos, T.F. Quatieri, J.F. Kaiser, Speech nonlinearities, modulations, and energy operators, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, 1991, pp. 421–424.

[12] P. Maragos, J. Kaiser, T. Quatieri, On amplitude and frequency demodulation using energy operators, IEEE Transactions on Signal Processing 41 (4) (April 1993) 1532–1550.

[13] A. Potamianos, P. Maragos, Speech formant frequency and bandwidth tracking using multiband energy demodulation, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, May 1995, pp. 784–787.

[14] A. Potamianos, P. Maragos, Time–frequency distributions for automatic speech recognition, IEEE Transactions on Speech and Audio Processing 9 (3) (March 2001) 196–200.

[15] T. Thiruvaran, E. Ambikairajah, J. Epps, Speaker identification using FM features, in: Proceedings of 11th Australasian International Conference on Speech Science and Technology, Auckland, New Zealand, 2006, pp. 148–152.

[16] D.V. Dimitriadis, P. Maragos, A. Potamianos, Robust AM–FM features for speech recognition, IEEE Signal Processing Letters 12 (9) (September 2005) 621–624.

[17] C.R. Jankowski Jr., T.F. Quatieri, D.A. Reynolds, Measuring fine structure in speech: application to speaker identification, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, May 1995, pp. 325–328.

[18] A. Potamianos, P. Maragos, Speech analysis and synthesis using an AM–FM modulation model, Speech Communication 28 (3) (1999) 195–209.

[19] R. Sussman, M. Kahrs, Analysis and resynthesis of musical instrument sounds using energy separation, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, May 1996, pp. 997–1000.

[20] T. Virtanen, A. Klapuri, Separation of harmonic sound sources using sinusoidal modeling, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, 2000, pp. II765–II768.

[21] O. Yilmaz, S. Rickard, Blind separation of speech mixtures via time–frequency masking, IEEE Transactions on Signal Processing 52 (7) (July 2004) 1830–1847.

[22] L. Atlas, C. Janssen, Coherent modulation spectral filtering for single-channel music source separation, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)-05, vol. 4, March 2005, pp. iv/461–iv/464.

[23] S. Disch, B. Edler, Multiband perceptual modulation analysis, processing and synthesis of audio signals, in: Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 2009, pp. 2305–2308.

[24] Y. Litvin, I. Cohen, D. Chazan, Separation of speech and music sources from a single-channel mixture using discrete energy separation algorithm, in: International Workshop on Acoustic Echo and Noise Control, IWAENC, 2010.

[25] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley & Sons, Inc., New York, 2001.

[26] R. Gribonval, L. Benaroya, E. Vincent, C. Févotte, Proposals for performance measurement in source separation, in: Proceedings of the 4th International Symposium on ICA and BSS (ICA2003), Nara, Japan, April 2003, pp. 763–768.

[27] C. Févotte, R. Gribonval, E. Vincent, BSS_EVAL toolbox user guide revision 2.0, Technical Report 1706, IRISA, Rennes, France, April 2005 [Online]. Available: 〈http://www.irisa.fr/metiss/bsseval/〉.

[28] E. Vincent, R. Gribonval, M.D. Plumbley, Oracle estimators for the benchmarking of source separation algorithms, Signal Processing 87 (8) (2007) 1933–1950.

[29] L. Benaroya, F. Bimbot, R. Gribonval, Audio source separation with a single sensor, IEEE Transactions on Audio, Speech, and Language Processing 14 (1) (January 2006) 191–199.