Separation of speech and music sources from a single-channel mixture using discrete energy separation algorithm

Yevgeni Litvin[1], Israel Cohen[1], and Dan Chazan[2]

[1]Department of Electrical Engineering
Technion - Israel Institute of Technology
Technion City, Haifa 32000, Israel
{elitvin@tx, icohen@ee}.technion.ac.il

[2]IBM Research Laboratory
Haifa, Israel

***Abstract*—In this paper, we address the problem of monaural source separation of a mixed signal containing speech and piano components. We use Discrete Energy Separation Algorithm (DESA) to estimate frequency-modulating (FM) signal energy. We design a time-varying filter in the time-frequency domain for rejecting the interfering signal. An estimation of the FM signal energy employs instantaneous signal properties that are localized both in time and frequency. We present experimental results which demonstrate the advantages of the proposed method using real audio signals.**

I. 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]; Computational Auditory Scene Analysis (CASA) [3]; Non-negative Matrix Factorization (NMF) [4] 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. Many existing solutions produce satisfactory results in special cases, the general problem of single-audio-channel BSS remains unsolved.

Teager and Teager [5] studied airflow and fluid dynamics of human speech apparatus, and described several nonlinear phenomena as well as their sources. Later, Kaiser [6] formulated the Teager Energy Operator (TEO). In [7] 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.

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. Our algorithm uses AM-FM analysis and the properties of the FM signal to differentiate between audio signal classes. First we filter the input signal by a short time Fourier Transform (STFT) filterbank. Then we use the DESA algorithm 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 (TF) bins in the STFT domain are classified into one of the target signal classes using EFMS values. The interfering signal is suppressed by zeroing TF bins attributed to the interfering signal. Finally, we reconstruct the separated component by inverting the STFT. Repeating the process twice, each time selecting a different audio source class as interfering, we recover segregated signals. The method is described in [8] in greater details.

The remainder of this paper is structured as follows. In Section II, we describe the TEO operator and the DESA algorithm used for the AM-FM analysis. In Section III, we explain why the proposed method should perform well in the separation task. Section IV defines the evaluation procedure of the EFMS. Section V describes a simple training procedure used to learn EFMS features and a Bayesian approach used for the creation of an STFT domain binary mask. In Section VI we present experimental results .

II. Discrete energy separation algorithm

In this section, we introduce mathematical notations and define AM-FM analysis using TEO (DESA algorithm [7]).

Let $x_c\left(t\right)$ be a continuous time signal and $x\left(n\right)=x_c\left(nT\right)$ be its sampled version with sampling period of T. We assume the following signal model

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

where n is a discrete time index, Ω_c is an angular frequency of a carrier, θ is some constant phase value, and $a\left(n\right)$ and

$q(n)$ are the amplitude and frequency modulating signals, respectively.

A discrete version of TEO is an operator $\Psi[x(n)]$ defined as:

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

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

$$\begin{aligned}\hat{\Omega}_i(n) &\approx \frac{1}{2}\arccos\left(1 - \frac{\Psi[x(n+1) - x(n-1)]}{2\Psi[x(n)]}\right) &(3)\\ &\approx \Omega_c + q(n) &(4)\\ |\hat{a}(n)| &\approx \frac{2\Psi[x(n)]}{\sqrt{\Psi[x(n+1) - x(n-1)]}}\;. &(5)\end{aligned}$$

These approximations are valid if some mild conditions on highest non-zero angular frequencies of $a(n)$, $q(n)$ and AM modulation index hold [7]. This version of DESA algorithm is called DESA-2 [7].

III. 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 TF discriminating factor and used for rejecting the interfering source.

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. In one of its forms it can be written as:

$$X_k(m) = e^{-j\frac{2\pi}{N}mM}\,(x * w_a)(mM)\;. \tag{6}$$

where $w_a(n)$ is an analytic bandpass filter and M is time subsampling factor.

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.

IV. EFMS calculation

Assume the AM-FM model (1) for the l-th harmonic partial $x_l(n)$ and assume that almost all the energy of $x_l(n)$ resides in the k-th subband of the STFT filterbank. The following procedure describes evaluation of the EFMS. Let $\alpha \in \mathbb{R}, 0 < \alpha < 1$. Each STFT frequency band $X_k(m)$ is modulated to an intermediate frequency $\Omega_{\text{if}} = \alpha\pi$ by multiplying $X_k(m)$ by $e^{j\Omega_{\text{if}}m}$. DESA operates on the real valued signals, we use only the in-phase component of the modulated filterbank output $\tilde{X}_k(n) = \Re\left(X_k(n)\,e^{j\Omega_{\text{if}}n}\right)$. It can be shown [8] that M has to satisfy $M \leq \min\{\alpha N, (1-\alpha)N\}$ in order to avoid aliasing. DESA estimator (3) can now be used to find the instantaneous frequency $\hat{\Omega}_{i,k}(m)$ in each frequency band.

The instantaneous frequency $\hat{\Omega}_{i,k}(m)$ also includes a constant term that originates from the carrier frequency. To remove it we filter $\hat{\Omega}_{i,k}(m)$ with a high-pass filter h_q and get an estimate of $q(n)$. Note that Ω_c is not necessarily constant in time, but we assume that it changes slowly compared to $q(n)$,

$$\hat{q}(n) \quad \approx \quad \left(\left(\tilde{\Omega}_c + \Omega_{\text{if}} + q(n)\right) * h_q\right)(n)\,.$$

We define the EFMS by

$$\hat{E}_k(m) \quad \triangleq \quad \left(u * \hat{q}_k^2\right)(m)\;, \tag{7}$$

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.

The upper pane of Fig. 1 shows the 50 lower frequency bands of the STFT filterbank output for a speech utterance. 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 algorithm. There are several amplitude peaks corresponding to voiced phonemes. The third pane shows the $\hat{\Omega}_{i,16}$ estimate. The lowest pane shows a plot of $\hat{E}_{16}(m)$. In 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 TF locations. As a result, the values of EFMS at the location of unvoiced phonemes are high. The piano play fragment depicted in Fig. 2 contains several piano notes. The

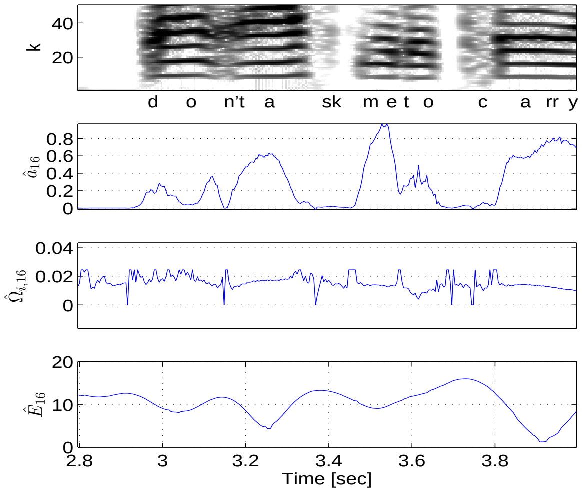

Figure 1. The spectrogram (50 lower frequency bands) of the speech utterance (vertical axis labels show frequency band numbers); the estimated AM component $\hat{a}_{16}$, the estimated instantaneous frequency $\hat{\Omega}_{i,16}$ and the EFMS ($\hat{E}_{16}\left(n\right)$) of the 16-th frequency band .

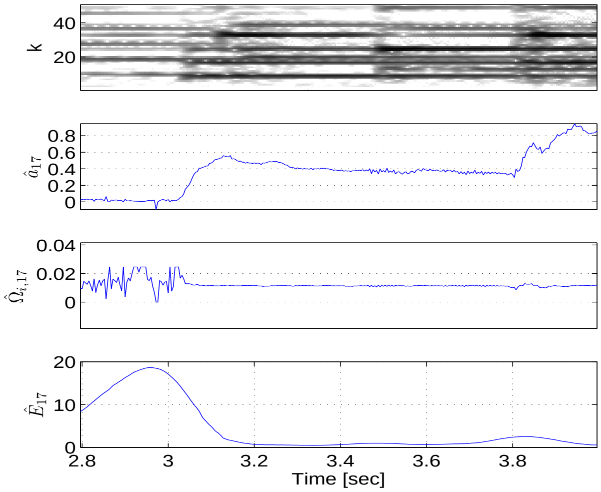

Figure 2. The spectrogram (50 lower frequency bands) of the piano play signal (vertical axis labels show frequency band numbers); the estimated AM component $\hat{a}_{17}$, the estimated instantaneous frequency $\hat{\Omega}_{i,17}$ and the EFMS ($\hat{E}_{17}\left(n\right)$) of the 17-th frequency band .

$\hat{E}_{17}\left(m\right)$ values are low while the note is being played. We conclude that two signals can be distinguished by comparing a one-dimensional value of the EFMS.

V. Source separation procedure

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

$$x\left(n\right) = s_1\left(n\right) + s_2\left(n\right) .$$

A linear mixture of two signals is a realistic assumption in some real-life scenarios. It is irrelevant whether s_1 or s_2 are filtered by some channel (convolutive mixture model) as long as the training set signals undergo same filtering.

In the training stage we estimate the empirical probability density functions $\hat{p}\left(\hat{E}|H^{(1)}\right)$ and $\hat{p}\left(\hat{E}|H^{(2)}\right)$ using normalized histograms. Large non overlapping areas indicate that a separation of these signals using only $\hat{E}\left\{x\right\}$ values should be possible. Yilmaz et al. [9] defined approximate W-disjoint orthogonality (W-DO) as an approximate "disjointness" of several signals in the STFT domain. They introduced a quantitative WDO measure and provided evidence of the high level of the W-DO for several speech signals. Since the EFMS is a local TF property, the approximate W-DO of signals guarantees robust EFMS estimation in the mixture. We verify that speech and piano play signals have high value of WDO in Section VI.

In the separation stage we use $\hat{p}\left(\hat{E}|H^{(1)}\right)$ and $\hat{p}\left(\hat{E}|H^{(2)}\right)$ to define a minimum risk decision rule for classification of the STFT TF bins based on $\hat{E}\left\{x\right\}$. Let $\eta_k\left(m\right) \triangleq \frac{\hat{p}\left(\hat{E}\left\{x\right\}_k(m)|H^{(1)}\right)p\left(H^{(1)}\right)}{\hat{p}\left(\hat{E}\left\{x\right\}_k(m)|H^{(2)}\right)p\left(H^{(2)}\right)}$. The $p\left(H^{(1)}\right)$ and $p\left(H^{(2)}\right)$ reflect prior belief of either class to be present in a TF bin. Let λ_{ij} be a penalty for assigning a TF bin to class i when in fact the sample belongs to class j and λ_r is a penalty for not assigning a TF to neither class. Using Bayes risk minimization, the decision rule can be written as

$$\begin{aligned}
\mathcal{R}_1 &= \left\{\left(k,m\right)|\frac{\lambda_{12}}{\lambda_{21}} < \eta_k\left(m\right) \cap \frac{\lambda_r}{\lambda_{12}} > \frac{1}{1+\eta_k\left(m\right)}\right\} ,\\
\mathcal{R}_2 &= \left\{\left(k,m\right)|\frac{\lambda_{12}}{\lambda_{21}} > \eta_k\left(m\right) \cap \frac{\lambda_r}{\lambda_{12}} > \frac{1}{1+\eta_k\left(m\right)}\right\} ,\\
\mathcal{R}_r &= \left\{\left(k,m\right)|\frac{\lambda_r}{\lambda_{12}} \leq \frac{1}{1+\eta_k\left(m\right)} \cap \frac{\lambda_r}{\lambda_{21}} \leq \frac{1}{1+1/\eta_k\left(m\right)}\right\} ,
\end{aligned}$$

where $\mathcal{R}_1$ and $\mathcal{R}_2$ are sets of TF bins assigned to different audio classes and $\mathcal{R}_r$ is a set of rejected TF bins [10].

A binary mask in the STFT domain is defined as

$$M_k^{(c)}\left(m\right) = \left\{\begin{array}{ll} 1 & \gamma_k\left(m\right) \in \mathcal{R}_c \\ 0 & \text{otherwise} \end{array}\right. , c \in \left\{1,2\right\} . \quad (8)$$

For the binary mask to be effective, we assume that approximate W-disjoint orthogonality [9] holds. The interfering source is removed by multiplying the STFT transform of the mixture by $M^{(c)}$

$$\hat{X}_k^{(c)}\left(m\right) = M_k^{(c)}\left(m\right) X_k\left(m\right) . \quad (9)$$

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

$$\hat{x}^{(c)}\left(n\right) = \text{ISTFT}\left\{\hat{X}_k^{(c)}\left(m\right)\right\} . \quad (10)$$

VI. Experimental Results

In this section 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 seconds of speech (either male or female) taken from TIMIT database sampled at 16 KHz and Chopin's prelude for piano Opus 28 No. 6 for GMM training. We use 1024 points STFT transform,

Table I
SEPARATION PERFORMANCE ANALYSIS.

	SDR$_1$	SIR$_1$	SAR$_1$	LSD$_1$	SDR$_2$	SIR$_2$	SAR$_2$	LSD$_2$
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

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 WDO value [9] for the pair of signals used in our experiment is 0.94, which according to [9] guaranties perceptually perfect separation using "oracle" masks defined in [11]. 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 signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), signal-to-artifact ratio (SAR) [12] and log spectral distance (LSD) were used for the performance evaluation. The results are shown in Table I. A 0 dB mixture of test signals was used in all experiments. 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.

Smaller amounts of interfering signal is audible in signals recovered by the proposed method compared to the GMM based algorithm. The overall audio quality is also more plausible. The most disturbing artifact in the recovered piano signal is the missing piano note onsets. 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.

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 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 sec). A certain amount of musical noise is also present. Applying the algorithm to a clean piano play signal reveals that most of the leaking signal results from the piano hammer strikes. This conclusion was confirmed by examination of spectrograms of the recovered signals.

VII. 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.

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. [13].

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.

REFERENCES

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

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

[3] F. R. Bach and M. I. Jordan, "Blind one-microphone speech separation: A spectral learning approach," in *NIPS*, Vancouver, 2004.

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

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

[6] J. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr 1990, pp. 381–384 vol.1.

[7] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Signal Process.*, vol. 41, no. 10, pp. 3024–3051, Oct 1993.

[8] Y. Litvin, I. Cohen, and D. Chazan, "Monaural speech/music source separation using discrete energy separation algorithm," *submitted for publication*.

[9] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, Inc., 2001.

[11] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

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

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