SINGLE-CHANNEL SOURCE SEPARATION OF SPEECH AND MUSIC USING SHORT-TIME SPECTRAL KURTOSIS

Yevgeni Litvin[1]*, Israel Cohen*[1]*, and Jacob Benesty*[2]

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

[2] INRS-EMT
Universite du Quebec
Montreal, QC H5A 1K6, Canada
benesty@emt.inrs.ca

ABSTRACT

In this paper, the problem of blind monaural speech/music source separation is addressed using short-time spectral kurtosis (STSK). An estimator for STSK is introduced, and a source separation algorithm is formulated that relies on the spectral kurtosis differences of distinct signal classes. The performance of the proposed algorithm is evaluated on mixtures of speech signals and various types of music signals. The results are compared to those obtained by a competing monaural source separation algorithm, which is based on a Gaussian mixture model (GMM).

1. INTRODUCTION

High order statistics are frequently used in the task of multichannel source separation. In particular, kurtosis is used as a measure of non-Gaussianity of the recovered mixture components. Spectral kurtosis (SK) is a tool capable of locating non-Gaussian components including their location in the frequency domain. SK was first introduced by Dwyer [1]. He defined it as a kurtosis value of the real part of the STFT filterbank output. Antoni [2] introduced a different formalization of the SK by means of Wold-Cramér decomposition which gave a theoretical ground for the estimation of the SK of non-stationary processes. He also showed practical applications of his approach in the field of machine surveillance and diagnostics [3, 4]. Other applications of spectral kurtosis include SNR estimation in speech signals [5], denoising [6], and subterranean termite detection [7].

In this paper we show how the SK of a mixture relates to the SK of its components. We define the short time spectral kurtosis (STSK) as a time localized version of the SK. We define a simple STSK estimator and show its application in the task of speech and music monaural separation. The proposed algorithm uses STSK analysis to assign time-frequency bins of a mixture to the correct source. A binary mask is used to reject the interfering source in the STFT domain. In the experimental results we study the separation performance of the proposed algorithm on mixtures of speech and musical excerpts played by various instruments. We show improved performance of the proposed algorithm compared to a competing GMM based algorithm [8].

The remainder of this paper is structured as follows. In Section 2 we present the concept of SK. Section 3 extends the idea of SK to non-stationary signals. In Section 4 we describe a simple source separation algorithm based on the SK analysis. An experimental study is given in Section 5, followed by a short discussion in Section 6.

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

2. SPECTRAL KURTOSIS

In this section, we present the SK. We analytically evaluate the SK for some common probability distributions, and show how the SK of an instantaneous mixture relates to the SK of its components.

Let $x(n)$ be a real, discrete time, stationary random vector. Let X_k be its N points discrete Fourier transform (DFT) defined by:

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}, \quad (1)$$

where k is the frequency index. Due the circularity of X_k and following the reasoning in [9], the only way to define a spectral kurtosis for $x(n)$ that does not vanish is

$$\mathcal{K}_x(k) = \frac{\kappa_4\{X_k, X_k^*, X_k, X_k^*\}}{\left(\kappa_2\{X_k, X_k^*\}\right)^2}, \quad (2)$$

with κ_r being an r-th order cumulant. Using the circularity the definition can be simplified to:

$$\mathcal{K}_x(k) = \frac{\mathbb{E}\{|X_k|^4\}}{\left(\mathbb{E}\{|X_k|^2\}\right)^2} - 2. \quad (3)$$

Let $x_{\mathrm{WG}}(n)$ be a white Gaussian signal. Its DFT is a complex normally distributed vector. All cumulants of an order greater than 3 are zero for Gaussian and complex Gaussian random variables. By eq. (2) the SK of $x_{\mathrm{WG}}(n)$ is zero for all k.

Let a be an amplitude and m_0 a frequency index. Let $x_{\sin}(n) = ae^{j\left(2\pi\frac{m_0}{N}n+\varphi\right)}$. If $\varphi \sim U(0, 2\pi)$, $x_{\sin}(n)$ is a stationary process. We note that $\mathbb{E}\{|X_k|^4\} = \left(\mathbb{E}\{|X_k|^2\}\right)^2 = (Na)^4$. It follows that $\mathcal{K}_{x_{\sin}}(k) = -1$.

In this work we use the instantaneous mixture model:

$$x(n) = s_1(n) + s_2(n). \quad (4)$$

Assume that $s_1(n)$ and $s_2(n)$ are statistically independent stationary processes. Let $\phi_s(k) \triangleq \mathbb{E}\left(|S_k|^2\right)$ and $\gamma(k) \triangleq \phi_{s_1}(k)/\phi_{s_2}(k)$

where S_k is the k-th coefficient of DFT of s. Since S_k are circular processes, it can be shown that

$$\mathcal{K}_x(k) = \left|\frac{1}{1+1/\gamma(k)}\right|^2 \mathcal{K}_{s_1}(k) + \left|\frac{1}{1+\gamma(k)}\right|^2 \mathcal{K}_{s_2}(k). \quad (5)$$

When $\gamma(k) \gg 1$, a mixture SK is approximately the SK of the first component, i.e. $\mathcal{K}_x(k) \approx \mathcal{K}_{s_1}(k)$. Similarly, when $\gamma(k) \ll 1$, $\mathcal{K}_x(k) \approx \mathcal{K}_{s_2}(k)$.

2.1. Kurtosis Estimation

Let $\{X(i)\}_{i=1}^{L_K} \in \mathbb{R}^N$ denote a set of samples. In case $\{X(i)\}$ are i.i.d, Vrabie et al. [9] proposed the following unbiased estimator of the SK:

$$\hat{\mathcal{K}}_X = \frac{L_K}{L_K-1}\left(\frac{(L_K+1)\sum_{i=1}^{L_K}|X(i)|^4}{\left(\sum_{i=1}^{L_K}|X(i)|^2\right)^2} - 2\right). \quad (6)$$

Antoni [2] proposed another SK estimator assuming Wold-Cramér decomposition of non-stationary process. It is based on the STFT transform and requires the analyzed signal to be quasi-stationary at the scale of the STFT analysis windows. The estimator is defined using $2n$-th moment empirical estimator:

$$\hat{S}_{2nX}(k) \triangleq \left\langle |X_k(m)|^{2n}\right\rangle_m, \quad (7)$$

where $X_k(m)$ is the STFT transform of $x(n)$, m and k are the time and frequency indices, respectively, and $\langle . \rangle_t$ is the averaging operator with respect to t. The STFT based estimator of the SK is defined as

$$\hat{\mathcal{K}}_X(k) \triangleq \frac{\hat{S}_{4X}(k)}{\hat{S}_{2X}^2(k)} - 2. \quad (8)$$

The analysis of the statistical properties of this estimator can be found in [2].

The SK estimator (8) has no time localization. In order for the SK analysis be applicable to speech or music processing, we propose to localize the SK estimation in time. We define a time localized $2n$-th order empirical spectral moment of $|X_k(m)|$ as:

$$\hat{S}_{2nX,k}(m) \triangleq \sum_{i=-\lfloor L_K/2\rfloor}^{\lfloor L_K/2\rfloor} w_{\mathcal{K}}(m+i)\,|X_k(i)|^{2n}, \quad (9)$$

where $w_{\mathcal{K}}(m)$ is an averaging window and $\sum_m w_{\mathcal{K}}(m) = 1$. Definitions (7) and (9) are identical except for the time localization in (9). Finally we define:

$$\hat{\mathcal{K}}_{X,k}(m) \triangleq \frac{\hat{S}_{4X,k}(m)}{\hat{S}_{2X,k}^2(m)} - 2. \quad (10)$$

We note that this estimator is biased and its bias depends on the overlap of the STFT analysis windows. In the rest of this paper we refer to (10) as the short time spectrum kurtosis (STSK) estimator. The set of samples used for the estimation of a single STSK value is referred to as the STSK estimation window.

Loosely speaking, the relation between the STSK and SK is similar to the relation between Fourier transform and the STFT.

2.2. Physical Interpretation

Following [2], (7) can be written as

$$\hat{\mathcal{K}}_X(k) \triangleq \frac{\left\langle |X_k(m)|^4\right\rangle_m - \left\langle |X_k(m)|^2\right\rangle_m^2}{\left\langle |X_k(m)|^2\right\rangle_m^2} - 1. \quad (11)$$

This expression can be interpreted as a normalized empirical variance of the signal energy in different frequency bands (up to a subtracted constant -1) which is as a measure for the time dispersion of $|X_k|^2$. Similar interpretation can be applied to the STSK.

3. SHORT TIME SPECTRAL KURTOSIS OF NATURAL AUDIO SIGNALS

In this section, we demonstrate an STSK analysis of speech and piano play signals.

In the following examples we use audio signals sampled at 16 KHz. The STFT analysis is performed using $N = 1024$, $M = 128$. We set $L_K = 31$ and $w_{\mathcal{K}}$ to be a rectangular window

$$w_{\mathcal{K}}(m) = \begin{cases} 1/L_K, & -\lfloor L_K/2\rfloor \le m \le \lfloor L_K/2\rfloor \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

i.e. a single STSK value is estimated from a bandpass signal within a time span window of roughly 1/4 second.

Figure 1 shows the STFT and STSK of speech, piano play and their mixture respectively. We observe that low values dominate the STSK of the piano play signal and high values dominate the speech STSK.

When the STSK estimation window contains a speech phoneme onset or offset, plosive phoneme or a musical note onset or offset, the time dispersion of the energy is high (within that window), hence the STSK value is also high. The STSK values in the time-frequency regions that contain fricative phonemes are approximately zero since they are well modeled by a complex Gaussian noise.

An harmonic partial in the STFT domain is well modeled by a complex sine. The STSK value will be close to -1 if the STSK estimation window contains an harmonic partial. In most cases, the pitch rate of a natural speech changes constantly. If the STSK estimation window is long enough, it is likely to accommodate the inclusion or exclusion of an harmonic partial in the STFT frequency band. As a result, high values of the STSK will be estimated in time-frequency regions where voiced phonemes are present.

The W-DO property in an approximate sense [10] implies that it is unlikely that the large STFT coefficients coincide in the same time-frequency bins. We assume that the W-DO property in an approximate sense holds for speech and music mixtures. The W-DO property implies that $\gamma(k) \gg 1$ or $\gamma(k) \ll 1$ holds for almost all time-frequency bins. Using (5) we conclude that each time-frequency bin can be assigned to the correct source by testing its STSK value. This allows us to formulate a source separation algorithm as described next.

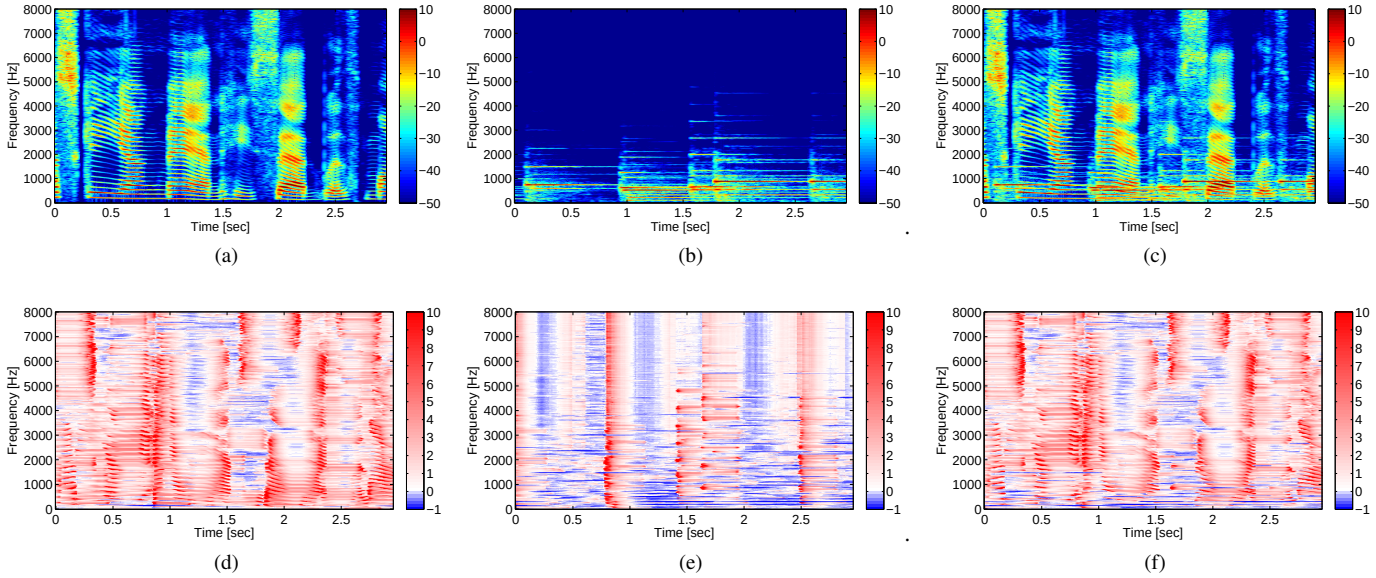

Fig. 1: Power spectrum (upper row) and STSK analysis (lower row) of (a),(d) a speech signal; (b),(e) a piano play signal; (c),(f) a speech and piano play mixture.

4. SOURCE SEPARATION USING STSK

In the previous section we observed that the STSK values of a speech signal are generally higher than the STSK values of a piano play. We also saw that pitch tracks of a piano play have low STSK values. This motivates us to define the following binary masks in the STFT domain:

$$M_k^{(1)}(m) = \begin{cases} 1 & \hat{\mathcal{K}}_{x,k}(m) > \delta_{\mathrm{SK}} \\ 0 & \text{otherwise} \end{cases}, \quad (13)$$

$$M_k^{(2)}(m) = 1 - M_{1,k}(m), \quad (14)$$

where δ_{SK} is a threshold chosen manually.

We reconstruct the mixture components by masking the interfering signal in the time-frequency domain followed by inverse short time Fourier (ISTFT) transform

$$\hat{S}_{c,k}(m) = M_k^{(c)} X_k(m), \quad (15)$$

$$\hat{s}_c(n) = \mathrm{ISTFT}\left(\hat{S}_c\right). \quad (16)$$

5. EXPERIMENTAL RESULTS

In this section we describe computer simulation and informal listening test results. We compare the performance of the proposed algorithm to a GMM monaural separation algorithm [8].

The GMM algorithm requires training sequences. We use the first minute of the test sequence file for training. The performances of both algorithms are evaluated on the second minute of the same mixture. Speech excerpts are taken from the TIMIT database are sampled at 16 KHz. Musical excerpts are taken from free Internet sources and downsampled to 16 KHz. We used a GMM model of order 26.

Table 1: Algorithm Parameters.

Sampling frequency	16KHz
STFT analysis window length	1024
STFT overlap (samples)	128
Short time SK estimation window length (samples/secs)	71 ($\sim$0.5 sec)
δ_{SK}	1

An instantaneous mixture of signals was obtained by adding two signals in the computer program. No noise was added to the mixture. The signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and signal-to-artifact ratio (SAR) measures [11] were used for the performance evaluation. The energies of the speech and piano signals were normalized prior to mixing, hence the SDR and SIR of the mixture are approximately zero and the SAR is very high.

The parameters of the proposed algorithm are shown in Table 1. The value of δ_{SK} was chosen experimentally in a way that makes the SDR of both extracted sources approximately equal.

Table 2 compares the performance of the proposed algorithm evaluated on speech signal mixed with different musical excerpts. The same table displays the performance of the GMM based separation algorithm. In all cases the STSK based algorithm outperforms the GMM based algorithm.

The objective measures show different quality of separation for different types of musical excerpts. Wind quartet is separated best, followed by piano solo, guitar solo, and orchestra sequences.

Listening to the extracted speech and music components reveals that the GMM algorithm produces a "jumpy" signal. This is the result of GMM state switching between consequent frames. The STSK based algorithm results in a more fluent and natural sound. Higher amount of residua and distortion is audible in the

Table 2: Comparison of the STSK-Based and the GMM-Based Separation Algorithms.

	Wind quartet		Piano solo		Guitar solo		Orchestra	
	STSK	GMM	STSK	GMM	STSK	GMM	STSK	GMM
SDR_1	8.9	2.2	6.4	2.4	5.7	−5.6	4.7	−5.2
SIR_1	29.4	6.3	17.2	8.3	15.5	20.5	13.3	19.4
SAR_1	8.9	5.3	6.9	4.4	6.4	−5.5	5.5	−5.2
SDR_2	9	1.6	6.6	2.6	5.7	0.9	4.7	0.9
SIR_2	17.8	8.8	16.9	7.6	15.5	1.6	13.8	1.8
SAR_2	9.7	3.1	7.2	4.9	6.3	11.2	5.4	10.3

signals extracted by the GMM algorithm.

The music residua found in the speech component extracted by the STSK based algorithm has non-harmonic nature, such as onsets of piano notes (caused by a felt covered hammer striking the strings), pluck sound in the guitar play and percussive instruments in the orchestra sequence. Speech residua in the extracted musical component contains traces of speech voiced phonemes.

All audio excerpts and the separation results used in this paper can be downloaded from `http://sipl.technion.ac.il/~elitvin/SK`.

6. DISCUSSION

The proposed algorithm fails to distinguish between non-harmonic components in music and speech since they both have relatively high values of the STSK, thus musical sources that contain less non-harmonic components (e.g. wind quartet) are separated to a greater extent than those with higher amounts of non-harmonic components (e.g. percussive instruments in the orchestra).

The fact that the proposed algorithm operates in time-frequency localized regions allows it to perform decently well in the case when the learning of the spectral shapes of a signal fails (e.g. orchestra, guitar solo). Perhaps, the combination of localized time-frequency and spectral information could further improve the separation performance.

7. CONCLUSIONS

Recent work on the spectral kurtosis mainly focused on machine surveillance and diagnostics. The value of the SK was estimated using large sets of independent samples. In our work we estimate time-localized SK values. We have defined a short time spectral kurtosis (STSK) and its ad-hoc estimator. The STSK should be seen as a frequency localized temporal feature. Each STSK value is estimated in a narrow band signal, hence the STSK carries information that is orthogonal to the spectral information.

We demonstrated the application of the STSK to monaural speech/music separation. In our experimental study we found that an algorithm that uses the STSK feature showed surprisingly good performance in separating instantaneous mixtures of speech and various musical excerpts. A convolutive mix scenario does not pose a problem for the proposed algorithm as long as the STSK values of two signals convolved with a respective channel remain significantly different.

The STSK can also serve as an additional feature in various audio processing tasks, such as signal enhancement and signal classification. A further study of the statistical properties of the STSK estimator is a subject for future research.

8. REFERENCES

[1] R. Dwyer, "Detection of non-gaussian signals by frequency domain kurtosis estimation," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-83*, vol. 8, 1983, pp. 607–610.

[2] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 282 – 307, 2006.

[3] J. Antoni and R. Randall, "The spectral kurtosis: application to the vibratory surveillance and diagnostics of rotating machines," *Mechanical Systems and Signal Processing*, vol. 20, no. 2, pp. 308 – 331, 2006.

[4] J. Antoni, "Fast computation of the kurtogram for the detection of transient faults," *Mechanical Systems and Signal Processing*, vol. 21, pp. 108–124, Jan. 2007.

[5] E. Nemer, R. Goubran, and S. Mahmoud, "SNR estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Process. Lett.*, vol. 6, no. 7, pp. 171–174, Jul. 1999.

[6] P. Ravier and P. O. Amblard, "Denoising using wavelet packets and the kurtosis: application to transient detection," in *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, 6–9 Oct. 1998, pp. 625–628.

[7] J. J. G. de la Rosa, C. G. Puntonet, and A. Moreno, "Subterranean termite detection using the spectral kurtosis," in *Proc. 4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications IDAACS 2007*, 2007, pp. 351–354.

[8] L. Benaroya and F. Bimbot, "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)*, Nara, Japan, Apr. 2003, pp. 957–961.

[9] V. Vrabie, P. Granjon, and C. Servière, "Spectral kurtosis: from definition to application," in *IEEE-EURASIP International Workshop on Nonlinear Signal and Image Processing*, Grado, Italy, 2003.

[10] 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.

[11] 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.