Monaural Source Separation

Y. Litvin

Department of Electrical Engineering

Supervised by Prof. Israel Cohen and Dr. Dan Chazan
Outline

1. Introduction
2. Subband Frequency Modulating Signal Modeling
3. Spectral Kurtosis
4. Bark-Scaled WPD
5. Conclusion
Outline

1. Introduction
   - Blind Source Separation
   - GMM Based Source Separation Algorithm
   - Distortion Measures

2. Subband Frequency Modulating Signal Modeling

3. Spectral Kurtosis

4. Bark-Scaled WPD

5. Conclusion
Blind Source Separation

Definition

Task of recovering a set of signals from a set of observed signal mixtures

- Number of sources
- Mixing model (instantaneous, echoic, convolutive, linear, non-linear)
- Number of observed mixtures
- Noise presence
- Training database
Problem Formulation

- Problem setup: single observation, two audio sources (speech and background music), no noise

\[ x(n) = s_1(n) + s_2(n) \]

- In the STFT domain (benefits: low inter-band correlation, sparse representation, binary masks)

\[ X_k(m) = S_{1,k}(m) + S_{2,k}(m) \]
Previous Work

Multichannel

(Comon, 1994)
Find demixing matrix by minimizing some measure of statistical independence (ICA).

(Zibulevsky & Pearlmutter, 2001)
Find demixing matrix by minimizing some measure of sparsity.
Previous Work

Single channel

(Hanson & Wong, 1984)
Estimate pitch of one of the talkers. Used harmonic information and spectral subtraction to suppress the other.

(Bach & Jordan, 2005)
Define distances between each T-F bins using CASA principles. Use clustering to group similar T-F bins together. Apply binary masking in the T-F domain.
Previous Work

Single channel

(Virtanen, 2003)

Use Non-negative Matrix Factorization to factor spectral magnitude matrix into frequency basis vectors and amplitudes: $\mathbf{X} \approx \mathbf{A}\mathbf{S}$. Cluster frequency basis vectors (columns of $\mathbf{A}$) and recreate mixture components using its frequency basis.
Wiener Based BSS Using GMM

Signal Model

- Introduced in (Benaroya & Bimbot, 2003)
- Mixture components are
  \[ s_1 \sim N(0, \Sigma_1); s_2 \sim N(0, \Sigma_2) \]
- Observed signal is
  \[ x = s_1 + s_2 \]
- Posterior Mean (PM) estimator for \( s_1(n) \) is
  \[ \hat{s}_1 = \Sigma_1 (\Sigma_1 + \Sigma_2)^{-1} x \]
Assume $s_1, s_2$ are stationary and approximately circular then Fourier transform $\mathcal{F}$ diagonalizes covariance matrix.

Denote $X \triangleq \mathcal{F} x, S_1 \triangleq \mathcal{F} s_1, S_2 \triangleq \mathcal{F} s_2$

$$S_1 \sim \mathcal{N}(0, \text{diag}(\sigma_1^2)); S_2 \sim \mathcal{N}(0, \text{diag}(\sigma_2^2))$$

$$X = S_1 + S_2$$

PM estimator for the case of vectors with diagonal covariance matrix

$$\hat{S}_1(i) = \frac{\sigma_1^2(i)}{\sigma_1^2(i) + \sigma_2^2(i)} X(i)$$
Wiener Based BSS Using GMM
Gaussian Mixture Model

- Assume $K$ Gaussian distributions $\{\mu^{(k)}, \Sigma^{(k)}\}_{k=1}^{K}$
- Probability of selecting $k$-th distribution is $\omega_k$ ($\sum_{k=1}^{K} \omega_k = 1$)
- GMM model defined by $\Lambda = \{\omega_k, \mu^{(k)}, \Sigma^{(k)}\}_{k=1}^{K}$
Wiener Based BSS Using GMM

- Assume $S_c(m)$ generated by $\Lambda_c$ ($c \in \{1, 2\}$ class index)
- Introduce hidden variables $q_c \in \{1, \ldots, K\}$
- Define posterior probability $\gamma_{j,k} = p(q_1 = j, q_2 = k | X)$
- When conditioned on $q_1, q_2$, mixture components $S_c \sim N(\mu^{(q_c)}, \Sigma^{(q_c)})$ and we may use PM

$$\hat{S}_1(i) = \sum_{i,j} \gamma_{i,j} \frac{\sigma_1^{(i)} (i)}{\sigma_1^{(i)} (i) + \sigma_2^{(j)} (i)} X(i)$$

- $\gamma_{j,k}$ estimated from mixture observation by exhaustive enumeration of $j, k \in \{1, \ldots, K\}$

$$\gamma_{i,j} \propto p(X|q_1 = j, q_2 = k) p(q_1 = j) p(q_2 = k)$$

$$= g(X; \Sigma_1^{(j)} + \Sigma_2^{(k)}) w_1^{(j)} w_2^{(k)}$$
Wiener Based BSS Using GMM

Separation
Wiener Based BSS Using GMM

Y. Litvin

Monaural Source Separation
Wiener Based BSS Using GMM

\[ \gamma_{i,j} \text{ is small} \]
Wiener Based BSS Using GMM

\[ \gamma_{i,j} \text{ is small} \]
\[ \gamma_{i,j} \text{ is large} \]
Distortion Measures

- $s_c$ the desired source, $s_{c'}$ the interfering source (Gribonval et al., 2003)

$$\hat{s}_c = y_c + e_{c,\text{interf}} + e_{c,\text{artif}}$$

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

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

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

<table>
<thead>
<tr>
<th>Distortion Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal to Distortion Ratio (SDR)</td>
<td>$10 \log_{10} \frac{|y_c|^2}{|e_{c,\text{interf}} + e_{c,\text{artif}}|^2}$</td>
</tr>
<tr>
<td>Signal to Interference Ratio (SIR)</td>
<td>$10 \log_{10} \frac{|y_c|^2}{|e_{c,\text{interf}}|^2}$</td>
</tr>
<tr>
<td>Signal to Artifact Ratio (SAR)</td>
<td>$10 \log_{10} \frac{|y_c + e_{c,\text{interf}}|^2}{|e_{c,\text{artif}}|^2}$</td>
</tr>
</tbody>
</table>
Outline

1. Introduction

2. Subband Frequency Modulating Signal Modeling
   - Motivation
   - AM-FM Demodulation using DESA
   - Energy of Frequency Modulating Signal
   - Source Separation Algorithm
   - Experimental Results

3. Spectral Kurtosis

4. Bark-Scaled WPD

5. Conclusion
Motivation

- Pitch track behavior is very different for speech and some “mechanically” generated sounds (e.g., music).
- Easily detected by examining the unwrapped phase of the subband signal.
Teager’s Energy Tracking Operator

- Undriven linear undamped oscillator with an amplitude $A$

\[
E_{osc} = \frac{1}{2} m\dot{x}_c^2 + \frac{1}{2} kx_c^2 = \frac{1}{2} m(A\omega_0)^2
\]
\[
\omega_o = \sqrt{k/m}
\]

- **Teager Energy Operator** (Teager & Teager, 1985)

\[
\Psi_c[x(t)] = (\dot{x}(t))^2 - x(t)\ddot{x}(t)
\]

Body position

\[
x(t) = A\cos(\omega_0 t + \theta)
\]
\[
\Psi_c[x(t)] = 2A^2\omega_0^2
\]

Approximately holds also for $A(t)$ and $\omega_0(t)$ (Maragos et al., 1993)

\[
x(t) \approx A(t)\cos(\omega_0(t) t + \theta)
\]
\[
\Psi_c[x(t)] \approx 2A(t)^2\omega_0(t)^2
\]
Energy Separation Algorithm (ESA)

- Continuous Energy Separation Algorithm (ESA) (Maragos et al., 1993)

\[
\omega_0(t) \approx \sqrt{\frac{\psi_c[\dot{x}(t)]}{\psi_c[x(t)]}}
\]

\[
|A(t)| \approx \frac{\psi_c[x(t)]}{\sqrt{\psi_c[\dot{x}(t)]}}
\]
Discrete ESA (DESA)

- Discrete TEO

\[ \Psi[x(n)] = x^2(n) - x(n-1)x(n+1) \]

- Discrete ESA (DESA)

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

\[ |\hat{a}(n)| \approx \frac{2\Psi[x(n)]}{\sqrt{\Psi[x(n+1) - x(n-1)]}} \]
Let $x_\ell$ be $\ell$-th harmonic partial. Assume AM-FM model.

$$x(n) = \sum_{\ell} a_\ell(n) \cos \left( \Omega_0 \ell n + \sum_{i=0}^{n} r(i) \ell \frac{1}{T} + \theta_\ell \right)$$
Energy of Frequency Modulating Signal
STFT subband

- At the output of the STFT filterbank

\[ X_k(m) \approx a(mM) e^{i(\tilde{\Omega}_c mM + \sum_{i=0}^{mM} r(i) \frac{1}{T})} \]

\[ \tilde{\Omega}_c = \Omega_c - \frac{2\pi}{N} k \]

Graph showing the frequency response with Amplitude on the y-axis and Frequency [Hz] on the x-axis.
Energy of Frequency Modulating Signal

**Intermediate Frequency**

- Modulate subband to some intermediate frequency $\Omega_{if} = \alpha \pi$, $0 < \alpha < 1$

$$\tilde{X}_k(m) = \Re(X_k(m)e^{j\Omega_{if}m}) = a(mM)\cos\left((\tilde{\Omega}_c + \Omega_{if})mM + \sum_{i=0}^{mM} r(i) \frac{1}{T}\right)$$
Energy of Frequency Modulating Signal

DESA

- Estimate instantaneous frequency using DESA

\[ \hat{\Omega}_{i,k}(m) \approx \frac{1}{2} \arccos \left( 1 - \frac{\psi \left[ \tilde{X}_k(m+1) - \tilde{X}_k(m-1) \right]}{2\psi \left[ \tilde{X}_k(m) \right]} \right) \]

\[ = \left( \tilde{\Omega}_c + \Omega_{if} \right) M + r( mM ) \frac{1}{T} \]

- Constant term is removed using high-pass filter \( h_r \) and Energy of Frequency Modulating Signal is obtained by smoothing \( r^2 (mM) \) using a Hamming window \( u(m) \) of length \( N_u \)

- Upper bound on \( M \leq \min \{\alpha N, (1 - \alpha) N\} \) (due to DESA assumption on signal bandwidth)
Energy of Frequency Modulating Signal

Block Diagram

\[ x(n) \xrightarrow{w_{a,k}(n)} x_a(n) \xrightarrow{e^{-j2\pi \frac{kn}{N}}} X_k(m) \xrightarrow{e^{j\Omega_0 m}} \tilde{X}_k(m) \]

\[ \xrightarrow{\text{STFT}} \]

\[ \xrightarrow{\text{DES A}} \tilde{\Omega}_{i,k}(m) \xrightarrow{h_q(m)} \hat{r}(m) \xrightarrow{(\cdot)^2} u(m) \xrightarrow{\hat{E}_k(m)} \]
EFMS of Real Audio Signals
EFMS of Real Audio Signals
Probability distribution of EFMS

![Graph showing empirical probability distribution of EFMS for Speech and Piano]

- **Speech**
- **Piano**

Y. Litvin
Monaural Source Separation
Source Separation Procedure

Classification

- $\xi$ - EFMS value
- $\lambda_{ij}$ - penalty for assigning a sample $\xi$ to class $i$ when in fact the sample belongs to class $j$
- $\lambda_r$ - penalty for rejecting a sample
- We look for regions $R_1, R_2, R_r$ that minimize loss function

$$L = \int_{R_1} \lambda_{12} p \left( H^{(2)} | \xi' \right) p (\xi') d\xi' +$$

$$+ \int_{R_2} \lambda_{21} p \left( H^{(1)} | \xi' \right) p (\xi') d\xi + \int_{R_r} \lambda_r p (\xi') d\xi'$$
Source Separation Procedure

Classification

Let \( \eta \triangleq \frac{p(\xi|H(1))p(H(1))}{p(\xi|H(2))p(H(2))} \)

Decision rules are

\[
\begin{align*}
\xi \in R_1 & \iff \left\{ \frac{\lambda_{12}}{\lambda_{21}} < \eta, \frac{\lambda_r}{\lambda_{12}} > \frac{1}{1+\eta} \right\} \\
\xi \in R_2 & \iff \left\{ \frac{\lambda_{12}}{\lambda_{21}} > \eta, \frac{\lambda_r}{\lambda_{21}} > \frac{1}{1+1/\eta} \right\} \\
\xi \in R_r & \iff \left\{ \frac{\lambda_r}{\lambda_{12}} \leq \frac{1}{1+\eta}, \frac{\lambda_r}{\lambda_{21}} \leq \frac{1}{1+1/\eta} \right\}
\end{align*}
\]
Source Separation Procedure

**Masking**

- Define binary mask for class $c \in \{1, 2\}$

  $$M_{k}^{(c)}(m) = \begin{cases} 1 & \xi_k(m) \in \mathcal{R}_c \\ 0 & \text{otherwise} \end{cases}$$

- Obtain masked STFT domain signal

  $$\hat{X}_{k}^{(c)}(m) = M_{k}^{(c)}(m) X_k(m)$$

- Back to the time domain

  $$\hat{x}^{(c)}(n) = \text{ISTFT}\left\{\hat{X}_{k}^{(c)}(m)\right\}$$
Experimental Results

Synthetic signals

\[ s_c(n) = \sum_{\ell=0}^{N_h} \cos(\ell \cdot 2\pi f_{c}^{(c)} n/f_s + \sum_{m=0}^{n} q_{\ell}^{(c)}(m) \frac{1}{T}) \]

\[ q_{\ell}^{(c)}(n) = \ell \cdot d^{(c)} \cos(2\pi f_{m}^{(c)} n/f_s) \quad d^{(1)} = 20, d^{(2)} = 1 \]
Experimental Results

Synthetic signals

- $s_1$ white noise, $s_2$ as before

![Graph of experimental results](image-url)
Experimental Results

Real signals

- Demo...
- \( N = 1024, \; M = 64, \; N_u = 121, \; \delta_E = 15\text{dB}, \; \lambda_{12} = \lambda_{21} = 1, \; \lambda_r = \infty, \; \alpha = 1/3 \)
- Oracle masks

\[
\tilde{M}^{(c)}_k(m) = \begin{cases} 
1 & |S_{1,k}(m)| \leq |S_{2,k}(m)| \\
0 & \text{otherwise}
\end{cases}
\]

<table>
<thead>
<tr>
<th></th>
<th>( \text{SDR}_1 )</th>
<th>( \text{SIR}_1 )</th>
<th>( \text{SAR}_1 )</th>
<th>( \text{LSD}_1 )</th>
<th>( \text{SDR}_2 )</th>
<th>( \text{SIR}_2 )</th>
<th>( \text{SAR}_2 )</th>
<th>( \text{LSD}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle mask</td>
<td>18.9</td>
<td>42.6</td>
<td>18.9</td>
<td>0.73</td>
<td>17.9</td>
<td>47.2</td>
<td>18.0</td>
<td>0.8</td>
</tr>
<tr>
<td>EFMS</td>
<td>6.1</td>
<td>11.8</td>
<td>7.8</td>
<td>2.4</td>
<td>6.4</td>
<td>19.7</td>
<td>6.6</td>
<td>2.0</td>
</tr>
<tr>
<td>GMM</td>
<td>2.4</td>
<td>9.3</td>
<td>3.8</td>
<td>2.9</td>
<td>2.6</td>
<td>7.9</td>
<td>4.8</td>
<td>2.5</td>
</tr>
</tbody>
</table>
Experimental Results
Real signals - Speech
Experimental Results
Real signals - Piano

GMM

EFMS

Y. Litvin
Monaural Source Separation
Experimental Results
Real signals - Residual signal

- Speech
- Piano
Outline

1. Introduction
2. Subband Frequency Modulating Signal Modeling
3. Spectral Kurtosis
   - Kurtosis
   - SK of Audio Signals
   - Separation Algorithm
   - Experimental Results
4. Bark-Scaled WPD
5. Conclusion
Real Kurtosis

- Measure of peakedness
- Kurtosis definition

\[
\text{Kurt}(X) = \frac{\kappa_4}{\kappa_2^2} = \frac{\mathbb{E}(X^4)}{\mathbb{E}(X^2)^2} - 3
\]

- \(X \sim N(\mu, \sigma^2) \Rightarrow \text{Kurt}(X) = 0\)
- \(X \sim \text{Laplace}(\mu, b) \Rightarrow \text{Kurt}(X) = 3\)

- Cumulant generating function of r.v. \(X\)

\[
g(t) = \log \mathbb{E}(e^{tX})
\]

- \(k\)-th cumulant is given by

\[
\kappa_k = \frac{d^k}{dt^k} g \bigg|_{t=0}
\]
Real Kurtosis

Peakiness of various distributions
Spectral Kurtosis Definition

- Let \( x(n) \) be a time domain signal, \( X_k \) the \( k \)-th coefficient of DFT and \( X_k^* \) its complex conjugate. SK \( \mathcal{K}_x \) is defined by (Vrabie et al., 2003)

\[
\mathcal{K}_x(k) = \frac{\kappa \{ X_k, X_k^*, X_k^*, X_k^* \}}{\left( \kappa \{ X_k, X_k^* \} \right)^2}
\]

- For circular processes

\[
\mathcal{K}_x(k) = \frac{\mathbb{E} \left\{ |X_k|^4 \right\}}{\left( \mathbb{E} \left\{ |X_k|^2 \right\} \right)^2} - 2
\]

- \( X \sim N(\mu, \sigma^2) \) \( \Rightarrow \) Kurt(\( X \)) = 0
- \( X = e^{j\Omega n + \theta}, \theta \sim U[0, 2\pi] \) \( \Rightarrow \) Kurt(\( X \)) = -1
Spectral Kurtosis of a Mixture

- Let \( \phi_A(k) \triangleq \mathbb{E}(|A_k|^2) \)
- Let \( \gamma \triangleq \phi_{s1}(k)/\phi_{s2}(k) \)

\[
\mathcal{K}_x(k) = \left| \frac{\phi_{s1}(k)}{\phi_{s1}(k) + \phi_{s2}(k)} \right|^2 \mathcal{K}_{s1}(k) + \left| \frac{\phi_{s2}(k)}{\phi_{s1}(k) + \phi_{s2}(k)} \right|^2 \mathcal{K}_{s2}(k)
\]

\[
= \left| \frac{1}{1 + 1/\gamma} \right|^2 \mathcal{K}_{s1}(k) + \left| \frac{1}{1 + \gamma} \right|^2 \mathcal{K}_{s2}(k)
\]

(Benesty, 2009)

- \( \phi_{s1}(k) \gg \phi_{s2}(k) \Rightarrow \gamma \gg 1 \Rightarrow \mathcal{K}_x(k) \approx \mathcal{K}_{s1}(k) \)
- \( \phi_{s1}(k) \ll \phi_{s2}(k) \Rightarrow \gamma \ll 1 \Rightarrow \mathcal{K}_x(k) \approx \mathcal{K}_{s2}(k) \)
- From W-DO, for each TF bin \( \gamma \ll 1 \) or \( \gamma \gg 1 \)
SK Estimation

- Let $X_k (m)$ be $k$-th frequency band of the STFT filterbank
- Assume $X_k (m)$ quasi-stationary i.i.d. process
- $L$ number of samples
- Spectral Kurtosis unbiased estimator (Vrabie et al., 2003)

\[
\hat{K}_x (k) = \frac{L}{L-1} \left[ \frac{(L+1) \sum_{i=1}^{L} |X_k (i)|^4}{\left( \sum_{i=1}^{M} |X_k (i)|^2 \right)^2} - 2 \right]
\]
Physical Interpretation

- Let $Y_k(m) = |X_k(m)|^2$
- SK can be rewritten as (Antoni, 2006)

$$\hat{K}_X(k) \triangleq \left[ \frac{\langle Y^2 \rangle_m - \langle Y \rangle_m^2}{\langle Y \rangle_m^2} \right] - 1$$

- Can be seen as normalized empirical variance with respect to time
- Similar to SK up to a constant additive element
STSK Estimation

- Short Time Spectral Kurtosis (STSK) is a SK localized in time.
- Let be 2n-th moment estimator

\[
\hat{S}_{2nX,k}(m) \triangleq \sum_{i=-[L_K/2]}^{[L_K/2]} w_{\mathcal{X}}(m+i)|X_k(i)|^{2n}
\]

- We define STSK

\[
\hat{\mathcal{K}}_{X,k}(m) \triangleq \frac{\hat{S}_{4X,k}(m)}{\hat{S}_{2X,k}(m)} - 2
\]
Speech and Piano Spectrograms
Speech and Piano STSK
Piano play (fast), Mix Spectrograms
Piano play (fast), Mix STSK
Separation

- Mask out time-frequency bins that belong to the interfering signal

\[
M_{1,k}(m) = \begin{cases} 
1 & \hat{K}_x(m,k) > \delta_1 \\
0 & \text{otherwise} 
\end{cases}
\]

\[
M_{2,k}(m) = \begin{cases} 
1 & \hat{K}_x(m,k) \leq \delta_2 \\
0 & \text{otherwise} 
\end{cases}
\]

- Recover desired signal (\(\circ\) element-wise multiplication)

\[
\hat{s}_c(n) = \text{ISTFT}(M_c \circ X)
\]
## Experimental Results

- Demo ...

<table>
<thead>
<tr>
<th>Method</th>
<th>SDR(^1)</th>
<th>SIR(^1)</th>
<th>SAR(^1)</th>
<th>LSD(^1)</th>
<th>SDR(^2)</th>
<th>SIR(^2)</th>
<th>SAR(^2)</th>
<th>LSD(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle mask</td>
<td>18.9</td>
<td>42.6</td>
<td>18.9</td>
<td>0.73</td>
<td>17.9</td>
<td>47.2</td>
<td>18.0</td>
<td>0.8</td>
</tr>
<tr>
<td>GMM</td>
<td>2.4</td>
<td>9.3</td>
<td>3.8</td>
<td>2.9</td>
<td>2.6</td>
<td>7.9</td>
<td>4.8</td>
<td>2.5</td>
</tr>
<tr>
<td>STSK</td>
<td>7.7</td>
<td>19.5</td>
<td>8.0</td>
<td>2.8</td>
<td>8.2</td>
<td>21.3</td>
<td>8.4</td>
<td>2.3</td>
</tr>
<tr>
<td>EFMS</td>
<td>6.1</td>
<td>11.8</td>
<td>7.8</td>
<td>2.4</td>
<td>6.4</td>
<td>19.7</td>
<td>6.6</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. Subband Frequency Modulating Signal Modeling
3. Spectral Kurtosis
4. Bark-Scaled WPD
   - Algorithm
   - Separation Algorithm
   - Experimental Results
5. Conclusion
Wavelet Packet Decomposition

- Discrete Wavelet Transform
- Wavelet Packet Decomposition

Y. Litvin
Monaural Source Separation
Bark scale

- Basilar membrane acts as non-uniform filterbank
- Accounts for non-uniform frequency sensitivity of human ear
- Bark scale follows center frequencies of critical bands (1 Bark apart)
- Frequency to Bark scale:
  \[ z = \frac{26.81}{1 + 1980/f} - 0.53 \]

Let \( L \) be depth of the WPD tree and \( 0 \leq l < L, 0 \leq n < 2^l \)

Center frequency of WPD node \((l, n)\) is

\[ f_{l,n} = 2^{-l} \left( GC^{-1}(n) + 0.5 \right) \frac{F_s}{2} \]
Bark-Scaled WPD

- Bark-Scaled WPD (BS-WPD) introduced in (Cohen, 2001)
- WPD with center frequencies located 1-Bark apart
- Critical band structured filterbank:
  - fine frequency resolution at low frequencies
  - coarse frequency resolution at high frequencies
- Various wavelet families may be used
- Improved frequency resolution by additional levels of decomposition
Constant Sampling Rate BS-WPD

- BS-WPD has different sampling rates at terminal nodes
- Stop decimating for nodes deeper than 6
- Total of 168 frequency bands comparing to 512 for STFT with similar bandwidth at low frequency bands
Mapping Based Complex Wavelet Transform

- DWT/WPD lack shift invariance
  - Two time domain signals $x(n), x_\Delta(n) = x(n - \Delta)$, small $\Delta$
  - Let $X_{l,n}(m), X_{\Delta,l,n}(m)$ be $(l,n)$ terminal node of DWT
  - $X_{l,n}(m)$ is significantly different from $X_{\Delta,l,n}(m)$
  - STFT transform: $\Delta$ mostly has influence on phase

- Reason: decimation in the decomposition tree
Mapping based Complex Wavelet Transform

- Introduced by (Fernandes et al., 2003)
- Achieves “approximate shiftability”
- Hardy-space $H^2 (\mathbb{R} \rightarrow \mathbb{C})$ is defined by
  \[
  H^2 (\mathbb{R} \rightarrow \mathbb{C}) \triangleq \{ f \in L^2 (\mathbb{R} \rightarrow \mathbb{C}) : \mathcal{F} f (\omega) = 0 \text{ for a.e. } \omega < 0 \} 
  \]
- $L^2 (\mathbb{R} \rightarrow \mathbb{R})$ isomorphic to Hardy-space
- Softy-space is an approximation for a Hardy-space and can be mapped using digital filter $h^+$

Y. Litvin  |  Monaural Source Separation
Time-Frequency Representation Comparison

STFT

Complex BS−WPD
Training

- $E_m (\tilde{S}_c) = 0 \Rightarrow \Lambda_c = \{ \omega_k, 0, \Sigma^{(k)} \}_{k=1}^K$
- Data points $\{ \tilde{S}_1 (m) \}_{m=1}^L, \{ \tilde{S}_2 (m) \}_{m=1}^L$
- Using EM to train GMM models $\Lambda_1, \Lambda_2$
Separation

- Assume $\tilde{S}_c (m)$ generated by $\Lambda_c$ ($c \in \{1, 2\}$ class index)
- Introduce variables $q_c \in \{1, \ldots, K\}$
- Define posterior probability \( \gamma_{j,k} = p (q_1 = j, q_2 = k | \tilde{X}) \)
- When conditioned on \( q_1, q_2 \), mixture components $\tilde{S}_c \sim N (\mu^{(q_c)}, \Sigma^{(q_c)})$ and we may use PM

\[
\hat{S}_1 (i) = \sum_{i,j} \gamma_{i,j} \frac{\sigma_{1}^{(i)}(i)}{\sigma_{1}^{(i)}(i) + \sigma_{2}^{(j)}(i)} \tilde{X}(i)
\]

- $\gamma_{j,k}$ estimated from mixture observation by exhaustive enumeration of $j, k \in \{1, \ldots, K\}$

\[
\gamma_{i,j} \propto p (\tilde{X} | q_1 = j, q_2 = k) p (q_1 = j) p (q_2 = k) = g (\tilde{X}; \Sigma_{1}^{(j)} + \Sigma_{2}^{(k)}) w_{1}^{(j)} w_{2}^{(k)}
\]
Synthetic Signals

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

\[ f_{1,1}^{(1)} = 220\text{Hz}, \quad f_{1,2}^{(1)} = 440\text{Hz}, \quad f_{1,1}^{(2)} = 300\text{Hz}, \quad f_{1,2}^{(2)} = 600\text{Hz} \]

<table>
<thead>
<tr>
<th></th>
<th>SDR$_1$</th>
<th>SIR$_1$</th>
<th>SAR$_1$</th>
<th>SDR$_2$</th>
<th>SIR$_2$</th>
<th>SAR$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>STFT</td>
<td>16</td>
<td>40</td>
<td>16</td>
<td>16</td>
<td>31</td>
<td>16</td>
</tr>
<tr>
<td>CSR-BS-WPD</td>
<td>20</td>
<td>35</td>
<td>20</td>
<td>22</td>
<td>40</td>
<td>22</td>
</tr>
</tbody>
</table>
Natural Signals

- Speech and piano play
- Compared to STFT based algorithm (Benaroya & Bimbot, 2003)
- Comparison parameters
  - GMM order
  - Wavelet family

![Graph showing SDR (Signal to Distortion Ratio) vs. GMM order for different wavelet families: dmey, STFT, coif3, db5. The graph illustrates the performance of the separation algorithm across varying GMM orders, with each wavelet family showing distinct trends.](image-url)
Results Analysis

- Comparing to STFT
  - Low orders of GMM: better than STFT or comparable
  - High orders of GMM: comparable
- Different wavelet families: \textit{dmey} superior to other wavelet families
- Approximate W-DO orthogonality (Yilmaz & Rickard, July 2004)
  - \textit{dmey} CSR-BS-WPD transform has
    - the most sparse coefficients compared to other wavelet families
    - sparseness comparable to STFT
    - good frequency localization properties
    - successfully used for speech enhancement (Cohen, 2001)
Outline

1. Introduction
2. Subband Frequency Modulating Signal Modeling
3. Spectral Kurtosis
4. Bark-Scaled WPD
5. Conclusion
Summary

EFMS

- Definition of new signal analysis domain
  - Demonstration of usefulness in the task of source separation
- Novel monaural separation algorithm
- Based on subband phase signal properties (EFMS)
- Accounts for subband time dynamics and not spectral shape
- Good perceptual quality
Summary

STSK

- High order statistics (short time spectral kurtosis) for single channel source separation
  - Based on unpublished work of J. Benesty
  - Ad-hoc definition and estimator
  - Demonstration of usefulness in the task of source separation
- Defined STSK
- Like EFMS, provides good local TF signal characterization
- Study of STSK statistical properties is necessary
- Good experimental results
Summary

CSR-BS-WPD

- Extension of Bark-Scaled Wavelet Packet Decomposition (Cohen, 2001)
  - Approximate shiftability
  - Constant sampling rate
- Constant Sampling Rate Bark-Scaled signal analysis introduced
  - Critical band structure
  - Approximate shiftability
  - Easy access to spectral shape at given time index
- GMM based single channel source separation algorithm introduced
  - Reduced dimension of data points (compared to STFT)
  - Reduced computational complexity
  - Improved performance compared to STFT based algorithm
Future Research

- “Edge preserving” EFMS estimation (bilateral filtering?)
- More “sophisticated” FM analysis (e.g. spectral analysis of FM signal)
- Non-uniform filterbank
- Varying values EFMS for different frequencies
- Soft instead of binary masks
- Incorporate spectral information into classification
- Rigorous definition of STSK and its statistical properties
- Additional applications of EFMS, STSK and CSR-BS-WPD analysis (e.g. signal classification)
Thank you!
The spectral kurtosis: a useful tool for characterising non-stationary signals.

Cambridge, MA: MIT Press.
For Further Reading II

Benaroya, L., & Bimbot, F. 2003 (Apr.).
Wiener Based Source Separation with HMM/GMM using a Single Sensor.

Benesty, J. 2009 (Jul).
private communication.

Enhancement of speech using bark-scaled wavelet packet decomposition.
*Pages 1933–1936 of: Eurospeech.*
Independent component analysis, a new concept? 

A new framework for complex wavelet transforms. 

Proposals for Performance Measurement in Source Separation. 
*Pages 763–768 of: Proc. 4th International Symposium on ICA and BSS (ICA2003).*
Hanson, B., & Wong, D. 1984 (Mar).
The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech.

On amplitude and frequency demodulation using energy operators.

A phenomenological model for vowel production in the vocal tract.
For Further Reading V


Blind separation of speech mixtures via time-frequency masking.

Blind Source Separation by Sparse Decomposition in a Signal Dictionary.
BASS Tasks Taxonomy

- **Following taxonomy** (Vincent et al., 2003)
- **AQO** - Audio quality oriented
  - One versus all
  - Audio scene modification
- **SO** - Significance oriented
Applications

- **One versus all**
  - track extraction from polyphonic music
  - speech enhancement
  - old recording restoration
  - karaoke
  - object-based audio coding

- **Audio scene modification**
  - remixing of existing recordings
  - signal enhancement in hearing aids

- **Significance oriented**
  - speaker identification
  - polyphonic music transcription
  - musical instrument identification in polyphonic music