A Study on the Frequency-Domain Primary-Ambient Extraction For Stereo Audio Signals

Jianjun He, Woon-Seng Gan, and Ee-Leng Tan

{ jhe007@e.ntu.edu.sg, ewsgan@ntu.edu.sg, etanel@ntu.edu.sg }

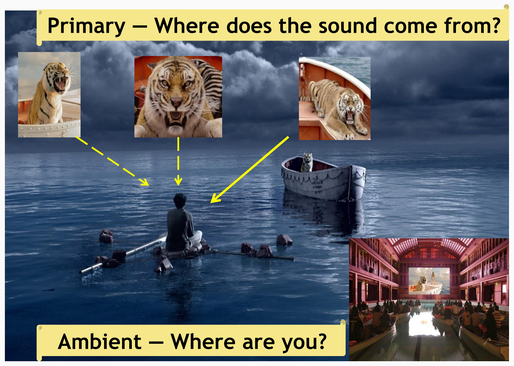
DSP Lab, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

Motivation

Objective: to obtain a new representation of sound scenes in digital media, which is both flexible and efficient in spatial audio reproduction for any playback systems.

- Existing sound scene representations:
  - **Channel-based**
    - ✓ Conventional, for a specific playback system;
    - ❑ **Lacks the flexibility to support different playback configurations.**
  - **Object-based**
    - ✓ Emerging, for any playback system;
    - ❑ **Lacks the efficiency: large storage and high transmission bandwidth.**
- **Primary-ambient based representation**
  - ✓ Inspired by human auditory system;
  - ✓ Facilitates flexible and efficient rendering.

- **Primary-ambient extraction (PAE)** from the channel-based audio (e.g., stereo).
  - ✓ Existing approaches: mainly for one dominant source in primary components;
  - ❑ **PAE with multiple sources (different directions) not well studied.**



Stereo Signal Model

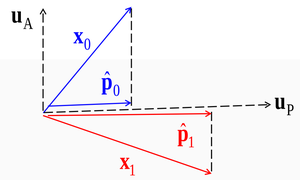
Signal = Primary + Ambient

$$\mathbf{x}_0 = \mathbf{p}_0 + \mathbf{a}_0$$

$$\mathbf{x}_1 = \mathbf{p}_1 + \mathbf{a}_1$$

Primary correlated	$\mathbf{p}_1 = k\mathbf{p}_0$
Ambient uncorrelated	$\mathbf{a}_0 \perp \mathbf{a}_1$
Primary ambient uncorrelated	$\mathbf{p}_i \perp \mathbf{a}_j$
Ambient power balanced	$P_{\mathbf{a}_0} \approx P_{\mathbf{a}_1}$

PCA based PAE



$$\hat{\mathbf{p}}_{\mathrm{PCA},0} = \frac{1}{1+k^2}\left(\mathbf{x}_0 + k\mathbf{x}_1\right),\ \hat{\mathbf{p}}_{\mathrm{PCA},1} = \frac{k}{1+k^2}\left(\mathbf{x}_0 + k\mathbf{x}_1\right)$$

Shifted PCA based PAE

To account for the partial primary correlation (0-lag) caused by the time difference, we proposed shifted PCA in [1]. Let *d* be the inter-channel time difference.

$$\hat{p}_{\mathrm{SPCA},0}(n) = \frac{1}{1+k^2}\left[x_0(n) + kx_1(n-d)\right],$$

$$\hat{p}_{\mathrm{SPCA},1}(n) = \frac{k}{1+k^2}\left[x_0(n+d) + kx_1(n)\right],$$

In frequency domain

$$\hat{P}_{\mathrm{SPCA},0}(f) = \frac{1}{1+k^2}\left[X_0(f) + kX_1(f)e^{-j2\pi fd/N}\right],$$

$$\hat{P}_{\mathrm{SPCA},1}(f) = \frac{k}{1+k^2}\left[X_0(f)e^{j2\pi fd/N} + kX_1(f)\right].$$

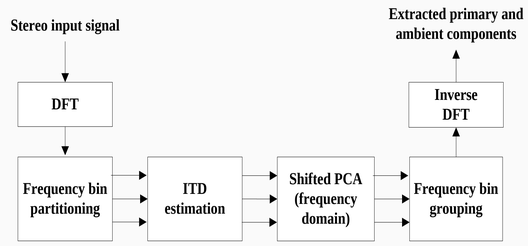
[1] J. He, E. L. Tan, and W. S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in Proc. ICASSP, Vancouver, Canada, 2013, pp. 266-270.

[2] J. Merimaa, M. M. Goodwin, and J. M. Jot, "Correlation-based ambience extraction from stereo recordings", in Proc. 123rd Audio Eng. Soc. Conv., New York, Oct. 2007.

[3] C. Faller and F. Baumgarte, "Binaural cue coding—part II: schemes and applications," IEEE Trans. Speech Audio Process., vol. 11, no. 6, pp. 520-531, Nov. 2003.

[4] J. He, E. L. Tan, and W. S. Gan, "Linear estimation based primary-ambient extraction for stereo audio signals," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 2, pp. 505-517, Feb. 2014.

Frequency Bin Partitioning



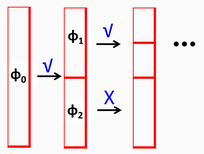
Ideally — No. of partitions = no. of sources; Each partition: the dominant frequency bins of each source → Unknown

In practice — Fixed partitioning: independent of input
- Uniform (2, 8, 32, etc. [2])
- Non-uniform (e.g. ERB [3])

Proposed — **Adaptive partitioning:** dependent of input
- **TD: Top-Down**
- Based on inter-channel cross-correlation coefficient (ICC), with two thresholds: φ_L , φ_H

$$\mathrm{ICC}: \boldsymbol{\varphi} = \frac{\left|\mathbf{X}_0^H \mathbf{X}_1\right|}{\sqrt{\mathbf{X}_0^H \mathbf{X}_0 \mathbf{X}_1^H \mathbf{X}_1}}$$

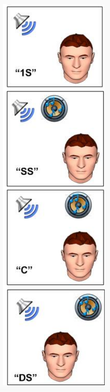
Conditions for partitioning continuation:
- $\varphi_0 < \varphi_H$, and
- $\max(\varphi_1, \varphi_2) > \varphi_0$, and
- $\min(\varphi_1, \varphi_2) > \varphi_L$.



Experimental Settings

Primary components: speech and music
Ambience: white Gaussian noise
Primary power/ signal power = 0.9

Source directions:
Amplitude panned and time shifted
- **1S**: one source
- **SS**: in the same side
- **C** : only one in the center
- **DS**: in different sides



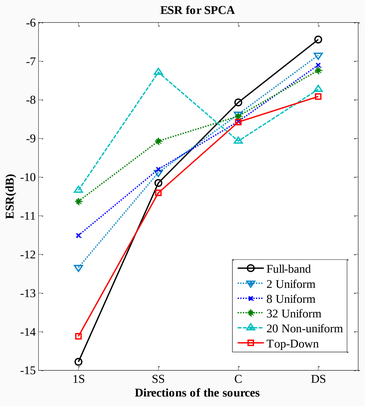
DFT: Size = 4096, Hanning window, 50% overlapping
TD parameters: φ_L = 0.05, φ_H = 0.7

Results

Performance evaluated by Error-to-Signal Ratio (ESR) [4]

$$\mathrm{ESR(dB)} = 10\log_{10}\left[0.5\left(\frac{\|\hat{\mathbf{p}}_0 - \mathbf{p}_0\|_2^2}{\|\mathbf{p}_0\|_2^2} + \frac{\|\hat{\mathbf{p}}_1 - \mathbf{p}_1\|_2^2}{\|\mathbf{p}_1\|_2^2}\right)\right].$$

ESR	Partitioning	1S	SS	C	DS
PCA	Full-band	-3.7	-4.2	-8.1	-4.7
	2 Uniform	-3.4	-4.0	-8.2	-5.0
	8 Uniform	-3.3	-3.9	-8.3	-5.2
	32 Uniform	-3.2	-3.9	-8.4	-5.5
	20 Non-uniform	-3.3	-4.0	-9.6	-6.9
	Top-Down	-3.7	-4.2	-8.4	-5.0
SPCA	Full-band	-14.8	-10.2	-8.1	-6.5
	2 Uniform	-12.3	-9.9	-8.4	-6.9
	8 Uniform	-11.5	-9.8	-8.6	-7.1
	32 Uniform	-10.6	-9.1	-8.4	-7.3
	20 Non-uniform	-10.3	-7.3	-9.0	-7.7
	Top-Down	**-14.1**	**-10.4**	**-8.6**	**-7.9**



CONCLUSIONS

1. Regardless of partitioning methods, SPCA outperforms PCA.
2. Frequency bin partitioning is unnecessary for one source, but is essential for multiple sources.
3. Not all partitioning methods yield better performance than the full-band method (i.e., no partitioning).
4. Generally, the best performance is obtained with the proposed ICC-based Top-Down adaptive partitioning method.