Multi-Shift Principal Component Analysis based Primary Component Extraction for Spatial Audio Reproduction

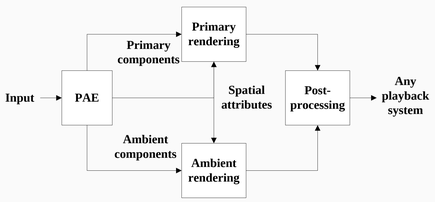
Jianjun He, and Woon-Seng Gan {jhe007@e.ntu.edu.sg, ewsgan@ntu.edu.sg}

Digital Signal Processing 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 of immersive spatial audio.



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

Stereo Signal Model[1]

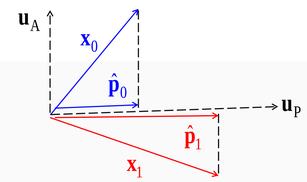
Signal = Primary + Ambient

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

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

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	$P_{\mathbf{a}_0} \approx P_{\mathbf{a}_1}$

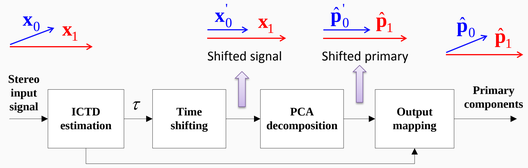
PCA based PAE [1, 2]



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

Shifted PCA based PAE[3]

To account for the partial primary correlation (0-lag) caused by the inter-channel time difference (ICTD) τ.



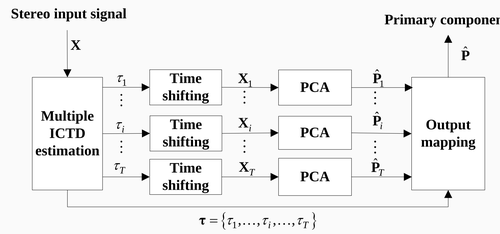
$$\hat{p}_{\mathrm{SPCA},0}(n) = \frac{1}{1+k^2}\left[x_0(n) + kx_1(n-\tau)\right],$$

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

Multi-Shift PCA (MSPCA) based PAE

MSPCA: for concurrent directional sound sources (from different directions) in the primary components.

Typical structure of MSPCA (MSPCA-T)



Consider a few selective shifts

$\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1\}$;

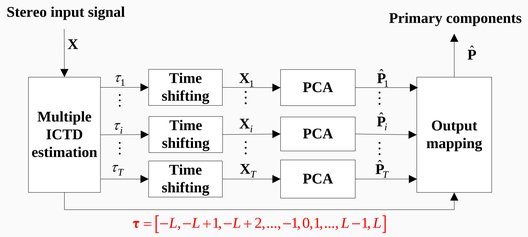
τ_i : ith estimated ICTD (T : total number of ICTDs)

$\mathbf{X}_i$: shifted signal

$\hat{\mathbf{P}}_i$: extracted shifted primary components

$\hat{\mathbf{P}} = \{\hat{\mathbf{p}}_0, \hat{\mathbf{p}}_1\}$: final output of the extracted primary components

Consecutive structure of MSPCA



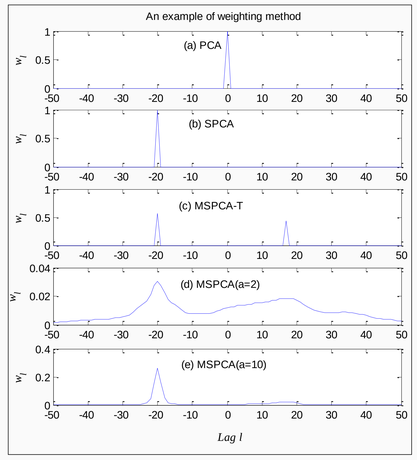
Consecutive shifting lag by lag, and apply different weights to different shifted versions. The weights are derived based on inter-channel cross-correlation coefficient (ICC).

$$\hat{P}(n) = \sum_{l=-L}^{L} w_l \hat{P}_l(n+l),$$

where $w_l = \phi_l^a \Big/ \sum_{l=-L}^{L} \phi_l^a$,

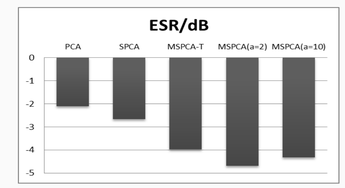
a : exponent of ICC

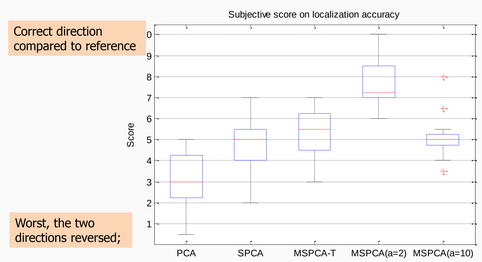
Results

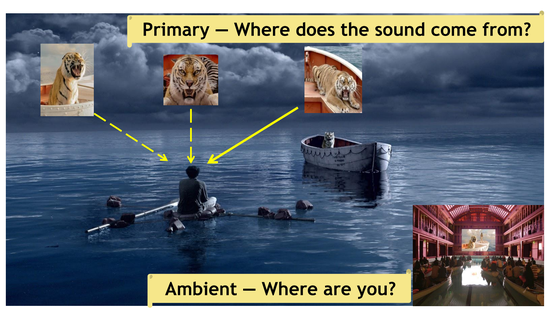


Performance evaluated by Error-to-Signal Ratio (ESR) [1]

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







Experimental Settings

Primary components: speech and music;
Ambience: white Gaussian noise;
Equal power among speech, music, ambience;
ICTD range: ±50 lags, (~2ms for fs=44.1 kHz);
Approaches:

- PCA, SPCA;
- MSPCA-T, MSPCA (a = 2, 10)

CONCLUSIONS

1. Proposed multi-shift PCA to handle multiple sources in primary component extraction;
2. MSPCA with typical structure (selected shifts), but its performance is degraded when ICTD estimation is inaccurate;
3. MSPCA with consecutive structure is more robust, by applying weights on every shifted versions.
4. The weighting method for different shifts is critical; in general, applying a proper exponent of the ICC yields good (objective and subjective) performance.

[1] 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.

[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] 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.

[4] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones," IEEE Signal Processing Magazine, vol. 32, no.2, pp. 100-113, Mar. 2015.