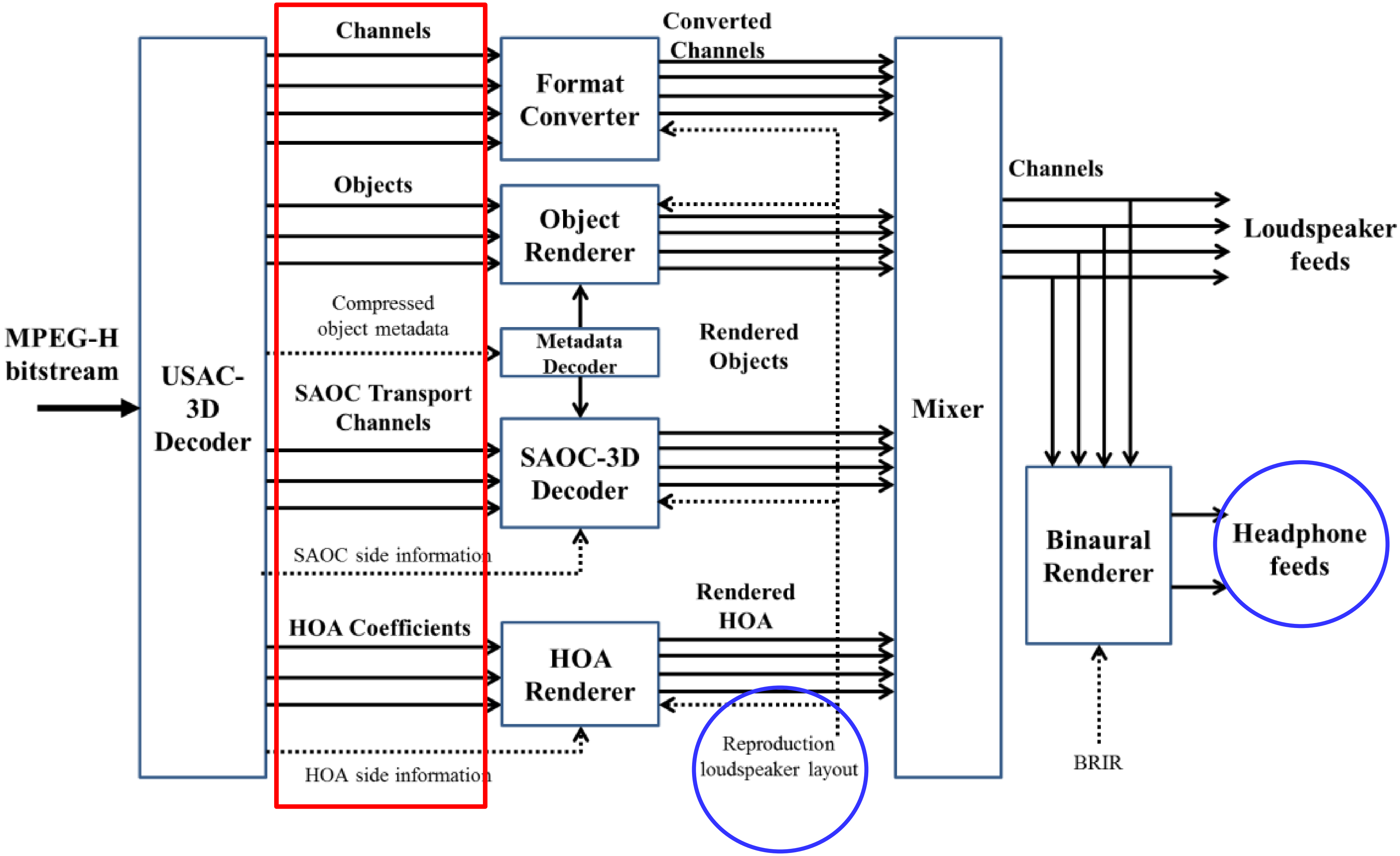
Applying Primary Ambient Extraction for Immersive Spatial Audio Reproduction

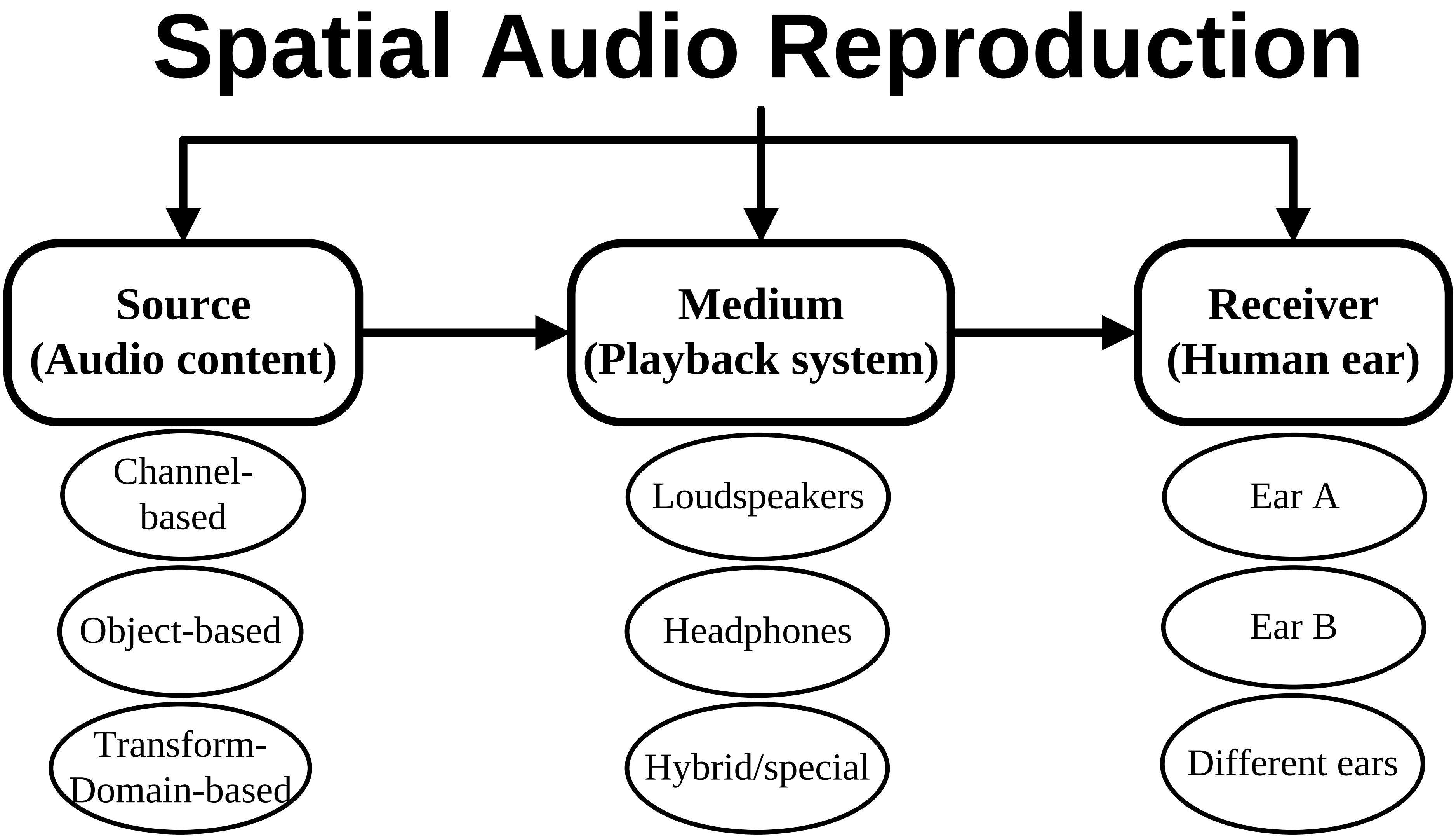
Jianjun He and Woon-Seng Gan

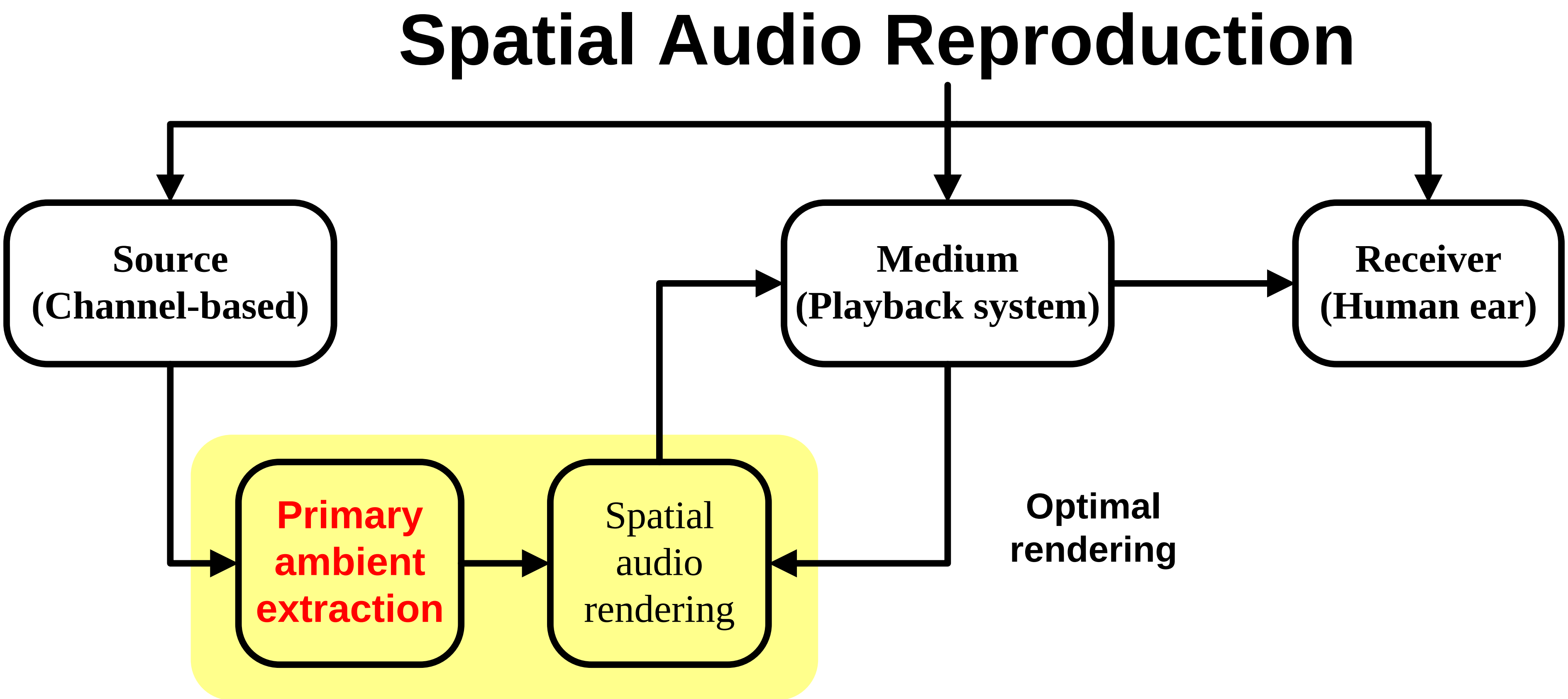
Digital Signal Processing Laboratory
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore

18th Dec, 2015

New 3D audio standard: MPEG-H 3D Audio

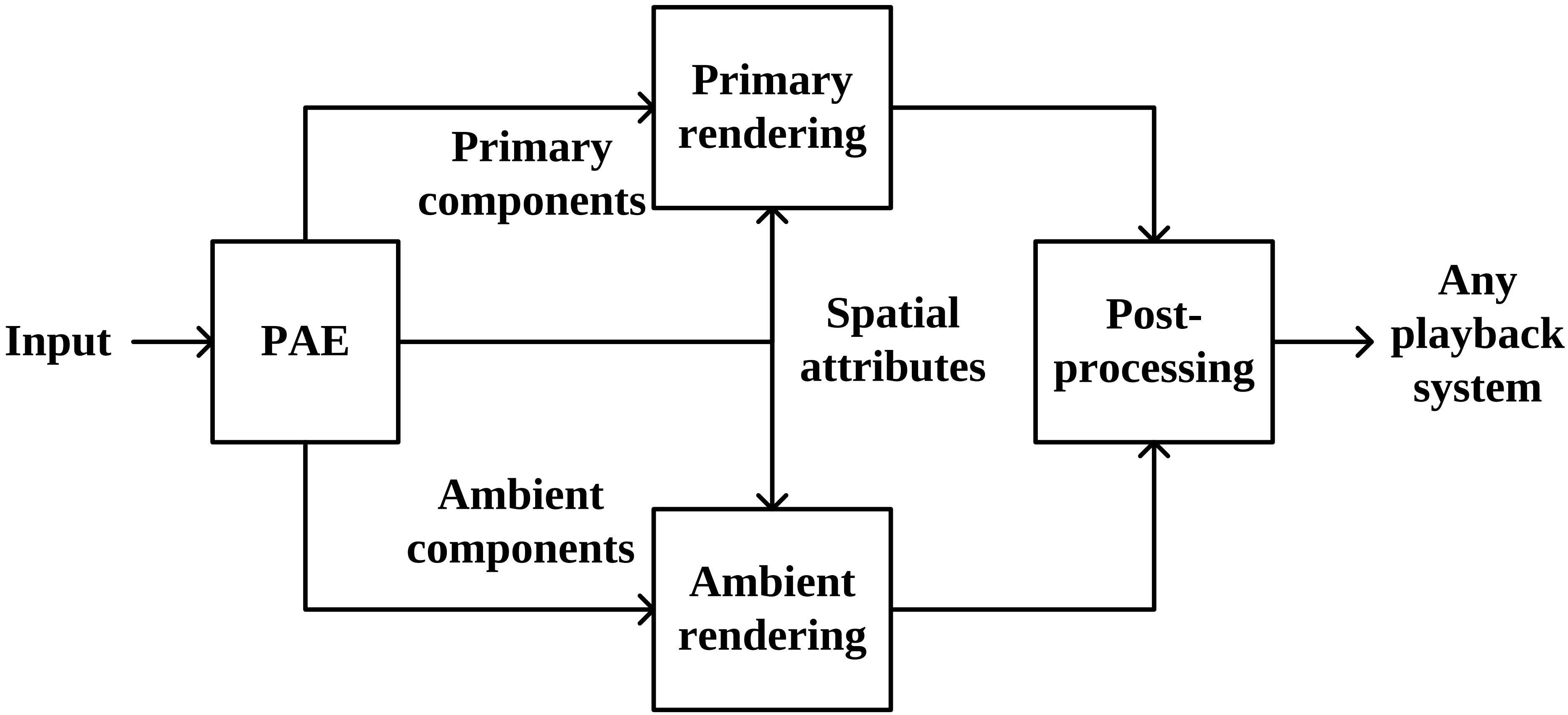



Spatial Audio Reproduction
Source
(Audio content)
Medium
(Playback system)
Receiver
(Human ear)
Channel-based
Object-based
Transform-Domain-based
Loudspeakers
Headphones
Hybrid/special
Ear A
Ear B
Different ears



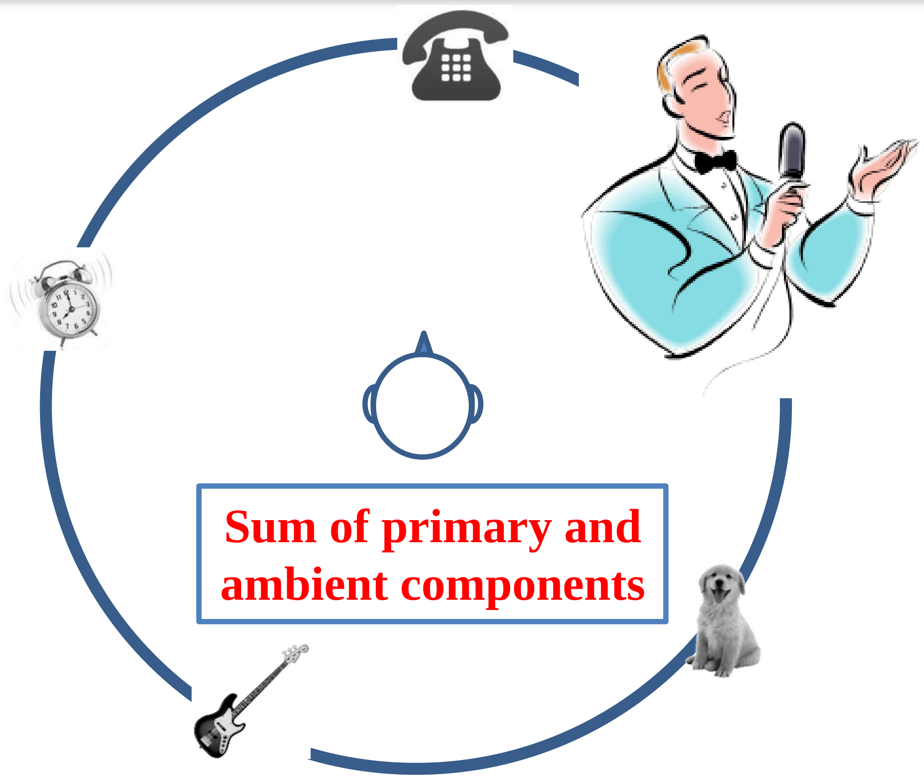
Essentially, PAE serves as a front-end to facilitate **flexible**, **efficient**, and **immersive** spatial audio reproduction.

PAE based spatial audio reproduction



Sound scene decomposition: PAE

Objective:
to extract the primary and ambient components from M (M = 2, stereo) mixtures



Mixtures = primary component + ambient component

$$x_m(n) = p_m(n) + a_m(n)$$

Definitions with Stereo Signal Model

Signal = Primary + Ambient

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

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

Assumptions

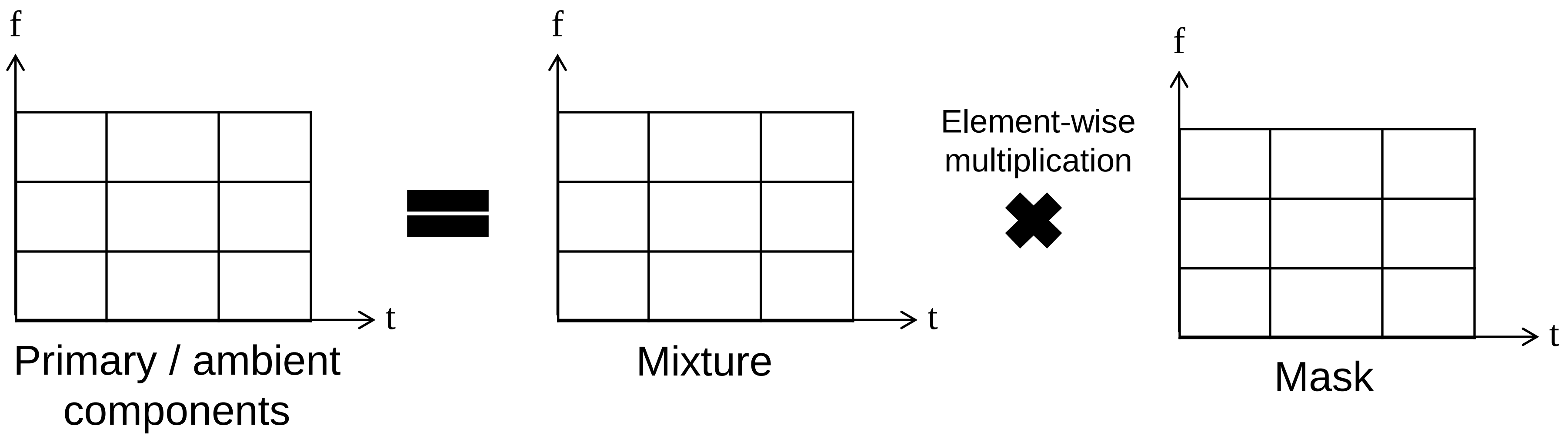
Primary components highly correlated	$\mathbf{p}_1 = k\mathbf{p}_0$
Ambient components uncorrelated	$\mathbf{a}_0 \perp \mathbf{a}_1$
Primary ambient components uncorrelated	$\mathbf{p}_i \perp \mathbf{a}_j$
Ambient power balanced	$P_{\mathbf{a}_0} = P_{\mathbf{a}_1}$

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.

Overview of recent work in PAE

No. of channels	Complexity of audio scenes		
	Basic (single source, only amplitude panning)	Medium (single source)	Complex (multiple sources)
Stereo	Time frequency masking: [53], [31], [49], [34] PCA: [54]-[58], [49], [26], [17]-[19], [46], [29] Least-squares: [45], [38], [36], [41], [29], [59] Ambient spectrum estimation: [60], [61] Others: [22], [32], [62]	LMS: [37] Shifted PCA: [63] Time shifting: [64]	PCA: [65], [40], [66]
Multichannel	PCA: [26] Others: [48], [67], [18], [68]	ICA and time-frequency masking: [69] Pairwise correlations: [70] Others: [27]	ICA: [69]
Single	NMF: [72] Neural network: [73]		

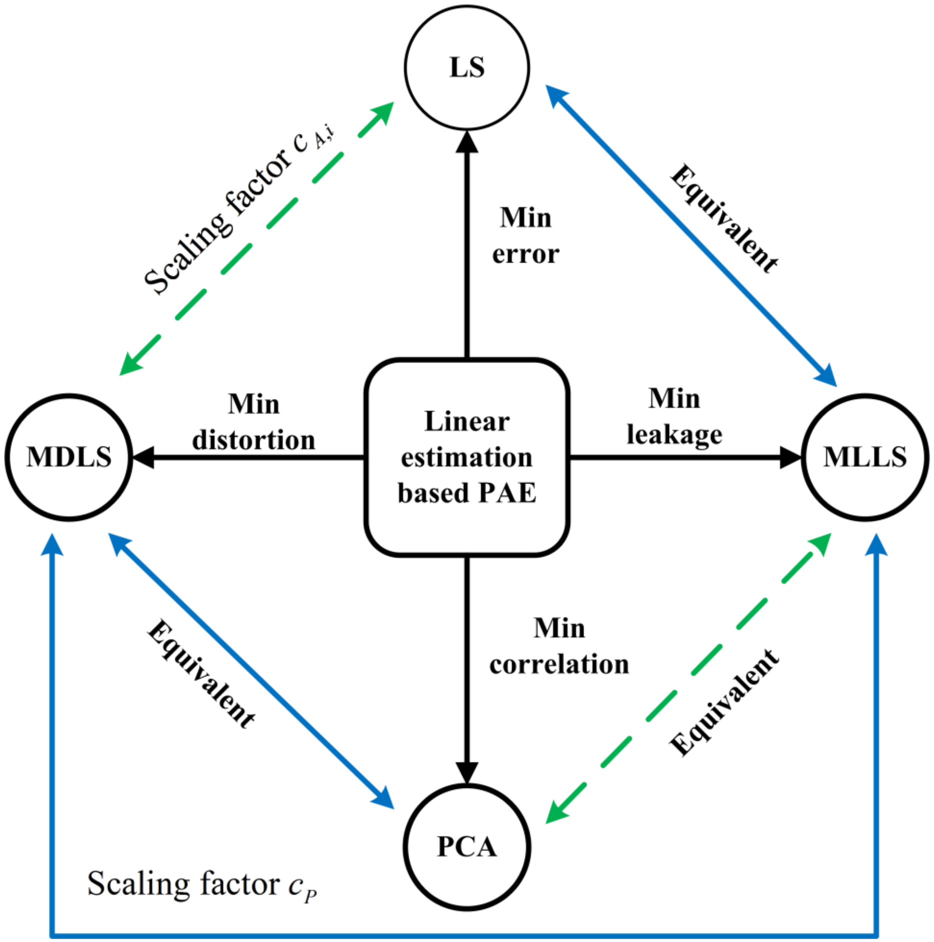
PAE: time frequency masking



Mask can be constructed using

- **Inter-channel coherence [Avendano and Jot, 2004]**
- **Pairwise correlation [Thompson et al., 2012]**
- **Equal level of ambience [Merimaa et al., 2007]**
- **Diffuseness [Pulkki, 2007]**

PAE: linear estimation



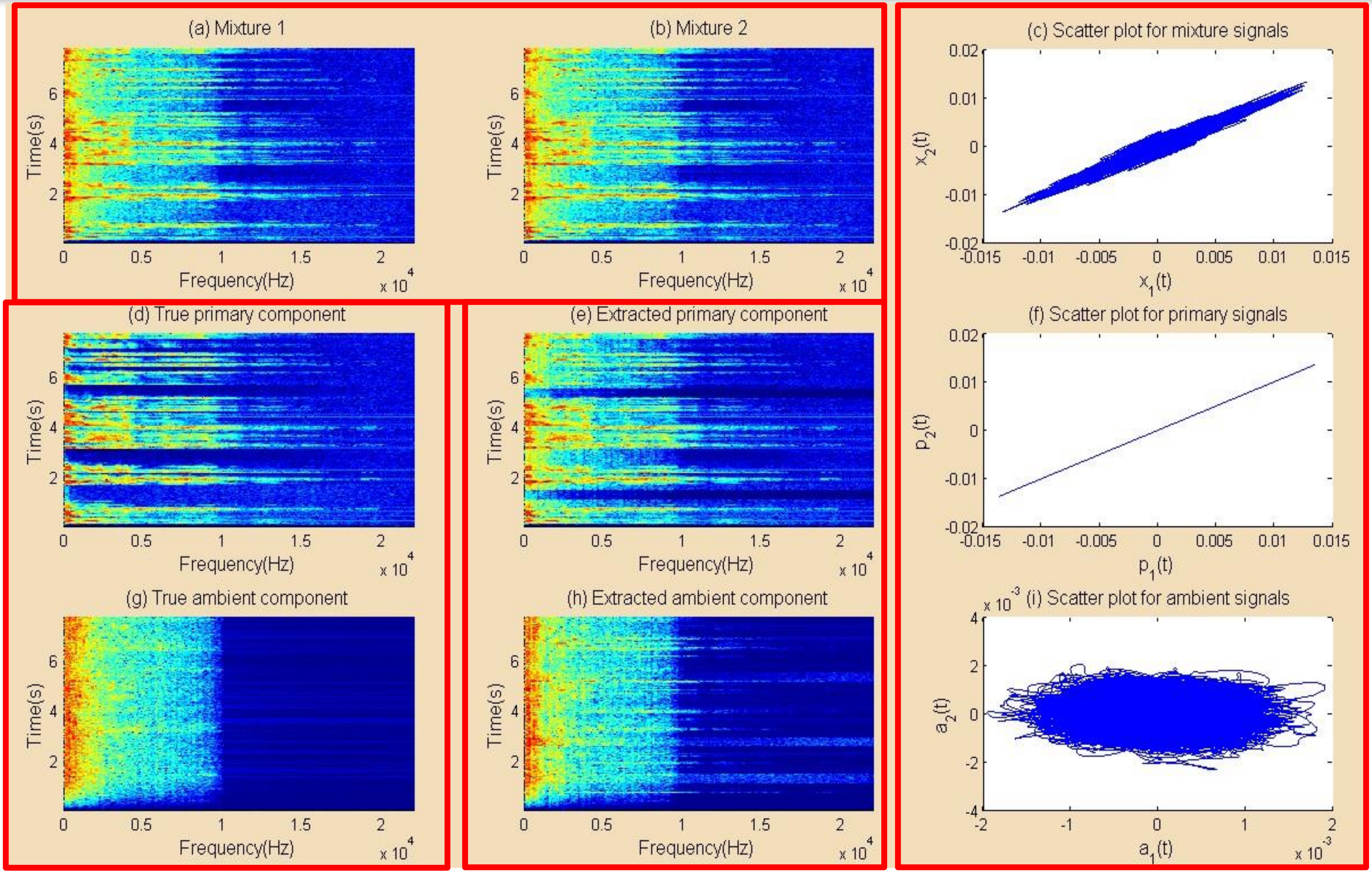
$$\begin{bmatrix} \hat{p}_0(n) \\ \hat{p}_1(n) \\ \hat{a}_0(n) \\ \hat{a}_1(n) \end{bmatrix} = \begin{bmatrix} w_{\mathrm{P}0,0} & w_{\mathrm{P}0,1} \\ w_{\mathrm{P}1,0} & w_{\mathrm{P}1,1} \\ w_{\mathrm{A}0,0} & w_{\mathrm{A}0,1} \\ w_{\mathrm{A}1,0} & w_{\mathrm{A}1,1} \end{bmatrix} \begin{bmatrix} x_0(n) \\ x_1(n) \end{bmatrix}$$

Objectives and relationships of four linear estimation based PAE approaches.

- **Blue** solid lines represent the relationships in the **primary** component;
- **Green** dotted lines represent the relationships in the **ambient** component.
- **MLLS**: minimum leakage LS
- **MDLS**: minimum distortion LS

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

PAE: an example from least-squares



PAE: ambient spectrum estimation

$$\mathbf{X}_0 = \mathbf{P}_0 + \mathbf{A}_0,\ \mathbf{X}_1 = \mathbf{P}_1 + \mathbf{A}_1 \quad |\mathbf{A}_0| = |\mathbf{A}_1| = |\mathbf{A}| \quad \mathbf{A}_c = |\mathbf{A}| \otimes \mathbf{W}_c,\ \forall c \in \{0,1\},$$

Ambient Phase Estimation (APE)

Ambient Magnitude Estimation (AME)

$$|\mathbf{A}| = (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0)$$

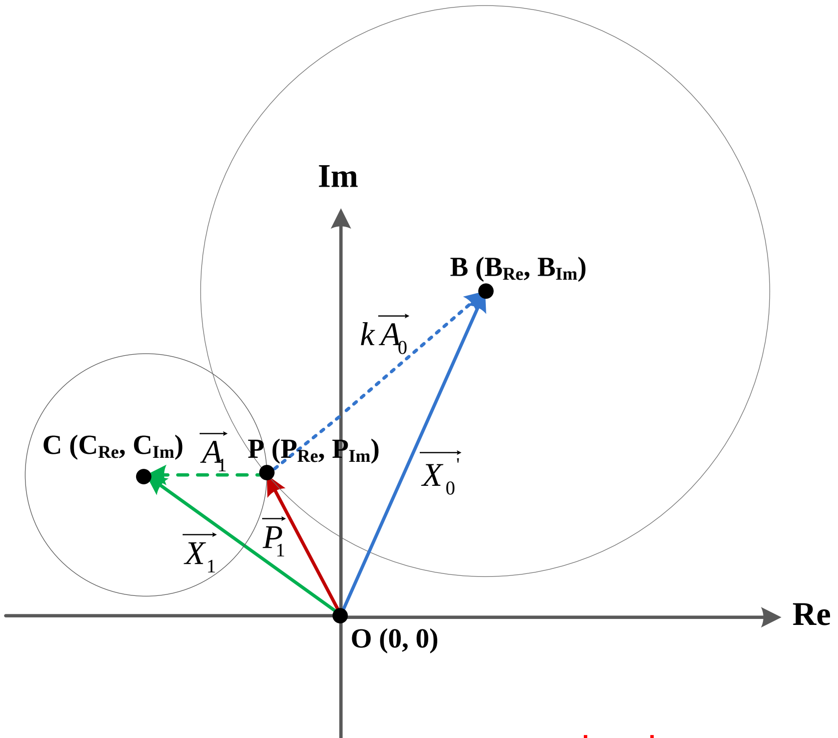
$$\mathbf{A}_c = (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0) \otimes \mathbf{W}_c,$$

$$\mathbf{P}_c = \mathbf{X}_c - (\mathbf{X}_1 - k\mathbf{X}_0)./(\mathbf{W}_1 - k\mathbf{W}_0) \otimes \mathbf{W}_c$$

$$\forall c \in \{0,1\}.$$

$$W_0(n,l) = e^{j\theta_0(n,l)}$$

$$W_1(n,l) = e^{j\theta_1(n,l)}$$



Find $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1$

Find $\mathbf{r} = |\mathbf{A}|$

J. He, E. L. Tan, and W. S. Gan, "Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 9, pp. 1431-1444, Sept. 2015.

PAE: ambient spectrum estimation using sparsity

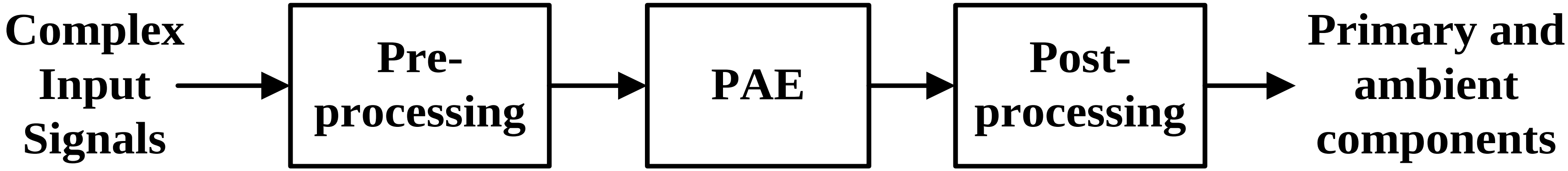
$$\text{APES}:\hat{\theta}_1^* = \arg\min_{\hat{\theta}_1} \left\| \hat{\mathbf{P}}_1 \right\|_1, \text{ or } \quad \text{AMES}:\hat{\mathbf{r}}^* = \arg\min_{\hat{\mathbf{r}}} \left\| \hat{\mathbf{P}}_1 \right\|_1,$$

Approximate efficient solution

$$\text{APEX}:\hat{\boldsymbol{\theta}}_1^* = \begin{cases} \angle \mathbf{X}_1 & , \forall k > 1 \\ \angle\left(\mathbf{X}_1 - \mathbf{X}_0\right), & \forall k = 1 \end{cases}$$

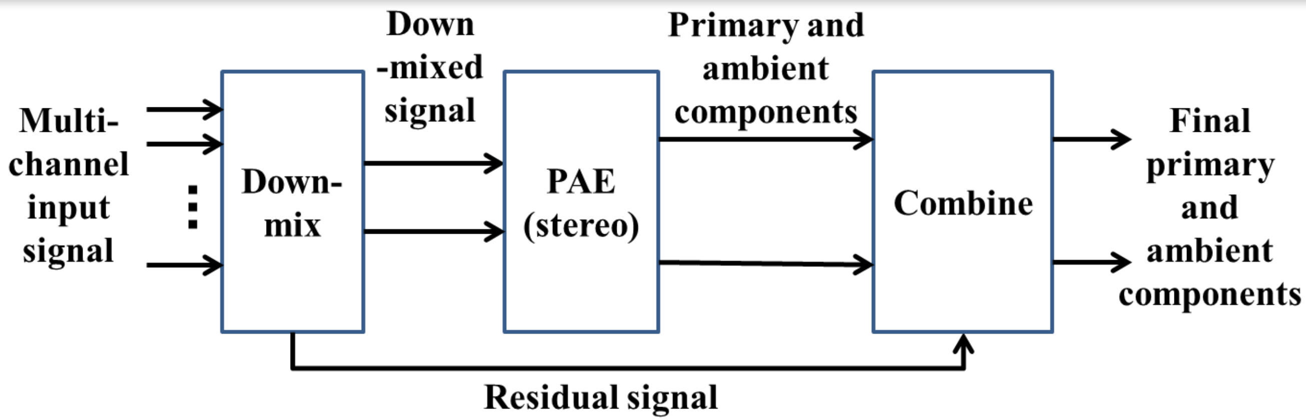
J. He, E. L. Tan, and W. S. Gan, “Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 9, pp. 1431-1444, Sept. 2015.

Framework of preprocessing and postprocessing on PAE

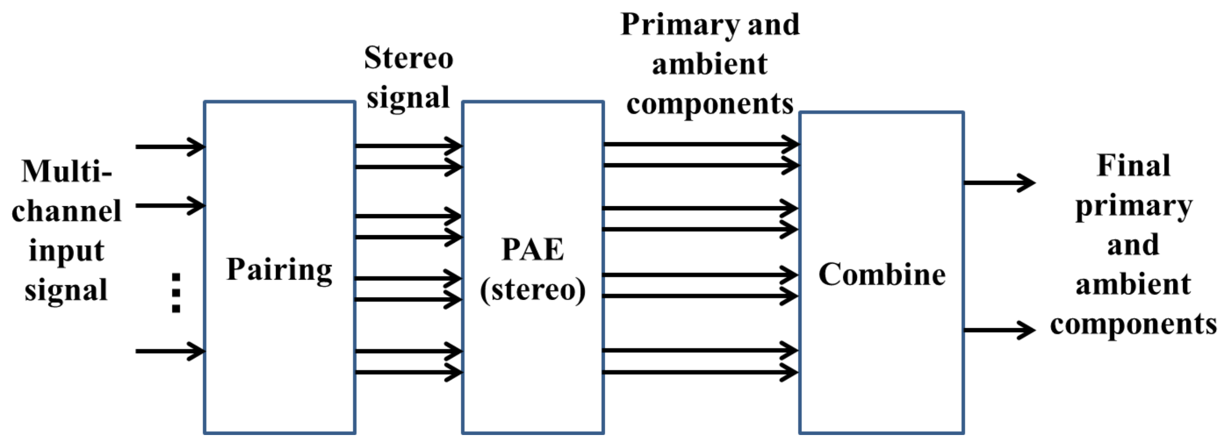


Preprocessing for Multichannel Signals

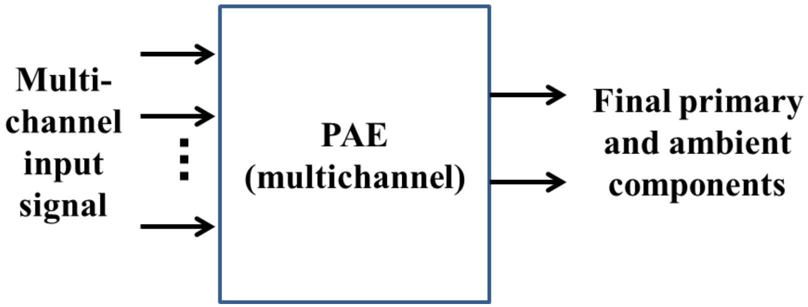
1. Using down-mix



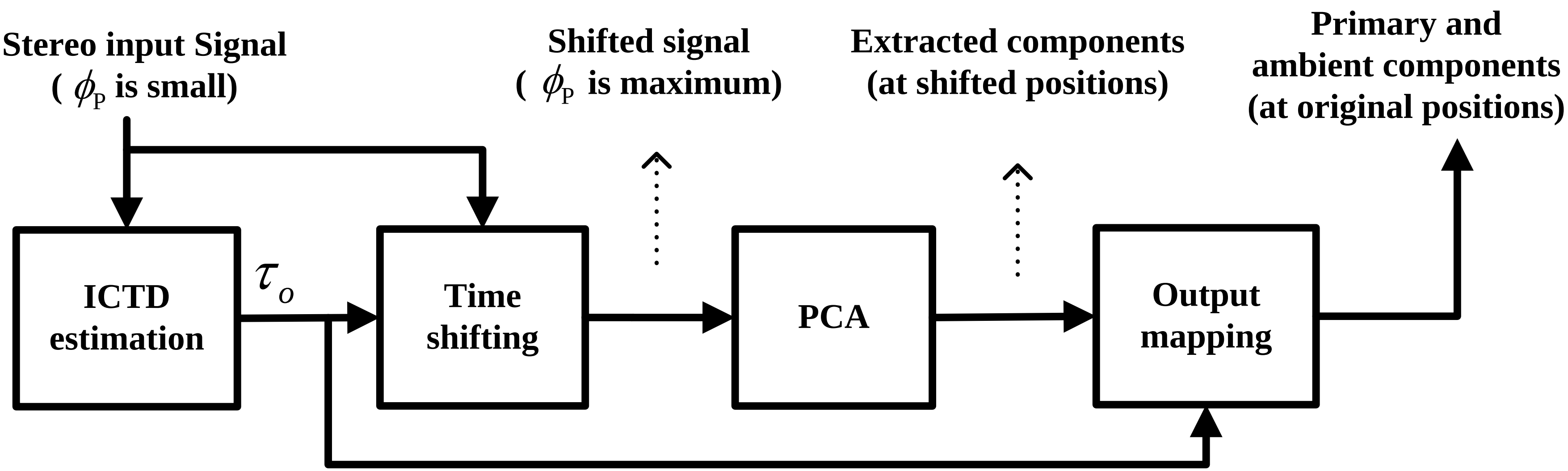
2. Using pairing



3. Direct



Preprocessing for time differences



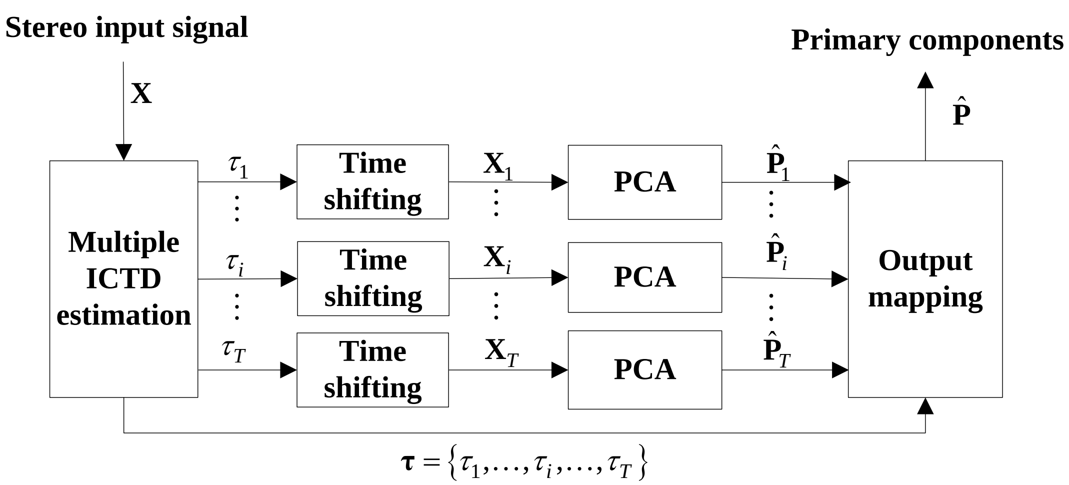
For mixture signals with partially correlated primary components

- **More accurate estimation of model parameter;**
- **Lower extraction error;**
- **Closer estimation of the spatial attributes;**
- **(Increase of computational load).**

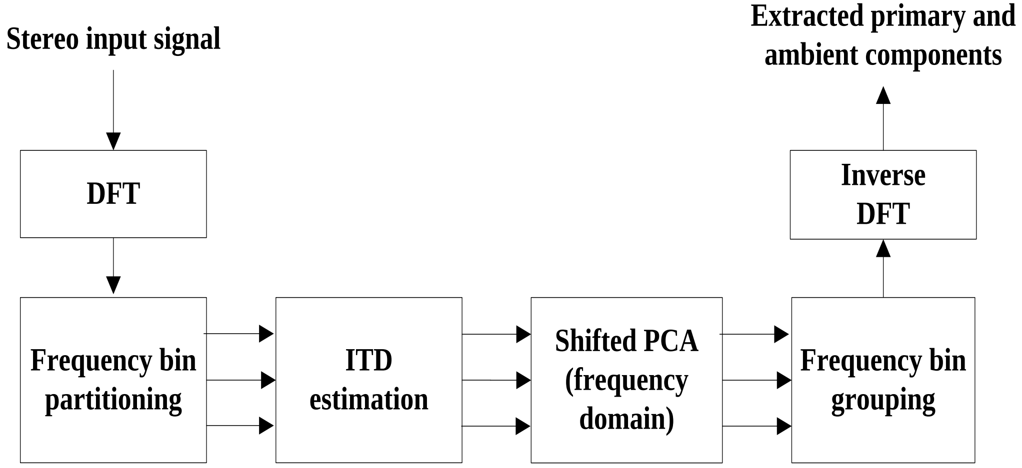
J. He, W. S. Gan, and E. L. Tan, "Time-shifting based primary-ambient extraction for spatial audio reproduction," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 10, pp. 1576-1588, Oct. 2015.

Preprocessing for Multiple Sources

Multi-shifting PAE with ICC based output weighting



Subband PAE with frequency bin partitioning



J. He, and W. S. Gan, "Multi-shift principal component analysis based primary component extraction for spatial audio reproduction," in Proc. ICASSP, Brisbane, Australia, Apr. 2015, pp. 350-354.
J. He, E. L. Tan, and W. S. Gan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in *Proc. ICASSP*, Florence, Italy, 2014, pp. 2892-2896.

Other preprocessing and postprocessing

- Preprocessing
  - Channel switch for $0 < k < 1$
  - Channel out-of-phase compensation for $k < 0$
  - Smoothing in model parameter estimation

- Postprocessing
  - Decorrelation
  - Scaling

Objective evaluation

Stimuli

- Primary component:
  - Speech, $k = 2$
- Ambient component:
  - Wave lapping sound
- Primary power ratio (PPR):
  - (0, 1) at an interval of 0.1
- FFT size: 4096

Approaches compared

- Masking
- PCA
- LS
- APEX

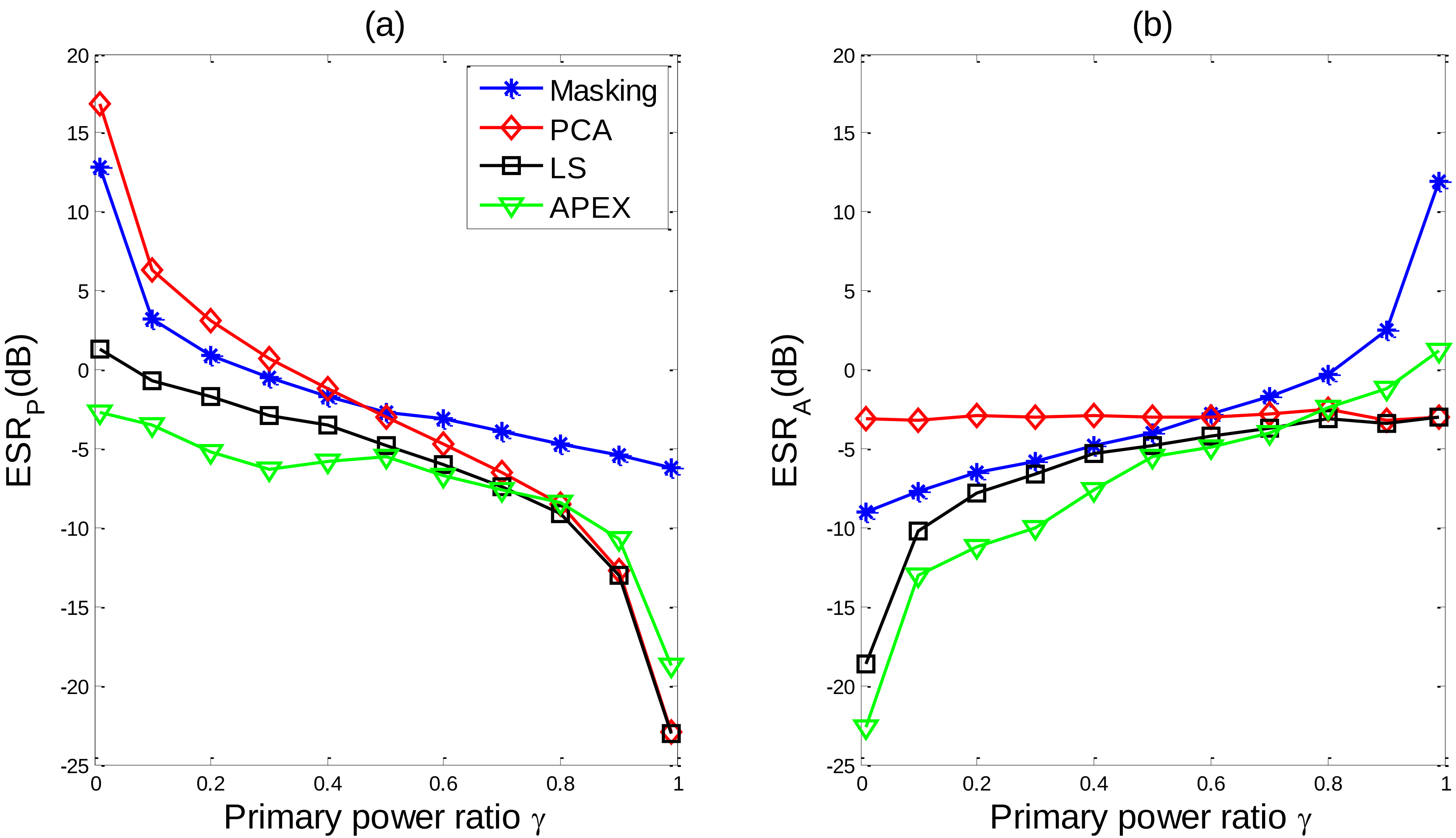
Performance evaluated

1. **Extraction accuracy**: ESR
2. **Spatial accuracy**: ICC, ICLD

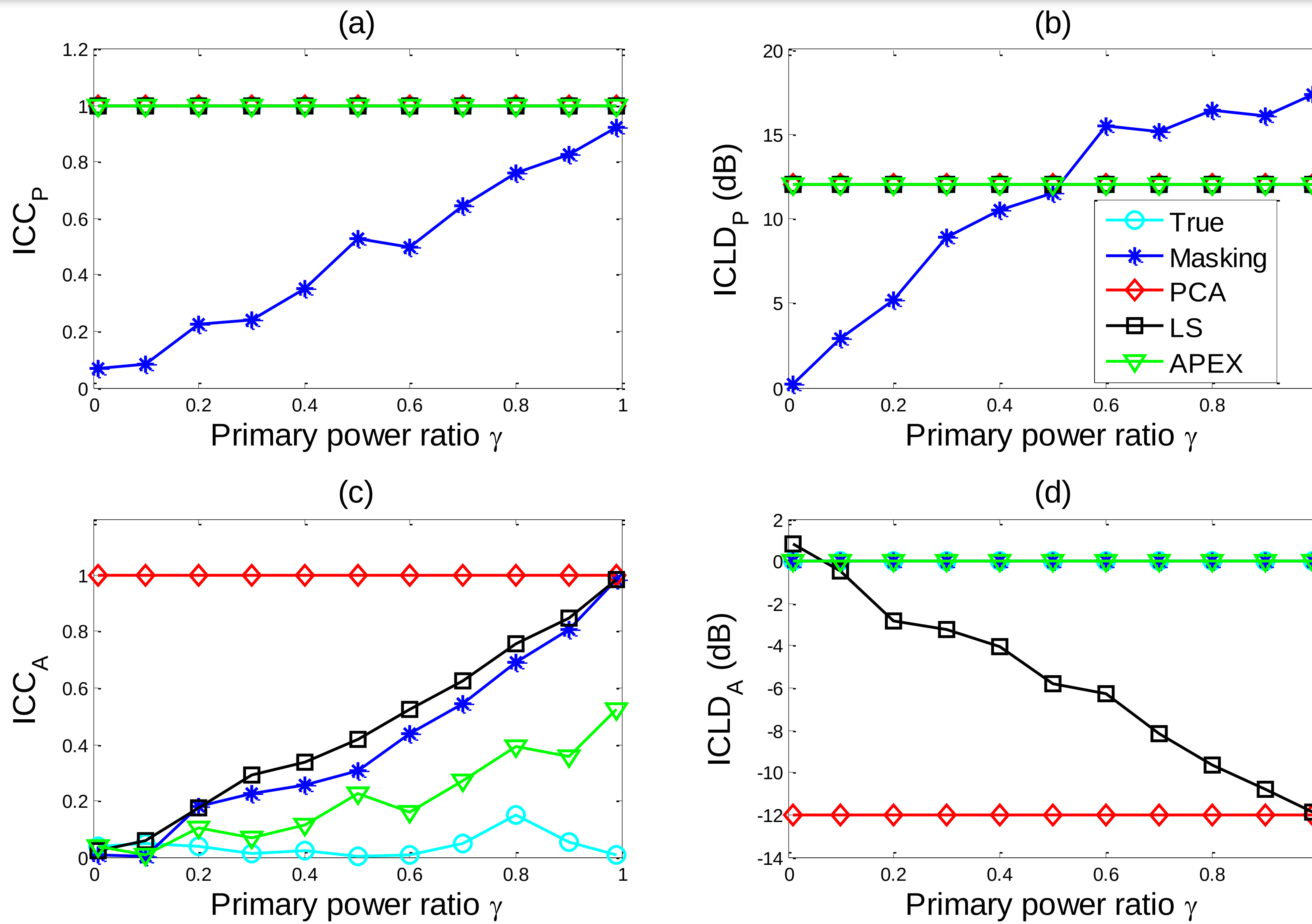
$$\mathrm{ESR}_{\mathrm{P}} = 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{p}}_c - \mathbf{p}_c\right\|_2^2}{\left\|\mathbf{p}_c\right\|_2^2}\right\},$$

$$\mathrm{ESR}_{\mathrm{A}} = 10\log_{10}\left\{\frac{1}{2}\sum_{c=0}^{1}\frac{\left\|\hat{\mathbf{a}}_c - \mathbf{a}_c\right\|_2^2}{\left\|\mathbf{a}_c\right\|_2^2}\right\}.$$

Extraction accuray



Spatial accuracy



Subjective evaluation

Stimuli

- Primary component:
  - speech, music, and bee sound, $k = 2$
- Ambient component:
  - forest, canteen, and waterfall sound
- Primary power ratio (PPR):
  - (0.3, 0.7)
- Duration: 2-4 seconds

Performance evaluated

1. Extraction accuracy
2. Ambient diffuseness

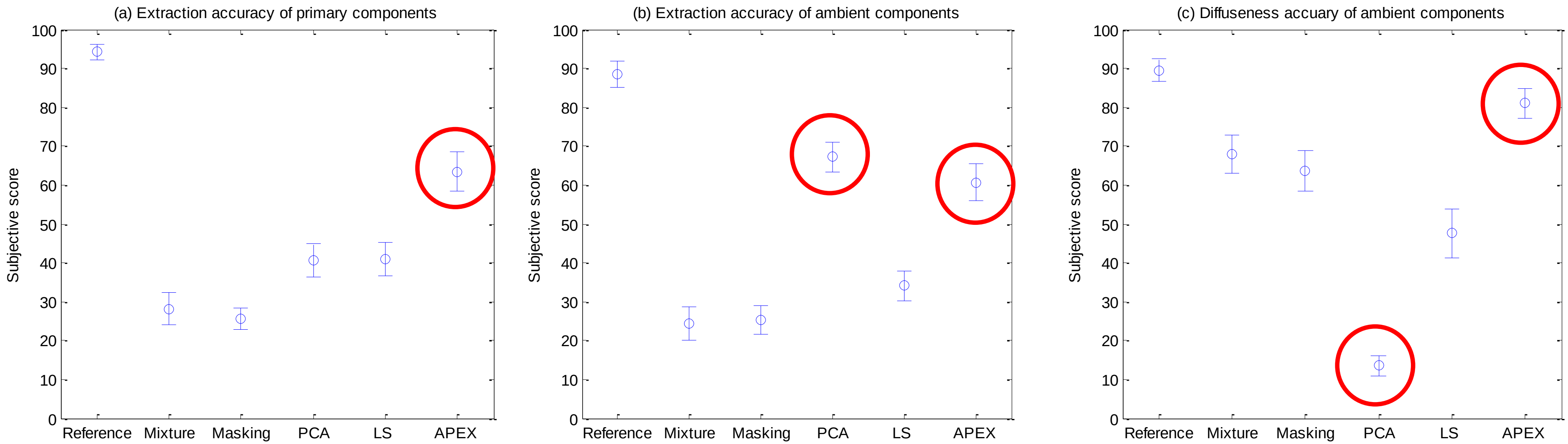
Approaches compared

- Masking
- PCA
- LS
- APEX
- Reference
- Mixture

Listening tests

- 17 subjects
- Headphone listening
- Procedure similar to MUSHRA

Subjective scores



Conclusions

- Introduced primary ambient extraction as a useful tool for immersive spatial audio reproduction;
- Reviewed PAE literature;
- Proposed a preprocessing and postprocessing framework for PAE to deal with complex input signals;
  - For multichannel signals
  - For time differences;
  - For multiple sources;
- Objective and subjective evaluation results provides suggestions on choosing PAE approaches.
- More thorough evaluations for PAE with complex signals.

Read more on primary ambient extraction

[1] J. He, W. S. Gan, and E. L. Tan, "Time-shifting based primary-ambient extraction for spatial audio reproduction," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 10, pp. 1576-1588, Oct. 2015.

[2] J. He, E. L. Tan, and W. S. Gan, "Primary-ambient extraction using ambient spectrum estimation for immersive spatial audio reproduction," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 9, pp. 1431-1444, Sept. 2015.

[3] J. He, W. S. Gan, and E. L. Tan, "Primary-ambient extraction using ambient phase estimation with a sparsity constraint," IEEE Signal Process. Letters, vol. 22, no. 8, pp. 1127-1131, Aug. 2015.

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

[5] 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.

[6] J. He, and W. S. Gan, "Applying primary ambient extraction for immersive spatial audio reproduction," 2015 Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference (invited), Hong Kong, Dec. 2015.

[7] J. He, and W. S. Gan, "Multi-shift principal component analysis based primary component extraction for spatial audio reproduction," in Proc. ICASSP, Brisbane, Australia, Apr. 2015, pp. 350-354.

[8] J. He, W. S. Gan, and E. L. Tan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in Proc. ICASSP, Florence, Italy, 2014, pp. 2892-2896.

[9] 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.

[10] W. S. Gan and J. He, "Assisted Listening for headphones and hearing aids: Signal Processing Techniques," Tutorial at APSIPA ASC 2015, Hong Kong.

[11] J. He, "Spatial audio reproduction using primary ambient extraction," PhD thesis, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Aug. 2015.

[12] J. He, "3D sound effects analysis, synthesis, and application design – a primary ambient extraction approach," Progress report for PhD qualifying examination, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, Jun. 2013.

Thank you and Contact us

Dr. Woon-Seng Gan:
ewsgan@ntu.edu.sg

Mr. Jianjun He:
jhe007@e.ntu.edu.sg

DSP Lab, School of EEE,
Nanyang Technological University,
Singapore