Natural Sound Rendering for Headphones

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

DSP Lab, School of EEE

Nanyang Technological University, Singapore

AudioBeam System
Putting Sound at Where You Want Only

3D
Creating Immersive Soundscape

3D
Headphones
Sound That Surrounds You
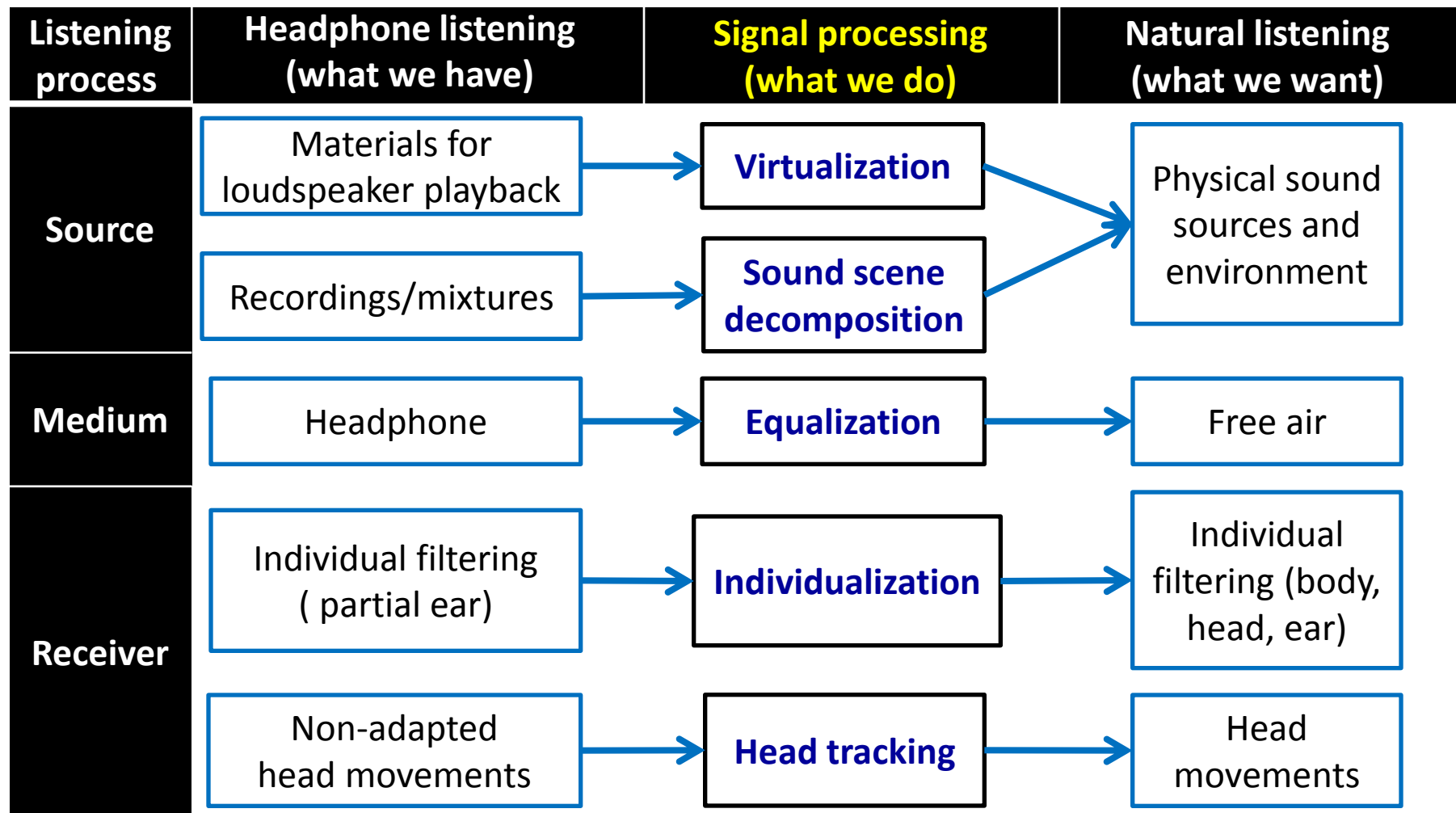
Dsp
Digital Signal Processing Lab

Natural sound

➢ Sound plays an integral part in our life, and it has advantages over vision.

➢ Applications: navigation, communication, medical, multimedia, virtual reality and augmented reality, etc.

➢ We listen to sound in digital media using headphone everyday.

➢ However, conventional headphone listening experience is inherently different from listening in physical world.

➢ It is advantageous to recreate a natural listening experience in headphones.

➢ Rendering natural sound in headphones has been the common objective in headphone industry.

Natural sound rendering essentially refers to rendering of the spatial sound using headphones to create an immersive listening experience and the sensation of "being there" at the venue of the acoustic event.

To achieve natural sound rendering in headphones

➤ **The differences between natural listening and headphone listening;**

➤ **Challenges for rendering sound in headphone to mimic natural listening;**

➤ **How can signal processing techniques help?**

- Virtualization;

- Sound scene decomposition;

- Individualization;

- Equalization;

- Head tracking;

➤ **How to integrate these techniques?**

➤ **Subjective evaluation**

➤ **Conclusions and future trends**

Challenges and solutions

| Listening process	Headphone listening (what we have)	Signal processing (what we do)	Natural listening (what we want)
Source	Materials for loudspeaker playback	**Virtualization**	Physical sound sources and environment
	Recordings/mixtures	**Sound scene decomposition**	
Medium	Headphone	**Equalization**	Free air
Receiver	Individual filtering (partial ear)	**Individualization**	Individual filtering (body, head, ear)
	Non-adapted head movements	**Head tracking**	Head movements

[1] D. R. Begault, *3-D sound for virtual reality and multimedia*: AP Professional, 2000.
[40] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones," in press, IEEE Signal Processing Magazine, Mar. 2015.

Signal processing techniques

1. **Virtualization**: to match the desired playback for the digital media content;

2. **Sound scene decomposition using** blind source separation (**BSS**) and primary-ambient extraction (**PAE**): to optimally facilitate the separate rendering of sound sources and/or sound environment;

3. **Individualization**: to compensate for the lost or altered individual filtering of sound in headphone listening;

4. **Equalization**: to preserve the original timbral quality of the source and alleviate the adverse effect of the inherent headphone response;

5. **Head tracking**: to adapt to the dynamic head movements of the listener.

Virtualization
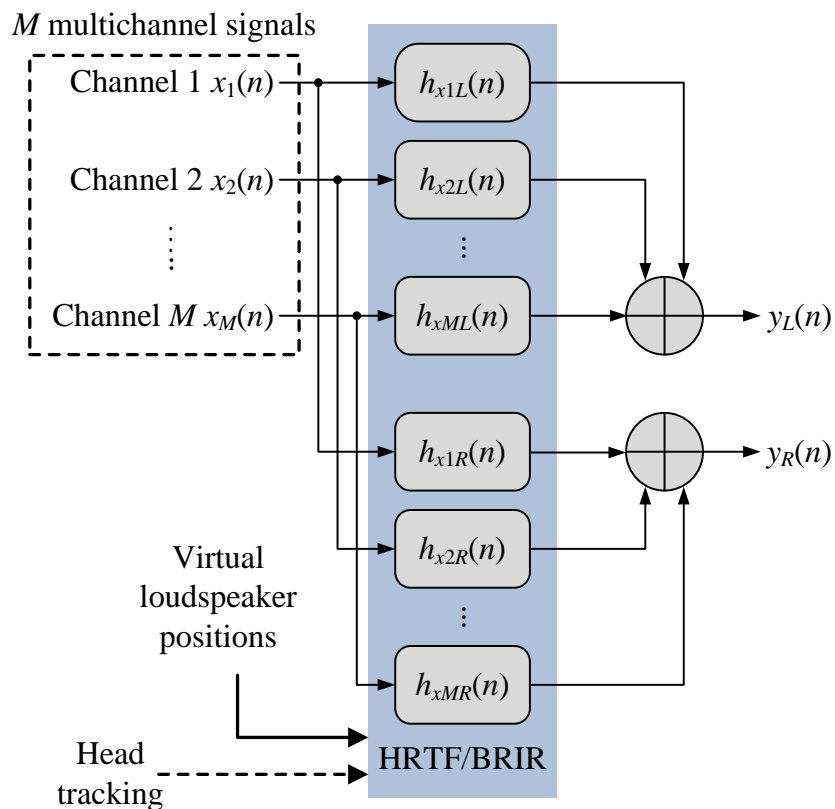
(a) Virtualization of multichannel loudspeaker signals
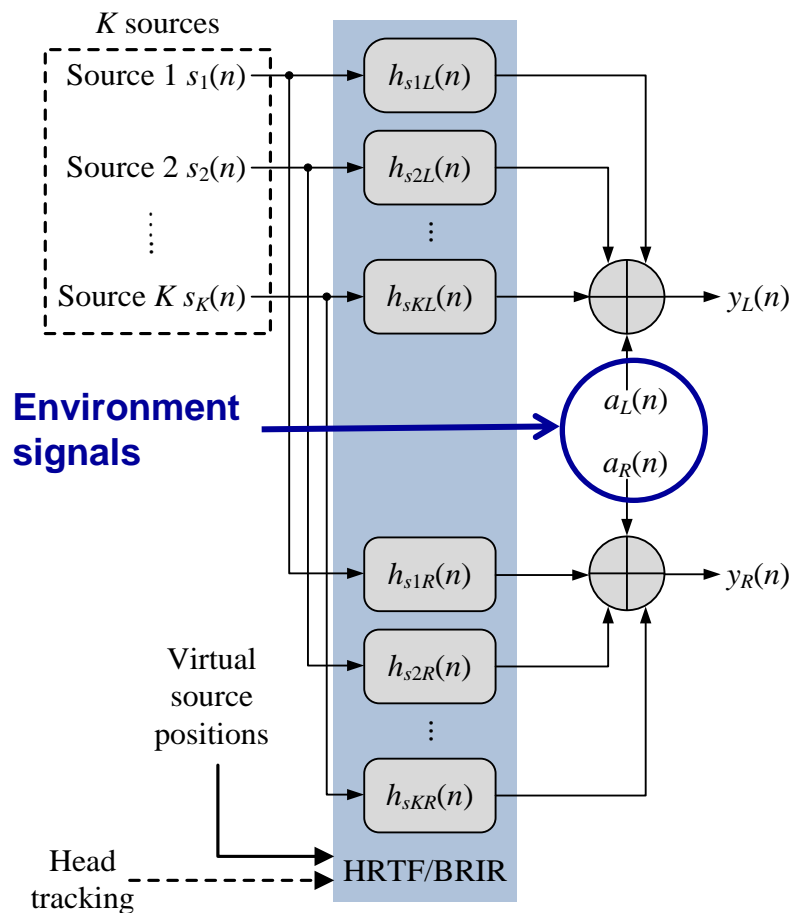
(b) Virtualization of source and environment signals

[6] J. Breebaart and E. Schuijers, "Phantom materialization: a novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 16, no.8, pp. 1503-1511, Nov. 2008.

Virtualization

➢ **Incorporate head tracking**

- Adapt to the changes of sound scene with respect to natural head movements;

- Reduce front-back confusions, azimuth localization errors;

- Concern of head tracking latency.

➢ **Adding reverberation**

- Externalization of the sound sources, and enhance depth perception;

- Rendering of the sound environment;

- How to select correct amount of reverberation.

[10] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.,* vol. 49, no. 10, pp. 904-916, Oct. 2001.
[12] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *Signal Processing Magazine, IEEE,* vol. 28, no. 1, pp. 33-42, Jan. 2011.

Sound scene decomposition

	Blind Source Separation	Primary-Ambient Extraction
Objective	To obtain useful information about the original sound scene from given mixtures, and facilitate natural sound rendering	
Basic model	1. Multiple sources sum together 2. Sources are independent	1. Dominant sources + Environmental signal 2. Primary components are highly correlated; 3. ambient components are uncorrelated
Common characteristics	1. Usually no prior information, only mixture signals 2. Perform extraction/separation based on various signal models 3. Require objective as well as subjective evaluation	
Typical applications	Speech, music	Movie, gaming
Limitations	1. Small number of sources 2. Sparseness/disjoint 3. No/simple environment	1. Small number of sources 2. Sparseness/disjoint 3. Low ambient power 4. Primary ambient uncorrelated

[40] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones," in press, IEEE Signal Processing Magazine, Mar. 2015.

Sound scene decomposition: BSS

Objective: to extract the K sources from M mixtures

Mixtures = function (gains, sources, time difference, model error)

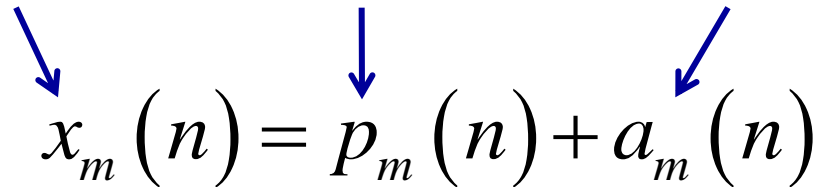$$x_m(n) = \sum_{k=1}^{K} g_{mk} s_k (n - \tau_{mk}) + e_m(n), \quad \forall m \in \{1, 2, \ldots, M\}$$

Case		Typical techniques
M = K		ICA
M > K		ICA with PCA, LS
M < K	M > 2	ICA with sparse solutions
	M = 2	Time-frequency masking
	M = 1	NMF, CASA

ICA: Independent component analysis
PCA: principal component analysis
LS: least squares;
NMF: non-negative matrix factorization;
CASA: computational auditory scene analysis

Sound scene decomposition: PAE

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

Mixtures = primary component + ambient component

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

| Case		Typical techniques
Basic model	Channel-wise estimation	Time-frequency masking
	Combine M channels	Linear estimation (PCA, LS, etc.)
More complex model		Classification, Time/phase shifting, Pairing up two channels, etc.
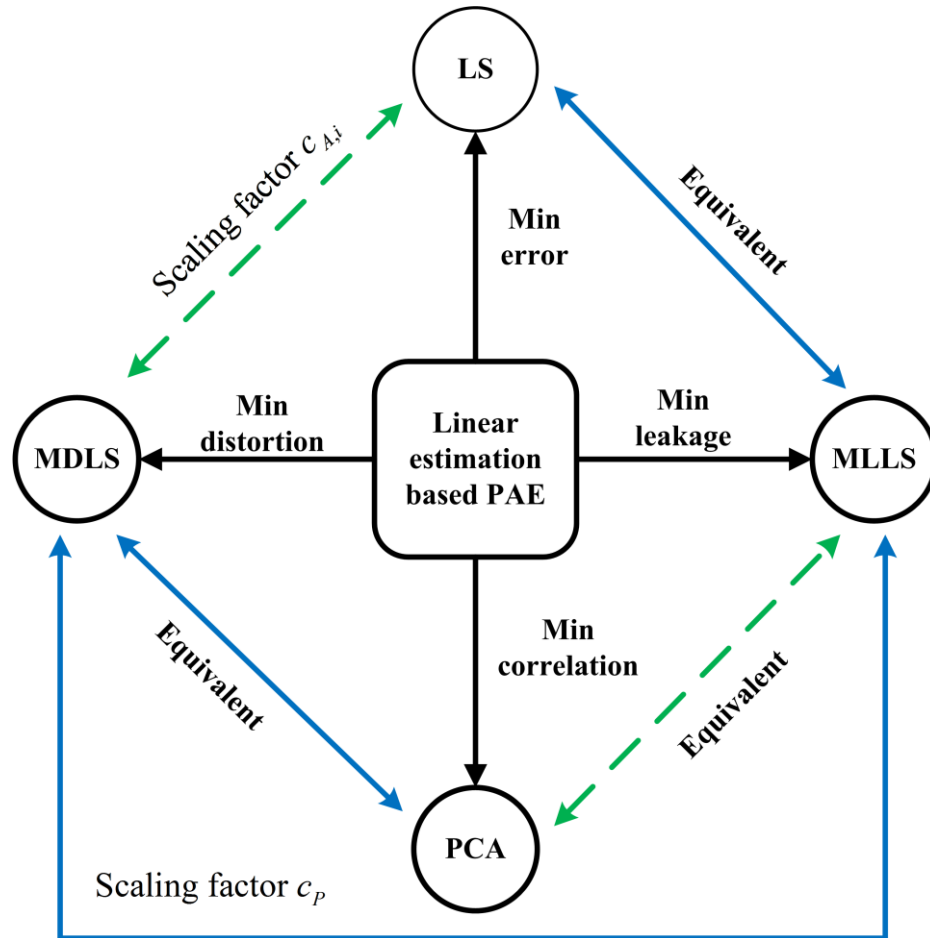
$$\begin{bmatrix} \hat{p}_0(n) \\ \hat{p}_1(n) \\ \hat{a}_0(n) \\ \hat{a}_1(n) \end{bmatrix} = \begin{bmatrix} w_{P0,0} & w_{P0,1} \\ w_{P1,0} & w_{P1,1} \\ w_{A0,0} & w_{A0,1} \\ w_{A1,0} & w_{A1,1} \end{bmatrix} \begin{bmatrix} x_0(n) \\ x_1(n) \end{bmatrix}$$

**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

[21] 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.
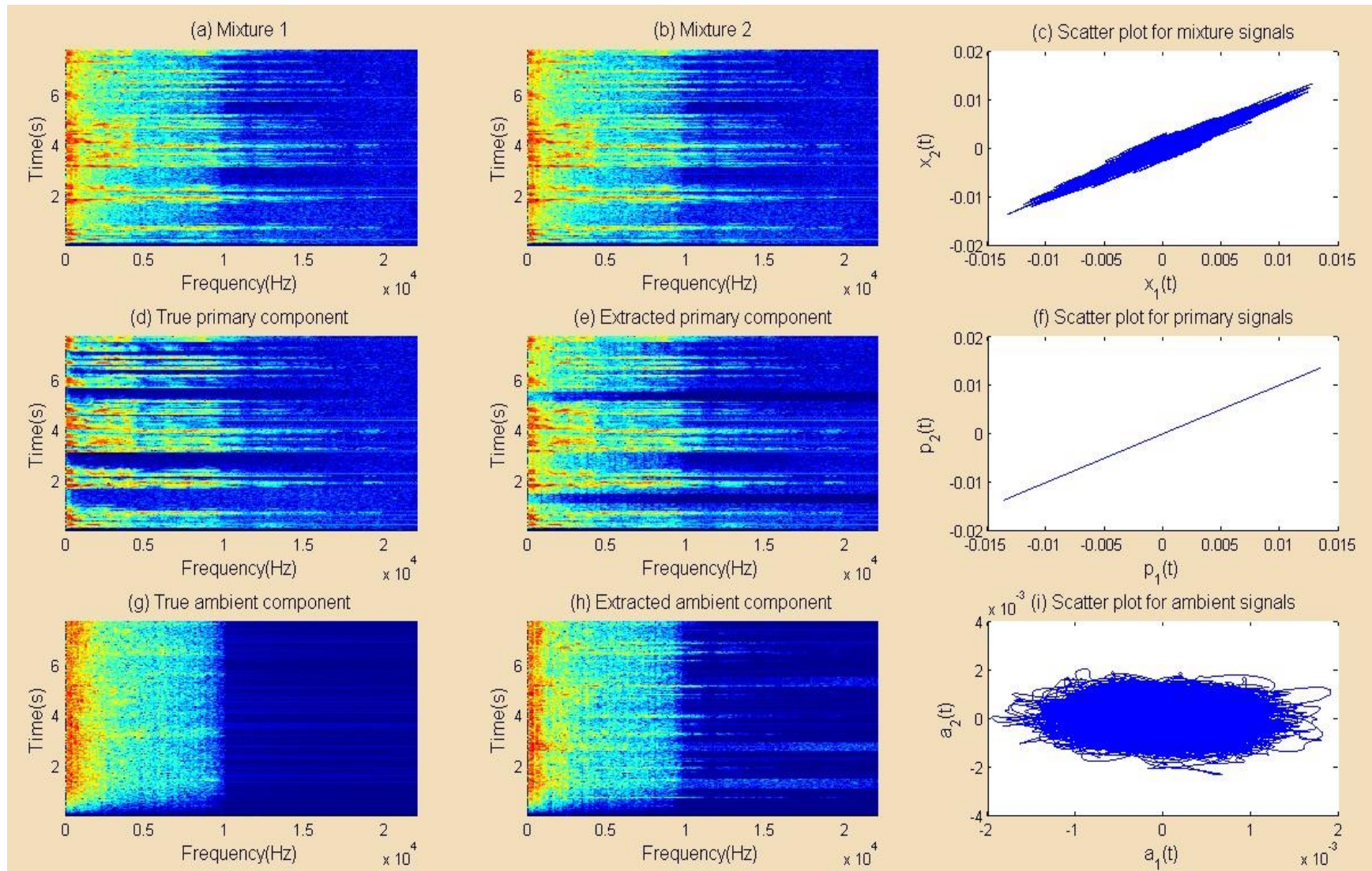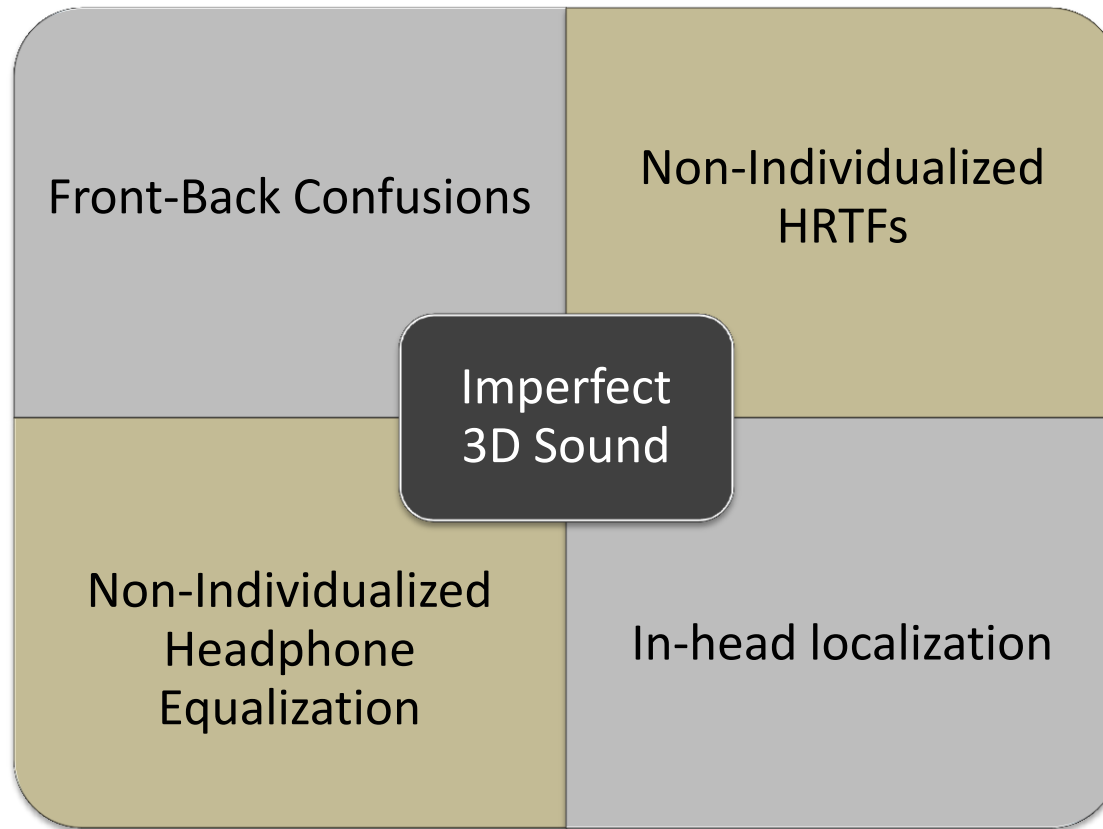
Individualization
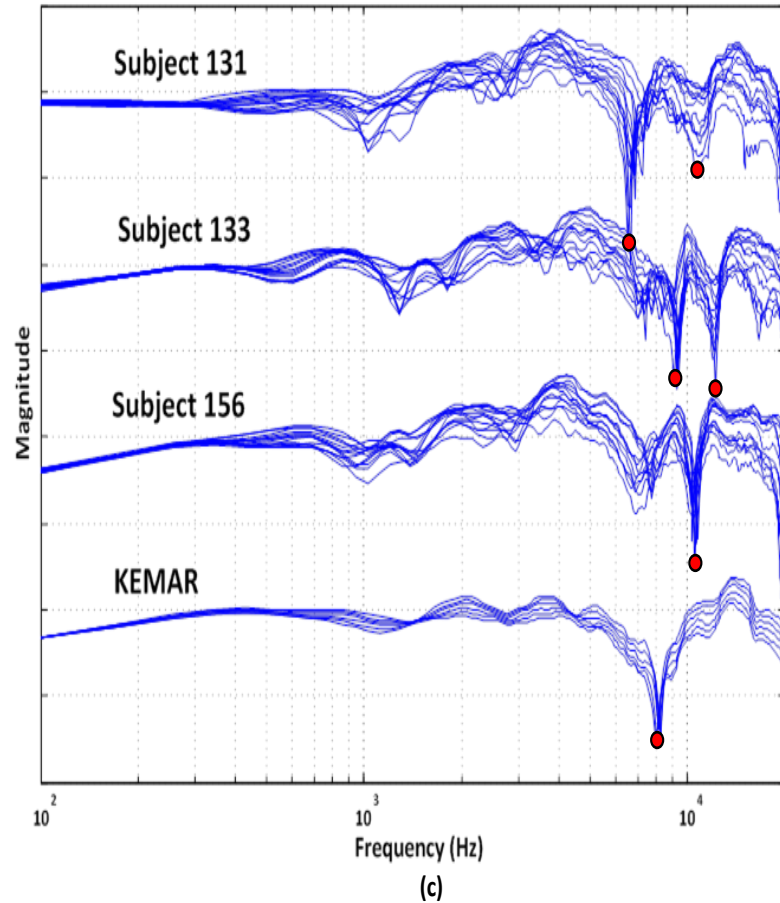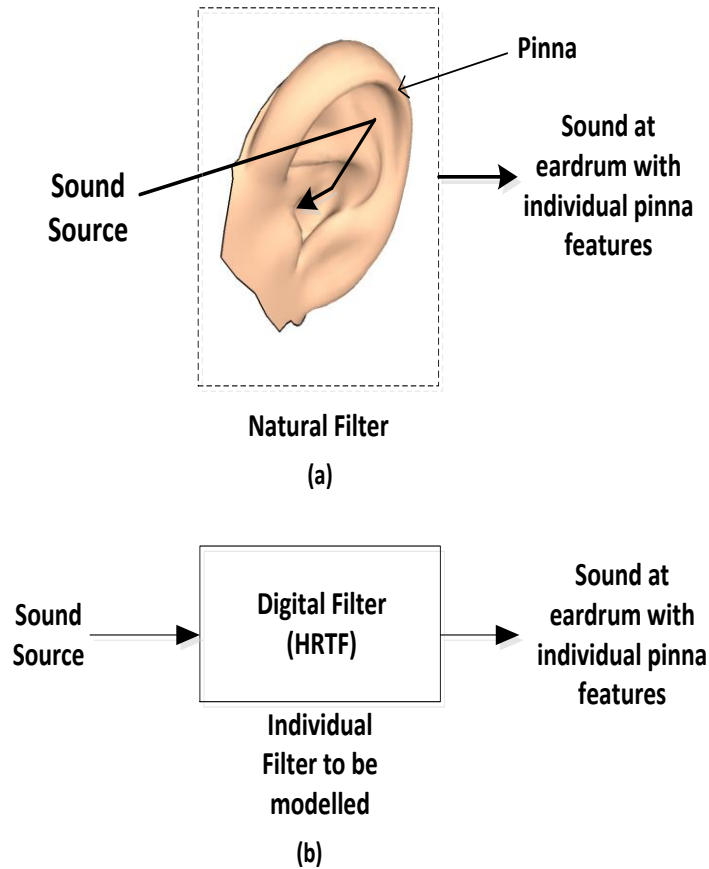
Variation of HRTFs (Idiosyncratic)

[26] S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: a review," in *Virtual Reality*, ed: Springer, 2007, pp. 397-407.

# Why is individualization necessary?

Front-Back Confusions

Non-Individualized HRTFs

Imperfect 3D Sound

Non-Individualized Headphone Equalization

In-head localization

Use of non-individual HRTFs degrades the veracity of the perception of 3D sound

Individualization

How to obtain Individual Features	Techniques	Pros and Cons	Performance
Acoustical Measurements	Individual measurements [25], IRCAM France, CIPIC, Tohuko uni., etc.	Ideal, accurate Tedious, requires high precision	Reference for individualization techniques
Anthropometric data	Optical Descriptors : 3D mesh, 2D pictures ; Numerical Solutions : PCA, FEM,BEM, ANN	Need a large database; Requires high resolution imaging; Expensive	Uses the correlation between individual HRTF and anthropometric data
Listening/ Training	PCA weight tuning, Tune magnitude Spectrum, Selection from Non-individualized HRTF database	directly relates to perception; requires regular training;	Obtains the best HRTFs perceptually
Playback Mode	Frontal Projection Headphone	No additional measurement, Type-2 EQ	Automatic customization, reduced front-back confusions
Non-individualized HRTF	Generalized HRTF	Easy to implement, Poor localization	Not an individualization technique

Frontal Projection Headphones

Conchae

FRONT EMITTER

SIDE EMITTER
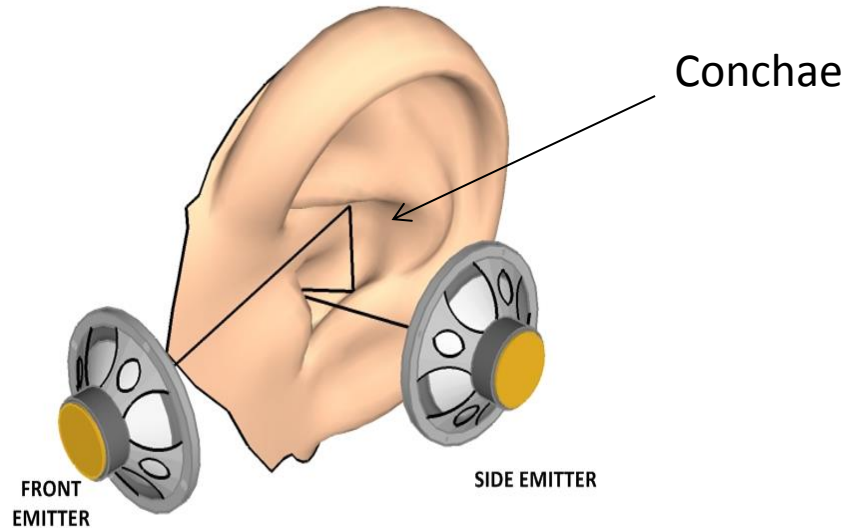
Motivation

➢ To accurately reproduce 3D audio over headphones catering to any individual without using individualized binaural synthesis

➢ To overcome the front-back confusions using non-individual HRTFs and thus improve the frontal image of the virtual auditory space
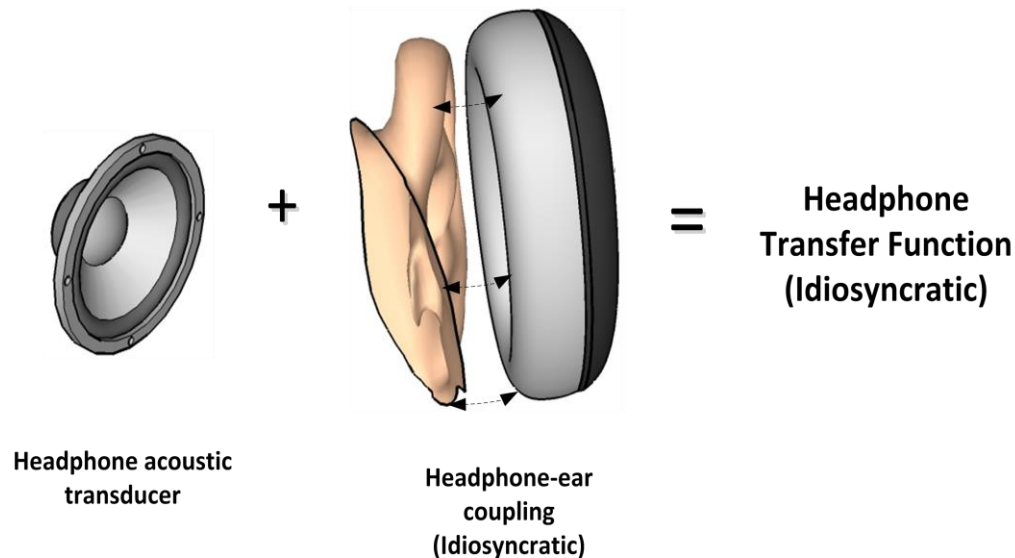
[33] K. Sunder, E. L. Tan, and W. S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.,* vol. 61, no. 12, pp. 989-1000, Dec. 2013.

Equalization

Headphone is not acoustically transparent:

1) Headphone colors the input sound spectrum;
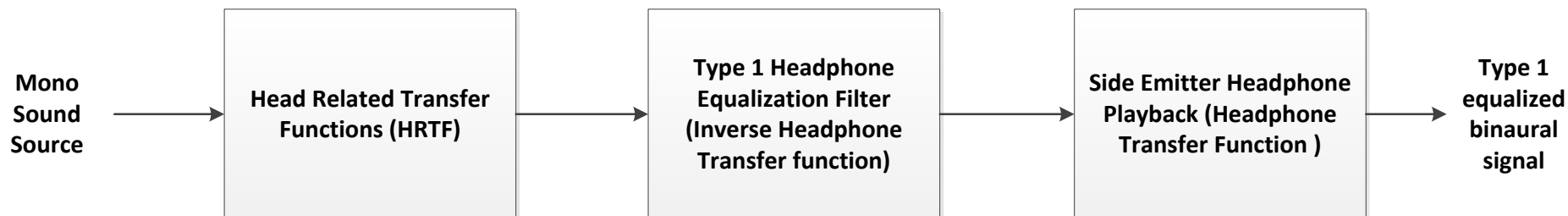2) Affects the free-field characteristics of the sound pressure at the ear

Breakdown of headphone transfer function (HPTF)

Equalization for binaural and stereo

Mode of Equalization	Aim	Types of Equalization and Target Response	Characteristics
Non-Decoupled (Binaural)	Spectrum at eardrum is the individual HRTF features	Conventional equalization (flat target response)	The spectrum at the eardrum has individual features (if individualized HRTF is used) Dependent on the individual's unique pinna features
		Type-2 equalization	Removes only the distortion due to the headphone emitter Independent of the idiosyncratic features of the ear
Decoupled (Binaural, Stereophony)	Emulate the most natural reproduction closer to the perception in a reference field	Free-field equalization (FF)	Target response is the free-field response corresponding to the frontal incidence
		Diffuse-field equalization (DF), Weighted DF, Reference Listening Room	Target response is the diffuse-field response, or response of a reference room Lesser inter-individual variability

Conventional Equalization (Type 1 EQ)

➢ Headphone is not acoustically transparent, therefore the effect of the headphone must be removed.

| Mono Sound Source | → | Head Related Transfer Functions (HRTF) | → | Type 1 Headphone Equalization Filter (Inverse Headphone Transfer function) | → | Side Emitter Headphone Playback (Headphone Transfer Function) | → | Type 1 equalized binaural signal |

*Equalization process : Removing the headphone transfer function*

$$Y(\omega) = S(\omega) \cdot HRTF(\omega) \cdot \frac{1}{HPTF(\omega)} \cdot HPTF(\omega)$$

Where, $Y(\omega)$ = Equalized Binaural Signal
$S(\omega)$ = Source Signal Spectrum
$HRTF(\omega)$ = Head Related Transfer Function (Left/Right)
$HPTF(\omega)$ = Headphone Transfer Function (Left/Right)

And, $\frac{1}{HPTF(\omega)}$ = Equalization Filter

Type 2 EQ (Frontal Projection Headphones)

➤ Equalizing to the free field response of the headphone with the ear-cup.

➤ Does not include headphone-ear coupling.

➤ Reflections/diffractions created by the interactions with the pinna due to the frontal projection are important and should be retained.

Mono Sound Source → Head Related Transfer Functions (HRTF) → Type 2 Equalization Filter (Inverse Free Field Response) → Frontal Emitter Headphone Playback (Headphone Transfer Function) → Type 2 equalized binaural signal
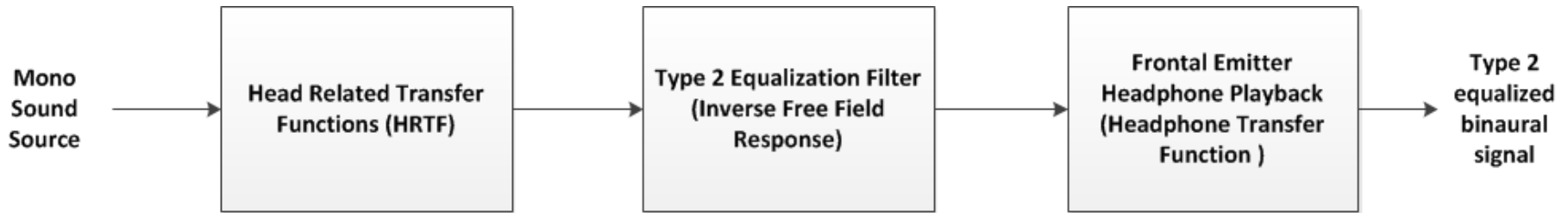
$$Y(\omega) = S(\omega) \cdot HRTF(\omega) \cdot \frac{1}{FFR(\omega)} \cdot HPTF(\omega)$$

$$HPTF(\omega) = FFR(\omega) \cdot PC(\omega)$$

Where, $Y(\omega)$ = Equalized Binaural Signal
$S(\omega)$ = Source Signal Spectrum
$HRTF(\omega)$ = Head Related Transfer Function (Left/Right)
$HPTF(\omega)$ = Headphone Transfer Function (Left/Right)
$FFR(\omega)$ = Free Field Response of the Frontal Emitter
$PC(\omega)$ = Personalized Pinna Cues generated by frontal projection

Free-air Equivalent Coupling Headphones

Presence of headphones affects the free-field characteristics of the sound pressure at the ear

$$G = \left(\frac{1}{M * HPTF}\right) * PDR,$$

G = Electrical Transmission Gain of the headphone
M = Microphone Transfer Function
HPTF = Headphone Transfer Function
PDR = Pressure Division Ratio

$$PDR = \frac{Z_{earcanal} + Z_{headphones}}{Z_{earcanal} + Z_{radiation}},$$

$Z_{radiation}$ = Free air radiation impedance as seen from the ear canal

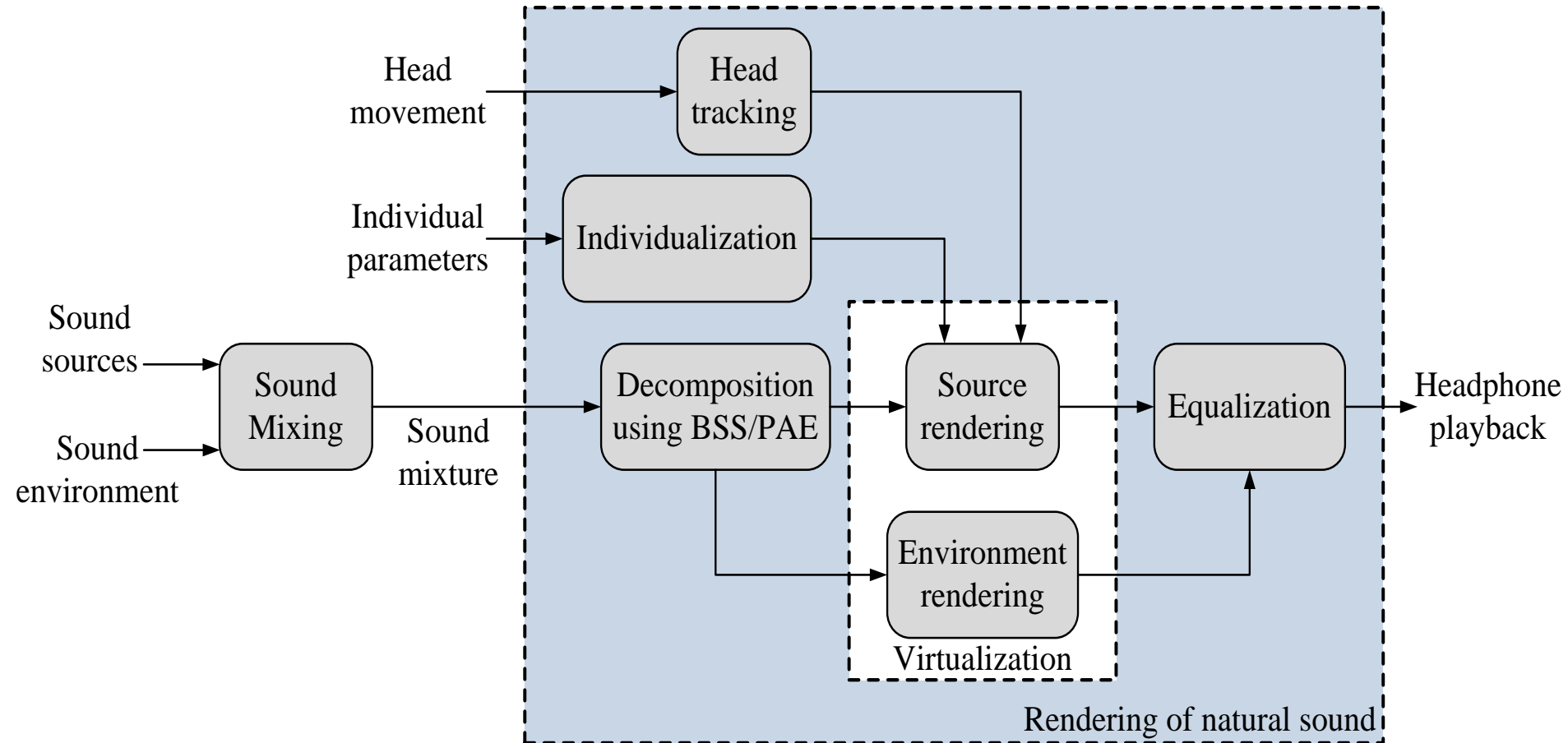$Z_{earcanal}$ = Impedance of the ear canal

$Z_{headphones}$ = Impedance of the headphones

PDR = 1 indicates that the pressure in the free field and in the presence of the headphones are equal (FEC headphone)

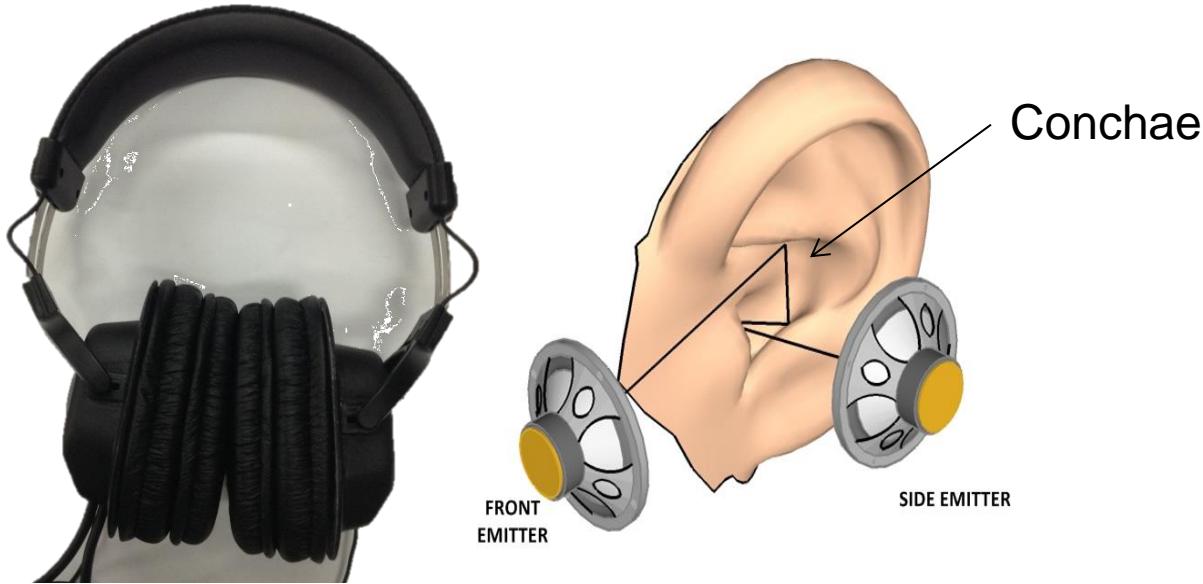BALL headphones : Any headphones at a certain distance from the ear
K1000 M, K1000 2 (AKG), DT 990 (Beyerdyanamic), Stax SR LAMBDA have close to FEC characteristics
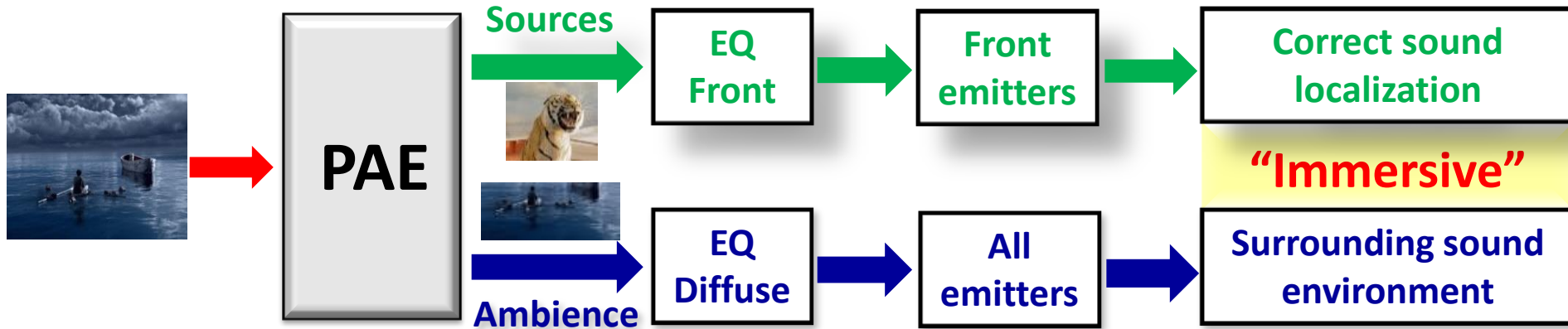
Integration

[40] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones," in press, IEEE Signal Processing Magazine, Mar. 2015.

3D Headphone: an example

FRONT EMITTER

SIDE EMITTER

Conchae

Key features
- Patented structure with strategic-positioned emitters;
- Individualization via frontal projection; no measurement or training required;
- Recreate an immersive perception of sound objects with surrounding ambience;
- Compatible with all existing sound formats.

PAE

Sources → **EQ Front** → **Front emitters** → **Correct sound localization**

"Immersive"

Ambience → **EQ Diffuse** → **All emitters** → **Surrounding sound environment**

[39] W. S. Gan and E. L. Tan, "Listening device and accompanying signal processing method," US Patent 2014/0153765 A1, 2014.

Subjective evaluation

➢ Conventional stereo system: stereo headphone
➢ Natural sound rendering system: 3D headphone
➢ Stimuli: binaural, movie and gaming tracks;
➢ 4 measures: Sense of direction, externalization, ambience, and timbral quality;
➢ 18 subjects, score of 0-100, and overall preference.

(a) MOS for 4 measures

(b) Scatter plot for all scores

(c) Preference of the tracks
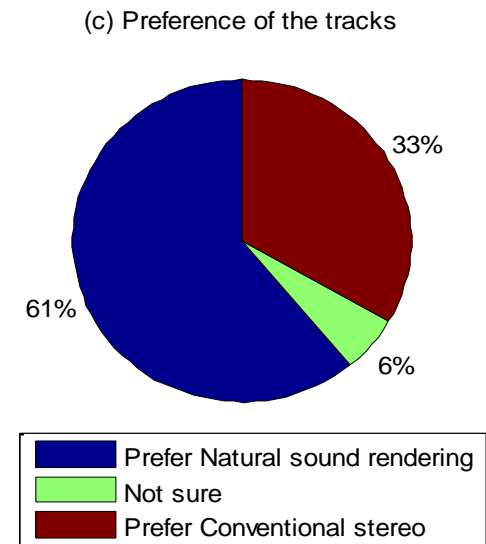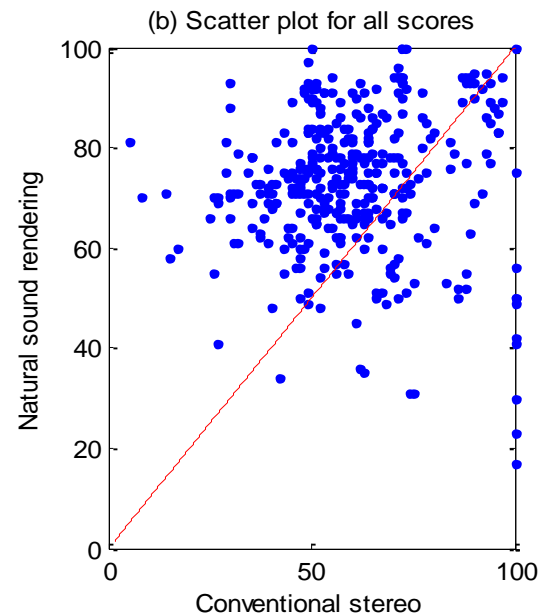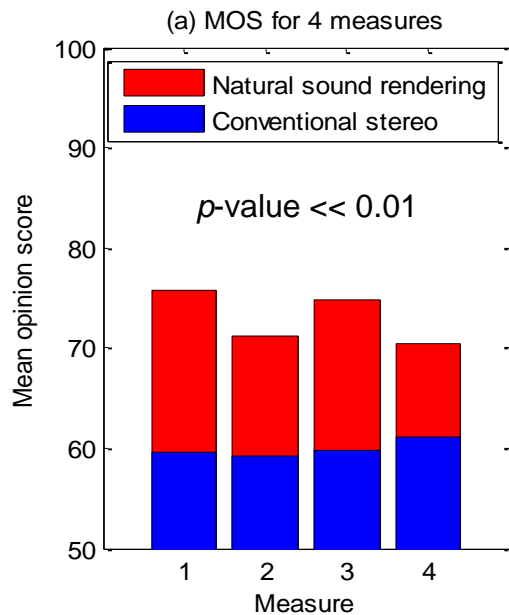
Conclusions

With these signal processing techniques applied in the sound rendering, headphone listening of digital media is becoming more natural, which assists and immerses listeners in the recreated sound scene, as if they were "being there".

➢ Advent of low cost, low power, small factor, and high speed multi-core embedded processor.

➢ 3D headphone: one example of such natural sound rendering system.

➢ Improved performance verified physically and validated psychophysically, compared to conventional headphone listening.

Future trends

- **From virtual reality to augmented reality**: integrate microphones and other sensors;

- **Headphone design (hardware):** more natural response, less coloration;

- **Headphone rendering (software):** object-based audio, prior mixing and environment information, advanced signal processing techniques, psychoacoustics.

- **A collaboration effort from the whole audio community!**

Future of headphone listening:

More intelligent and assistive, content-aware, location-aware, listener-aware.

References

[1] D. R. Begault, *3-D sound for virtual reality and multimedia*: AP Professional, 2000.

[2] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE,* vol. 101, no. 9, pp. 1920-1938, Sep. 2013.

[3] V. Pulkki, "Spatial sound reproduction with directional audio coding," *J. Audio Eng. Soc.,* vol. 55, no.6, pp. 503-516, Jun. 2007.

[4] S. Olive, T. Welti, and E. McMullin, "Listener Preferences for Different Headphone Target Response Curves," in *Proc. 134th Audio Engineering Society Convention*, Rome, Italy, May 2013.

[5] M. M. Goodwin and J. M. Jot, "Binaural 3-D audio rendering based on spatial audio scene coding," in *Proc. 123rd Audio Engineering Society Convention*, New York, Oct. 2007.

[6] J. Breebaart and E. Schuijers, "Phantom materialization: a novel method to enhance stereo audio reproduction on headphones," *IEEE Trans. Audio, Speech, and Language Processing,* vol. 16, no.8, pp. 1503-1511, Nov. 2008.

[7] C. Avendano and J. M. Jot, "A frequency-domain approach to multichannel upmix," *J. Audio Eng. Soc.,* vol. 52, no.7/8, pp. 740-749, Jul. 2004.

[8] C. Faller, "Multiple-loudspeaker playback of stereo signals," *J. Audio Eng. Soc.,* vol. 54, no.11, pp. 1051-1064, Nov. 2006.

[9] F. Menzer and C. Faller, "Stereo-to-binaural conversion using interaural coherence matching," in *Proc. 128th Audio Engineering Society Convention,* London, UK, May 2010.

[10] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Eng. Soc.,* vol. 49, no. 10, pp. 904-916, Oct. 2001.

[11] 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.

[12] V. R. Algazi and R. O. Duda, "Headphone-based spatial sound," *Signal Processing Magazine, IEEE,* vol. 28, no. 1, pp. 33-42, Jan. 2011.

[13] R. Nicol, *Binaural Technology*: AES, 2010.

[14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent component analysis*. New York: John Wiley & Sons, 2004.

[15] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proc. IEEE,* vol. 98, no. 6, pp. 995-1005, Jun. 2010.

[16] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Processing,* vol. 52, no.7, pp. 1830-1847, Jul. 2004.

References

[17] T. Virtanen, "Sound source separation in monaural music signals," PhD Thesis, Tampere University of Technology, 2006.

[18] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation," *IEEE Signal Processing Magazine,* vol. 31, no. 3, pp. 107-115, 2014.

[19] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications.* NJ: Wiley-IEEE Press, 2006.

[20] J. Merimaa, M. M. Goodwin, and J. M. Jot, "Correlation-based ambience extraction from stereo recordings," in *Proc. 123rd Audio Engineering Society Convention,* New York, Oct. 2007.

[21] 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.

[22] J. He, E. L. Tan, and W. S. Gan, "Time-shifted principal component analysis based cue extraction for stereo audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'13)*, Canada, May 2013, pp. 266-270.

[23] J. He, E. L. Tan, and W. S. Gan, "A study on the frequency-domain primary-ambient extraction for stereo audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'14)*, Florence, Italy, 2014, pp. 2892-2896.

[24] J. Thompson, B. Smith, A. Warner, and J. M. Jot, "Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations," in *Proc. 133rd Audio Eng. Soc. Conv.*, San Francisco, 2012.

[25] H. Møller, M. F. Sørensen, D. Hammershøi, and C. B. Jensen, "Head-related transfer functions of human subjects," *J. Audio Eng. Soc.,* vol. 43, no. 5, pp. 300-321, May 1995.

[26] S. Xu, Z. Li, and G. Salvendy, "Individualization of head-related transfer function for three-dimensional virtual auditory display: a review," in *Virtual Reality*, ed: Springer, 2007, pp. 397-407.

[27] G. Enzner, "3D-continuous-azimuth acquisition of head-related impulse responses using multi-channel adaptive filtering," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA),* Oct. 2009, pp. 325-328.

[28] M. Rothbucher, M. Durkovic, H. Shen, and K. Diepold, "HRTF customization using multiway array analysis," in *Proc. 18th European Signal Processing Conference (EUSIPCO'10)*, Aalborg, August 2010, pp. 229-233.

[29] R. O. Duda, V. R. Algazi, and D. M. Thompson, "The use of head-and-torso models for improved spatial sound synthesis," in *Proc. 113th Audio Engineering Society Convention*, Los Angeles, Oct. 2002.

[30] D. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis, "HRTF personalization using anthropometric measurements," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, New York, Oct. 2003, pp. 157-160.

References

[31] J. C. Middlebrooks, "Individual differences in external-ear transfer functions reduced by scaling in frequency," *J. Acoust. Soc. Amer.,* vol. 106, no. 3, pp. 1480-1492, Sep. 1999.

[32] K. J. Fink and L. Ray, "Tuning principal component weights to individualize HRTFs," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'12)*, Kyoto, Mar. 2012, pp. 389-392.

[33] K. Sunder, E. L. Tan, and W. S. Gan, "Individualization of binaural synthesis using frontal projection headphones," *J. Audio Eng. Soc.,* vol. 61, no. 12, pp. 989-1000, Dec. 2013.

[34] A. Bondu, S. Busson, V. Lemaire, and R. Nicol, "Looking for a relevant similarity criterion for HRTF clustering: a comparative study," in *Proc. 120th Audio Engineering Society Convention*, Paris, France, May 2006.

[35] H. Møller, D. Hammershoi, C. B. Jensen, and M. F. Sorensen, "Transfer characteristics of headphones measured on human ears," *J. Audio Eng. Soc.,* vol. 43, no. 4, pp. 203-217, Apr. 1995.

[36] V. Larcher, J. M. Jot, and G. Vandernoot, "Equalization methods in binaural technology," in *Proc. 105th Audio Engineering Society Convention,* SanFrancisco, Sep. 1998.

[37] A. Kulkarni and H. S. Colburn, "Variability in the characterization of the headphone transfer-function," *J. Acoust. Soc. Amer.,* vol. 107, no. 2, pp. 1071-1074, Feb. 2000.

[38] H. Møller, C. B. Jensen, D. Hammershøi, and M. F. Sørensen, "Design criteria for headphones," *J. Audio Eng. Soc.,* vol. 43, no. 4, pp. 218-232, Apr. 1995.

[39] W. S. Gan and E. L. Tan, "Listening device and accompanying signal processing method," US Patent 2014/0153765 A1, 2014.

[40] K. Sunder, J. He, E. L. Tan, and W. S. Gan, "Natural sound rendering for headphones," in press, IEEE Signal Processing Magazine, Mar. 2015.

Acknowledgements

We wish to thank the subjects that participated in our subjective experiments!

Email: ewsgan@ntu.edu.sg