A Source Separation Evaluation Method in Object-Based Spatial Audio

Qingju LIU*, Wenwu WANG*, Philip J. B. JACKSON*, Trevor J. COX†

*University of Surrey, †University of Salford , {q.liu, w.wang, p.jackson}@surrey.ac.uk, t.j.cox@salford.ac.uk

Future Spatial Audio for Immersive Listener Experience at Home Website: http://www.s3a-spatialaudio.org/

EUSIPCO 2015, Nice, Côte d'Azur, France, 31st August – 4th September 2015

What is object-based spatial audio (OSA)?

- Spatial audio (stereo, surround) gives the listener immersive spatial information, e.g. where the sound sources are and how reverberant the listening environment is.
- In OSA, sound scenes are represented in object format, e.g. each sound source is an audio object.

Why source separation (SS)?

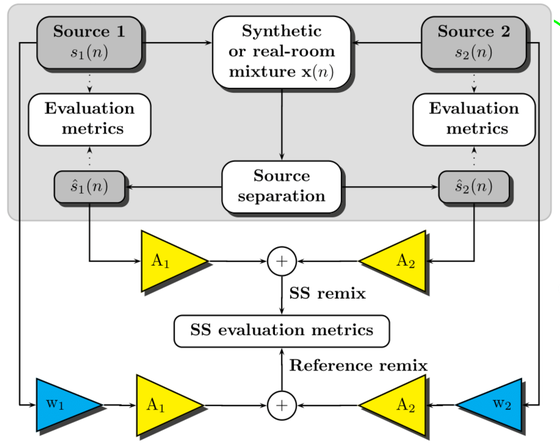
- SS provides a potentially useful and enabling tool for audio object extraction, e.g. blind source separation (BSS), beamforming and computational auditory scene analysis (CASA).

Limitations of existing SS evaluation methods

- Do not take into account the sound field reconstruction process.

Proposed evaluation method

Framework of the proposed SS evaluation method for object-based spatial audio



The conventional SS evaluation framework is highlighted in the shadowed area.

Amplitude panning

$A_1 + A_2 = 1$

The quality of the separated sources may not be good enough in terms of the evaluations using traditional metrics, but when they are remixed for spatial audio reproduction, perceptual quality of the generated spatial sound may well be satisfactory.

Principles of the proposed evaluation method

In spatial audio, we aim to reconstruct a sound field with a high quality, where the separated audio objects are likely to be **mixed down** using different rendering techniques such as stereo, surround, high order ambisonics (HOA) and wave field synthesis (WFS).

The benefit of introducing this re-mixing process is to relate the evaluation directly to what the listener hears when the source gains are adjusted in a remix.

1. Generate a new mixture (SS remix) to model the rendering process, where each source estimate is amplified and added together.
2. A reference mixture (reference remix) is obtained using the same remixing process.
3. Then the SS remix and the reference remix are compared using conventional SS metrics.

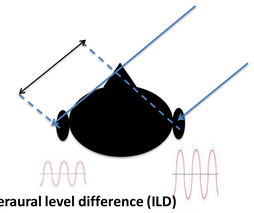
Experiments

Implementation of the baseline source separation methods

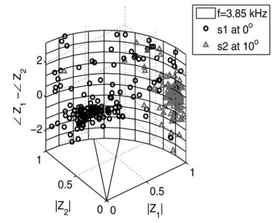
Two blind source separation (**BSS**) methods: "**Alinaghi**" and "**Sawada**"

A. Alinaghi, P. J. Jackson, Q. Liu, and W. Wang, "Joint mixing vector and binaural model based stereo source separation," IEEE/ACM Trans. Audio, Speech, Language Process. (ASLP), vol. 22, no. 9, pp. 1434–1448, September 2014.

H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," IEEE Trans. ASLP, vol. 19, no. 3, pp. 516–527, March 2011.



Interaural level difference (ILD)
Interaural time/phase different (ITD/IPD)



Mixing vector cues (MV)

Two beamforming methods: delay and sum (**DS**) and minimum variance distortionless response(**MVDR**)

DS is signal-independent, which directly compensates the delay from the target to each microphone.

MVDR is signal-dependent, where signal covariance estimation is involved for spatial filter calculation.

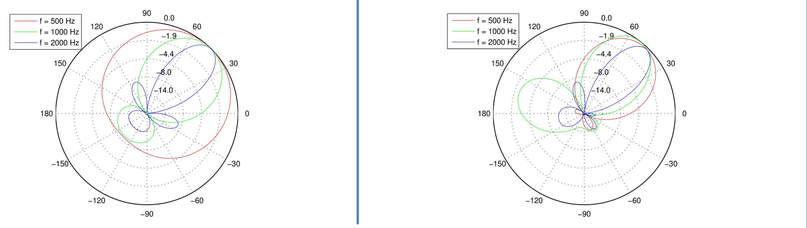
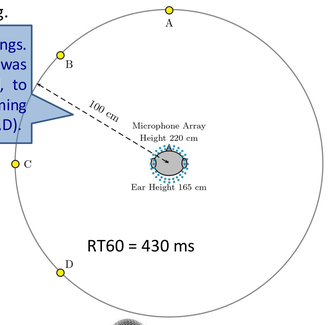


Illustration of the two beamforming algorithms enhancing sources from the 45 azimuth. For MVDR, the mixtures are generated by two concurrent speakers at azimuths 0 and 45 degree.

Data recording in VML

Two female speakers read randomly-chosen TIMIT sentences continuously for approximately 30 seconds at Position (A,B). This process was repeated twice for position pairs (A,C) and (A,D). Cortex MK2 and a 48-channel microphone array as well as some clip-on spot microphones are used for recording.

Setup for real-room speech recordings. The 48-channel microphone array was hung right above the dummy head, to record concurrent speech signals coming from position pairs (A,B), (A,C) and (A,D).



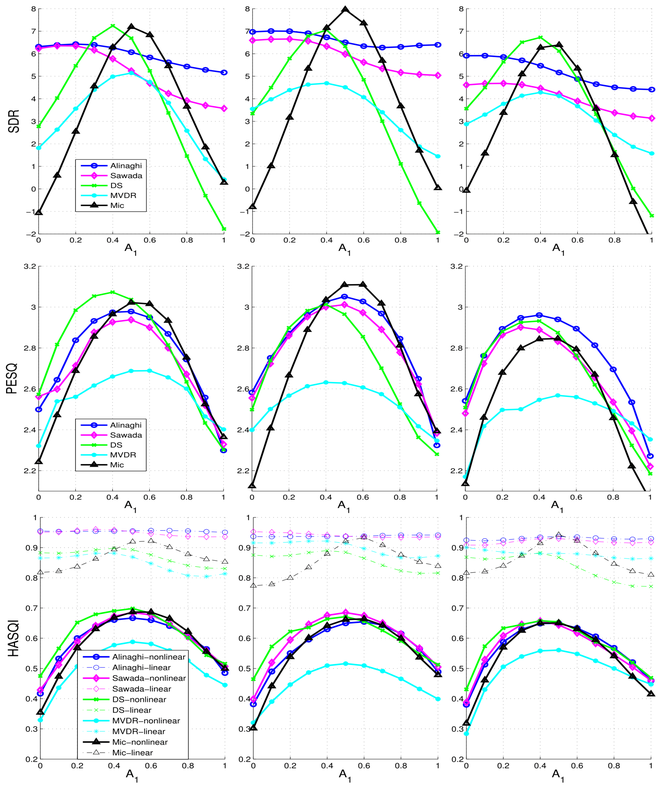
Cortex Manikin MK2 binaural head and torso simulator (Cortex MK2)

Spot microphone (clip-on, ground truth)

48-channel microphone array

Experimental results and analysis

Three different conventional SS evaluation metrics were integrated into our framework: signal to distortion ratio (SDR), perceptual evaluation of speech quality (PESQ) and hearing aid speech quality index (HASQI).



- The quality of the reconstructed sound field is similar to the quality of the isolated source estimate for BSS in terms of SDR.
- Beamforming remix gains a better quality than its separated sources, since the residual artefacts are masked by the reference mix.
- Source estimates after remix yield a better quality in terms of PESQ.
- HASQI-nonlinear is consistent with PESQ. HASQI-linear is consistent with SDR.

Conclusions, challenges and future work

- A new SS evaluation method in the context of spatial audio object separation.
- The conventional SS evaluation metrics are integrated into our proposed scheme.
- The proposed framework can be extended to scenarios with more than two sound sources.
- Experimental results show that remixed signals have the potential to deliver a higher quality as compared to the isolated source estimates, due to masking of residual artefacts.
- What kind of cues should be exploited to **develop new SS methods** that deliver a better reconstructed sound field in a wide range?
- To integrate **spatial quality metering** into the proposed scheme.

Acknowledgments. This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.