A SOURCE SEPARATION EVALUATION METHOD IN OBJECT-BASED SPATIAL AUDIO

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

[*] Centre for Vision, Speech and Signal Processing
University of Surrey, UK

[†] Acoustics Research Centre
University of Salford, UK

ABSTRACT

Representing a complex acoustic scene with audio objects is desirable but challenging in object-based spatial audio production and reproduction, especially when concurrent sound signals are present in the scene. Source separation (SS) provides a potentially useful and enabling tool for audio object extraction. These extracted objects are often remixed to reconstruct a sound field in the reproduction stage. A suitable SS method is expected to produce audio objects that ultimately deliver high quality audio after remix. The performance of these SS algorithms therefore needs to be evaluated in this context. Existing metrics for SS performance evaluation, however, do not take into account the essential sound field reconstruction process. To address this problem, here we propose a new SS evaluation method which employs a remixing strategy similar to the panning law, and provides a framework to incorporate the conventional SS metrics. We have tested our proposed method on real-room recordings processed with four SS methods, including two state-of-the-art blind source separation (BSS) methods and two classic beamforming algorithms. The evaluation results based on three conventional SS metrics are analysed.

Index Terms— Spatial audio, object-based, blind source separation, beamforming, evaluation

1. INTRODUCTION

Spatial audio provides immersive spatial information, e.g. where the sound sources are and how reverberant the environment is. Conventional spatial audio systems are often channel-based, where the auditory scene is represented by channel signals, which are transmitted to a specific reproduction system (e.g. a 5.1 loudspeaker array) to reconstruct the sound field. However, channel-based spatial audio lacks adaptivity to different reproduction systems, individual preference and listening environments. An emerging alternative to address the above limitations is object-based spatial audio, in which the auditory scene is represented by audio objects, with each audio object containing an audio stream as well as associated metadata [1]. A typical audio stream is a sound source, and the metadata describes properties of the sound source and the acoustic ambience, e.g. the 3D position of the sound source and the reverberation level of the environment. At the rendering (reproduction) stage, to reconstruct a sound scene, these audio objects are mixed down based on the reproduction system setup as well as the metadata. A listener may interact with the listening environment by manipulating the metadata.

An essential step in object-based spatial audio production is to represent the audio scene in terms of audio objects. This is challenging in real-room environments when there are concurrent sound signals. Source separation (SS) techniques can be applied to address this audio object separation problem, and there are many SS frameworks available. For instance, blind source separation (BSS) based on statistical cues such as mutual independence of sound sources [2] or spatial cues [3, 4]; beamforming methods [5, 6] based on the propagation model of sound signals; computational auditory scene analysis (CASA) [7] based on human auditory perception mechanisms.

A key question to ask is, however, that whether these SS techniques offer sufficient quality for object representation in spatial audio production and reproduction. Conventionally, SS algorithms are evaluated using the following metrics. For instance, signal-to-noise ratio (SNR)-based metrics such as (frequency-weighted) segmental SNR [8], weighted spectral slope measure [9], source to interference/artefact/distortion ratio (SIR, SAR, SDR) [10]; linear predictive coding (LPC)-based evaluations such as log-likelihood ratio (LLR) [11] and Itakura-Saito (IS) distance; auditory-motivated perceptual evaluation metrics such as perceptual evaluation of speech quality (PESQ) [12] and perceptual evaluation methods for audio source separation (PEASS) [13].

In spatial audio, however, the aim is to evaluate the quality of the reconstructed sound field, where the sources (audio objects) extracted via SS methods are manipulated and mixed down. Using the performance metrics mentioned above may not be able to truly assess the quality of the produced spatial audio. For instance, the quality of the separated sources may not be good enough in terms of the evaluations using the above metrics, but when they are remixed for spatial audio

The authors of the paper would like to acknowledge the support of 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.

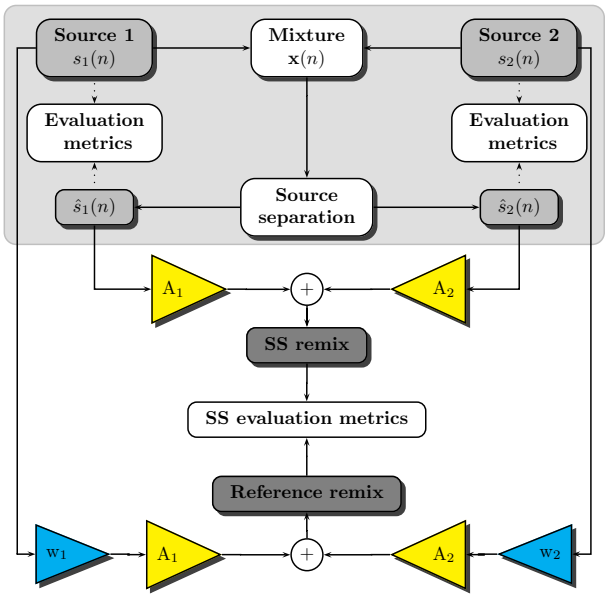

Fig. 1. Framework of the proposed SS evaluation method for object-based spatial audio. The conventional SS framework is highlighted in the shadowed area.

reproduction, the perceptual quality of the generated spatial sound may well be satisfactory. Therefore, to evaluate the performance of an SS algorithm in this context, an alternative metric is required. To this end, we propose a new method by comparing the remix of the separated sources (SS remix) with the ground truth remix from the original sources (reference remix). This strategy is similar to the amplitude panning law used for stereo sound. The previously-mentioned SS evaluation metrics are integrated into this method. More details of our method are introduced in the next section.

2. THE PROPOSED EVALUATION METHOD

We first introduce the framework of conventional source separation assessment. Take a 2×2 system as an example, the two original sources are denoted as $s_1(n)$ and $s_2(n)$, and their mixture is denoted as $\mathbf{x}(n)$. A SS method is applied to $\mathbf{x}(n)$ to obtain two source estimates $\hat{s}_1(n)$ and $\hat{s}_2(n)$. To evaluate the performance of the SS method, $\hat{s}_i(n)$ is directly compared with $s_i(n)(i = 1, 2)$ using existing SS evaluation metrics, assuming that $s_i(n)$ is known as a reference for performance evaluation. This framework is highlighted in the shadowed area in Figure 1.

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) [14] and wave field synthesis (WFS) [15]. Object-based spatial audio has the advantage of interactive listening, e.g., the listener can focus on one particular sound by turning up its volume and suppress the interfering sound. To evaluate the quality of the reconstructed sound field, a new SS evaluation method is proposed in this context, as shown in Figure 1. First we generate a new mixture (SS remix) to model the rendering process, where each source estimate is amplified and added together. Using the same remixing process, a reference mixture (reference remix) is obtained. Then the SS remix and the reference remix are compared using conventional SS metrics. Using again the 2×2 system as an example, the SS remix is obtained as $A_1\hat{s}_1(n) + A_2\hat{s}_2(n)$, s.t. $A_1 + A_2 = 1$, where A_i varies between [0, 1]. This strategy is similar to the classic amplitude panning [16] in spatial audio rendering. The reproduced sound field fades from $s_1(n)$ to $s_2(n)$ by decreasing A_1. When $A_1 = 1$, only the first source estimate is expected in the sound zone; when $A_1 = 0.5$, two source estimates are balanced. We need to stress that when $A_1 = 0$ or 1, the assessment is exactly the same as conventional SS evaluation methods. Note that, $\hat{s}_i(n)$ is a distorted version of $s_i(n)$ that $\hat{s}_i(n) \approx w_i * s_i(n)$ where $*$ denotes convolution, and w_i can be considered as a finite impulse response Wiener filter, whose estimation can be obtained via solving Wiener-Hopf equations. As a result, when generating the reference remix, we replace $s_i(n)$ with its contributions in $\hat{s}_i(n)$, i.e. $w_i * s_i(n)$, to cope with any short-term distortions and delays.

We have tested the proposed evaluation method on real-room speech recordings, where four different SS methods were used, and three existing SS evaluation metrics were integrated, as introduced in the next section.

3. EXPERIMENTS

3.1. SS algorithms

Two BSS algorithms and two classic beamforming algorithms were used for SS tasks.

Both BSS algorithms consider only time-invariant mixtures, i.e. sound sources are not moving. The first BSS algorithm, denoted as "Alinaghi" [3], works for stereo recordings. It is a time-frequency (TF) masking-based method, where the soft mask is generated based on the following three cues: interaural level difference (ILD), interaural phase difference (IPD) and mixing vectors (MV). A Gaussian mixture model (GMM) is applied to model these features for deriving the TF mask. The second BSS algorithm is denoted as "Sawada" [4]. With the sparsity assumption of speech signals at each TF point, the observation vector can be considered as a shifted version of the mixing vector associated with the dominant source, which can be probabilistically clustered to different sources. Assuming that the prior information of sound source number is available, both BSS algorithms were applied in the TF domain after 1024-point short time Fourier transform (STFT). "Alinaghi" initialises the GMM model based on the time delay estimation from the stereo recordings, then 16 expectation maximisation (EM) iterations are applied to update

these frequency-dependent GMM parameters in a bootstrap way. "Sawada" initialises the mixing vectors (MV) with k-means, and an EM algorithm is applied to update the MV cues with 50 iterations. Based on inter-frequency dependencies, the permutation problem is resolved before the time-domain reconstruction. These chosen parameters as used in [3] give satisfactory results under various reverberant conditions.

The two classic beamforming methods that we implemented are delay-and-sum (DS) and minimum variance distortionless response (MVDR) [5,6]. A beamformer requires a number of spatially distributed microphones, which can steer its beams to target directions for enhancement. DS depends on the positions of the microphones and the target sound, 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. Both beamforming methods were applied in the TF domain, with the same 1024-point STFT. When calculating the steering vector at each frequency bin, we used the ground truth positions of the sources and the microphone array. The power covariance was estimated from 200 segments with each segment lasting 20 ms. To avoid singular matrices, the estimated power covariance was compensated with an identity matrix scaled to the largest eigenvalue divided by 50.

3.2. Microphone setup

A 48-channel microphone array as well as the Cortex Manikin MK2 binaural head and torso simulator (Cortex MK2) were used to record data, shown in Figure 2, for beamforming methods and BSS methods respectively. The microphone array contains two circles with 24 microphones for each circle, with the inner and outer radius being 85 mm and 107 mm respectively. Both of the built-in microphones (NC-MK 231) in the dummy head and these in the microphone array (Countryman B3 Omnidirectional Lavalier) have smooth frequency responses ($<$ 1 dB variation) in the voice band of 300 Hz to 3400 Hz, which provides fair comparison for the BSS and beamforming technologies for speech signals. Besides that, two Countryman B3 microphones were used to record clean sound sources.

3.3. Data and recording setup

The recording room based in University of Surrey has a size of $244 \times 396 \times 242$ cm, with the reverberation time at about 430 ms. The dummy head stood in the centre of the room with ear height of 165 cm. The microphone array was hung on the ceiling, just above the dummy head at the height of 220 cm. Four positions were labelled as A, B, C and D, as shown in Figure 3, and their input azimuths relative to the dummy head are 0°, 45°, 90° and 135° respectively. Two female speakers were involved for recording data standing at positions A and B respectively, both reading randomly-chosen TIMIT sentences continuously for approximately 30

Fig. 2. The Cortex MK2 with built-in microphones at two ears and the 48-channel two-circular microphone array.

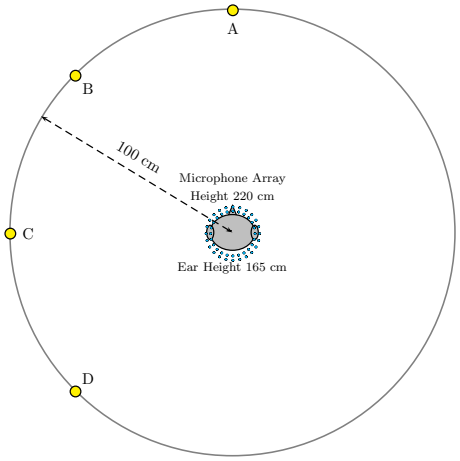

Fig. 3. Setup for real-room speech recordings. The 48-channel microphone array was hung right above the dummy head, to record concurrent speech signals coming from position pairs (A,B), (A,C) and (A,D).

seconds. This process was repeated twice for position pairs (A,C) and (A,D). Each subject wore a clip-on microphone to capture the ground truth[1]. The recorded data were sampled at 16 kHz, which covers the voiced band.

Then the previously introduced BSS and beamforming algorithms were applied to the dummy head mixtures and circular microphone-array mixtures respectively. After that, our proposed evaluation method is applied to these source estimates using the framework shown in Figure 1.

3.4. Results and analysis

The remix from the source estimates after SS and the reference remix from the ground truth were generated by changing A_1 from 0 to 1 with an increment of 0.1. Three different conventional SS evaluation metrics were integrated into our framework. The first one is signal-to-distortion ratio (SDR), which calculates the ratio of contributions from the reference

[1]Note that, the ground truth is not absolutely clean, since each close microphone might catch interfering information from the competing speaker.

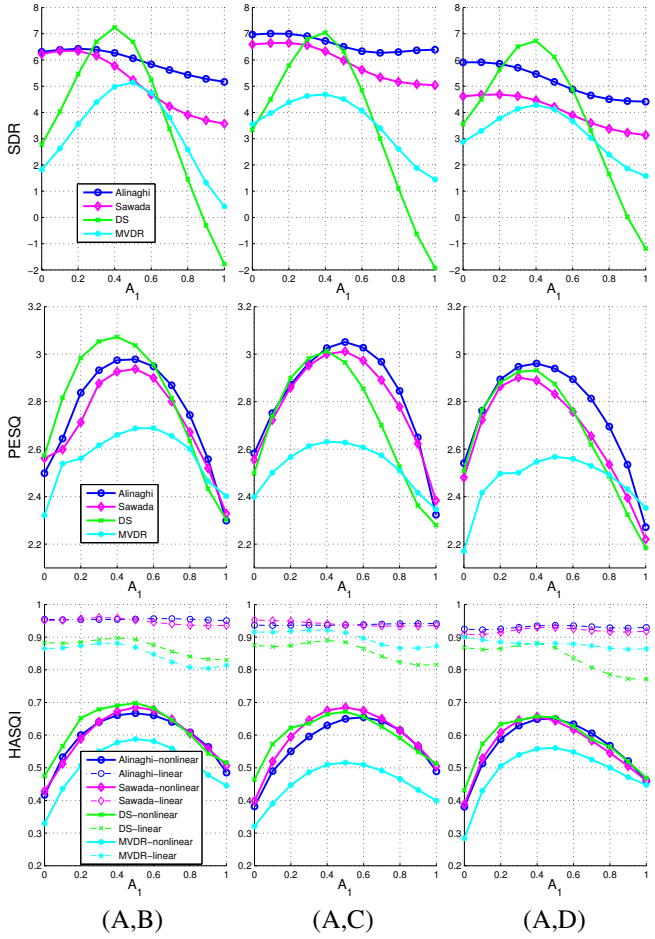

Fig. 4. The performance results of the SS algorithms evaluated by the proposed method. Three conventional SS evaluation metrics were integrated, which were SDR (row 1), PESQ (row 2) and HASQI (row 3) respectively. The proposed framework was tested on real-room recordings at three position pairs: (A,B) in column 1, (A,C) in column 2 and (A,D) in column 3.

remix to any other distortion components. The second one is perceptual evaluation of speech quality (PESQ) [12], which is auditory-motivated and widely used to evaluate the perceptual quality of speech signals. The third one is the hearing aid speech quality index (HASQI) [17], which copes with both normal-hearing and hearing-impaired listeners by adapting the cochlear model. The speech sound quality metric in HASQI was used, which has two terms: (1) the nonlinear distortion and (2) the linear distortion, introduced by short-term and long-term spectrum changes respectively.

The quantitative evaluation results are presented in Figure 4. First, we notice that the two BSS algorithms, denoted as "Alinaghi" and "Sawada", outperform the two beamforming algorithms in terms of SDR. In fact, the two beamformers fail to separate the sound sources, which can be seen by these very low SDR values at the two ends of these sub-plots in the top

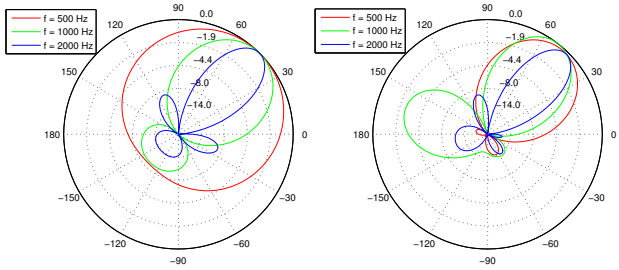

Fig. 5. Illustration of the two beamforming algorithms enhancing sources from the $45°$ azimuth. For the MVDR beamformer, the mixtures are generated by two concurrent speakers at azimuths $0°$ and $45°$ respectively.

row. In other words, the source components are embedded by the distortion corruption.

To explore reasons why these beamforming methods fail to separate sounds, we plotted their directivity patterns when the target beam direction is $45°$, as shown in Figure 5. Beam patterns vary at different frequency bins. For the DS beamformer, the main lobe points exactly at the target direction. However, the lobe width is big, especially for low frequencies, which means interfering components from the neighbouring directions are not sufficiently suppressed. For the MVDR beamformer, the beams are much narrower at low frequency bins, and they cross at one point in the target direction. However, the beam peaks are shifted away from the target direction for the following reason. The inverse of the power spectrum is complexed-valued, whose multiplication with the steering vector (from the target direction) results in the shift.

For the top row sub-plots in Figure 4, when the remixing parameter A_1 varies from 0 to 1, the SDR curves for BSS smoothly vary from one end to the other without much fluctuation. Note that, at the two ends, the remix contains information from only one source estimate. In other words, source estimates are compared directly with clean sources without remixing. From this curve, the quality of the reconstructed sound field is similar to the quality of the isolated source estimate. However, the SDR curves for beamforming first increase and then decrease dramatically. This is reasonable since the interference residual at each beamforming output can be partially considered as contributions from the reference remix after the two outputs are mixed down. In other words, the residual artefacts are masked by the reference mix.

Comparing the *linear distortion* measurements in HASQI (the dash-dot curves in the sub-plots of the bottom row, denoted as HASQI-linear) with the SDR results, we notice that they are consistent for BSS. This is because both SDR and HASQI-linear evaluate long-term distortions, with SDR on the signal magnitude in the time domain, and HASQI-linear on the signal envelope in the frequency domain. However, the remix advantage that the beamformers show in SDR almost disappears in HASQI-linear. This is because linear filtering

affects the HASQI-linear measurements, whilst the beamforming methods are essentially linear-filtering techniques. The soft masking-based BSS algorithms, on the other hand, are essentially nonlinear filtering techniques and therefore not affected.

However, SDR is not very consistent with subject speech quality evaluations. For instance, if we distort a signal by slowly lowering its volume, then we will get a very low SDR result, but the important information within the signal is not greatly affected. PESQ, the "prediction of the perceived quality that would be given by subjects in a subjective listening test" [12], addresses this limitation and gains more reliable results. We found that the source estimates after remix yield a better quality in terms of PESQ. Take the BSS measurement at position (A,B) in column 1 as an example, if we directly compare the two source estimates with their associated clean signals, we get the PESQ evaluations of about 2.5 and 2.3 respectively (results at two ends). However, if we remix them by taking their average ($A_1 = 0.5$), we get the PESQ result around 3. This phenomenon confirms that SS might fail to produce satisfactory results, but the reconstructed sound field from these source estimates may offer satisfactory perceptual quality. This also verifies that conventional SS evaluation metrics alone do not suffice for the evaluation of object-based representations.

The *nonlinear distortion* measurements in HASQI (the solid lines in the sub-plots of the bottom row, denoted as HASQI-nonlinear) are consistent with the PESQ results. This is reasonable since they both evaluate short-term distortions, with PESQ on the perceptual model representations, and HASQI on the cochlear model, and both models are auditory-motivated.

4. SUMMARY

We have proposed a new SS evaluation method in the context of spatial audio object separation. Source estimates obtained by SS are mixed down using a strategy similar to the amplitude panning law. Then conventional SS evaluation metrics are applied to the remixed signals. 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. An arising question is what kind of cues should be exploited to develop new SS methods that deliver a better reconstructed sound field in a wide range, i.e., the range where we can vary the value A_1 without sacrificing performance. This requires further study in the future.

REFERENCES

[1] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H audiothe new standard for universal spatial/3D audio coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, Dec. 2014.

[2] P. Comon, "Independent component analysis, a new concept?," *Sign. Proces.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[3] 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, Sept. 2014.

[4] 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, Mar. 2011.

[5] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[6] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1702–1715, July 2003.

[7] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[8] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. ASLP*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[9] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, May 1982, vol. 7, pp. 1278–1281.

[10] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL Toolbox User Guide – Revision 2.0," Technical report, 2005.

[11] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice Hall Englewood Cliffs, NJ, 1988.

[12] "ITU-T Rec.P. 862," Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

[13] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. ASLP*, vol. 19, no. 7, pp. 2046–2057, Sept. 2011.

[14] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format," in *Int. Conf. Signal Process. Audio Recording and Reproduction*, May 2003.

[15] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. Acoust. Soc. Am.*, vol. 93, no. 5, pp. 2764–2778, 1993.

[16] A. D. Blumlein, "British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems)," *J. Audio Eng. Soc.*, vol. 6, no. 2, pp. 91–98, Apr. 1958.

[17] J. M. Kates and K. H. Arehart, "The hearing-aid speech quality index (HASQI)," *J. Audio Eng. Soc.*, vol. 58, no. 5, pp. 363–381, May 2010.