Perceptual Evaluation of blind source separation in object-based audio production

Philip Coleman[1,2], Qingju Liu[2], Jon Francombe[1,3], and Philip J. B. Jackson[2]

[1] Institute of Sound Recording, University of Surrey, UK
p.d.coleman@surrey.ac.uk
[2] Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[3] BBC Research & Development, Salford, UK

Abstract. Object-based audio has the potential to enable multimedia content to be tailored to individual listeners and their reproduction equipment. In general, object-based production assumes that the objects—the assets comprising the scene—are free of noise and interference. However, there are many applications in which signal separation could be useful to an object-based audio workflow, e.g., extracting individual objects from channel-based recordings or legacy content, or recording a sound scene with a single microphone array. This paper describes the application and evaluation of blind source separation (BSS) for sound recording in a hybrid channel-based and object-based workflow, in which BSS-estimated objects are mixed with the original stereo recording. A subjective experiment was conducted using simultaneously spoken speech recorded with omnidirectional microphones in a reverberant room. Listeners mixed a BSS-extracted speech object into the scene to make the quieter talker clearer, while retaining acceptable audio quality, compared to the raw stereo recording. Objective evaluations show that the relative short-term objective intelligibility and speech quality scores increase using BSS. Further objective evaluations are used to discuss the influence of the BSS method on the remixing scenario; the scenario shown by human listeners to be useful in object-based audio is shown to be a worse-case scenario.

1 Introduction

Research into blind source separation (BSS), where an estimate of a clean audio source can be obtained knowing only a mixture of sounds, has been active for many years. Generally, the performance of such approaches has been evaluated in terms of the quality of the estimated audio signal after processing, suppression of interference, and absence of artefacts, using tools such as BSS Eval [21] and PEASS [3]. However, in *remixing*, where the estimated audio signal is combined with other audio before being presented to the listener, some separation artefacts may be masked, increasing the utility for source separation techniques in contexts such as broadcast where high-quality content is required.

Opportunities for source separation are also emerging in the context of object-based audio. Here, instead of content being mastered for a particular production

format, a set of *audio objects* is transmitted. Audio objects usually comprise a single 'clean' audio channel and corresponding metadata that describe how the audio should ideally be rendered for the end user. Object-based audio allows for customisation of content to each listener's reproduction setup, and personalisation of content to their personal preferences. However, clean audio objects may not always be available. In this paper, we investigate applications of BSS for clean object estimation in the context of an object-based workflow.

Other recent work has also sought to exploit the potential for using BSS as part of a remix. In [12], perceptual model results were used to show that the speech quality achieved by remixing estimated sources was higher than the quality of the estimated sources in isolation. In [20], subjective tests were conducted to investigate the extent to which users were satisfied by personalising object-based content, with a source separation scenario considered. The MARuSS (Musical Audio Repurposing using Source Separation) project has worked on the problem of musical remix and upmix using deep learning-based BSS [16], including separation of vocals from the remainder of the mix [18], and perceptual evaluation of BSS in the context of remixing [17,22].

The work described in this paper extends the work by Coleman *et al.* [2, Sec. VII.C] in three ways. First, we investigate two additional BSS algorithms; second, we extend the presentation and discussion of the objective metrics; third, we evaluate the effects of remixing both talkers instead of just the quieter talker as in [2]. The paper is organised as follows. In Sec. 2, we motivate the use of source separation in the context of an object-based production workflow and present the background theory for the BSS approaches implemented. In Sec. 3, we present the results of subjective and objective experiments for speech stimuli. Finally, in Sec. 4 we conclude.

2 Background

In this section, the application scenario for BSS in object-based audio is described, and the BSS methods under test are briefly introduced.

2.1 Object-based production workflow

An object-based scene is composed of a number of audio signals, together with metadata describing how they should be rendered for the end user. Traditionally, it is assumed that the audio signals are clean, that is, not contaminated with artefacts or interference from other sources. Then, in a standard object-based workflow, metadata would be manually authored by the sound designer in post-production. This process is time consuming, both in terms of capturing the required source signals and authoring the metadata. Consequently, a new workflow stage of *objectification* has recently been proposed [2], wherein audio objects and their metadata are estimated from audio and video signals that may form part of the final audible production or may serve purely as production aids.

Although a strictly object-based production would encode each individual sound source as an object, a commonly used pragmatic approach is to mix close microphone object signals with a channel-based capture of the entire scene. This enables opportunities for editing, remixing or personalising content (compared to a traditional channel-based broadcast of the same mix) and is supported in current standards [4,7]. Furthermore, if close microphone signals are not available (for example, if there is limited time to set up equipment), BSS can potentially be used to estimate the object signals. In this case, remixing can still take place in post-production. Coleman *et al.* [2] explored two use-cases for audio separation algorithms in object-based production (BSS for speech; beamforming for music). The analysis of the results from the speech use case is extended here.

2.2 Blind source separation methods

Three BSS methods are considered in this paper. The first is a traditional time-frequency (TF) masking method statistically-characterised with a Gaussian mixture model (GMM), where binaural features of inter-aural level difference (ILD) and inter-aural phase difference (IPD) are exploited to iteratively refine the GMM parameters for the separation mask generation [13]. We denote this method as "Mandel". The second uses similar principles, yet takes into account ILD and IPD as well as mixing vector features [1], and is denoted as "Alinaghi". Unlike the above methods, with unsupervised learning processes, the third method is based on deep neural networks (DNNs), where the commonly used spectral features and non-linearly-transformed binaural spatial features are fed into a hybrid DNN structure, consisting of convolutional layers and fully-connected layers [11]. The spatial features are iteratively refined using the DNN output. This method is denoted as "Liu". The training process for Liu was performed on a simulated data set lasting around 12 hours in a reverberant room (RT60 640 ms). It is noteworthy that the mixing scenario for training the DNN used in Liu does not correspond to the conditions of the data recorded for the experiments reported in this paper: the talker positions, microphones, and balance between dominant and interfering speakers were all different.

3 Experiments

To investigate the utility of BSS to enable object-based remix of stereo speech content, listening tests were conducted, and objective scores were obtained using predictive perceptual models and signal-based metrics. In this section, the setup for each experiment is described and the results are presented and discussed.

3.1 Speech stimuli

Performances were recorded in a large recording studio (dimensions 14.55 × 17.08 × 6.50 m; RT60 1.1 s) using a number of microphone techniques [5]. TIMIT sentences [6] spoken simultaneously by two talkers were recorded with a pair of

high-quality omnidirectional microphones, 18 cm apart, approximately 4 m from the talkers. Lapel microphone signals were also recorded, to provide close reference signals for the objective evaluation. In the stereo recording, one talker was 4.6 dB louder than the other, according to the relative estimated signal-to-interference ratios (SIRs) calculated by BSS Eval [21]. Therefore, for the subjective tests, the application scenario was to estimate the speech uttered by the *quieter* talker, i.e., to allow the talker at −4.6 dB SIR to be better level-balanced in post-production.

3.2 Subjective evaluation

Listening tests were conducted using a standardized "0+5+0" surround setup [8] with Genelec 8020B loudspeakers in an acoustically-treated listening room (RT60 conforming to ITU recommendation BS.1116-3 [10] above 400 Hz). In the subjective experiment (also reported by Coleman *et al.* [2]), listeners were presented with the stereo recording (left and right signals rendered directly to ±30 °) and a BSS-estimated object extracted by Mandel's method. They were asked to *"adjust the slider* [controlling the extracted object level] *until the target talker is as clear and easy to understand as possible, whilst ensuring that the overall audio quality remains at an acceptable level (compared to the reference)."* The BSS object was rendered at azimuths {0,15,30°}, with three repeats, giving nine ratings per listener. Additionally, a threshold of audibility was determined: listeners were presented with the same stimulus (object at 0°) and asked to *"adjust the [object] level to the point immediately before the mix is different to the reference."* This part also included three repeats. Ten experienced listeners completed the tests, of whom seven were native English speakers. The results are shown as boxplots for each participant (showing the range of the data, the quartiles, and medians with 95% confidence notches) in Fig. 1. It can be seen that for most participants, the thresholds of audibility and acceptability are significantly different. The results of participant 5 were removed from further analysis due to the large variance in threshold judgements. The results from the remaining participants were normally distributed, both for audibility (Lilliefors test, $p = 0.08$) and acceptability (Lilliefors test, $p > 0.50$). The mean mixing level averaged over azimuth (0.2 dB relative to the reference) differed significantly from the threshold of audibility (−14.9 dB) according to a two-sample t-test ($t = 9.73$, $p < 0.01$). There is therefore a region (15 dB range) in which the BSS-extracted object is audible and makes the target talker clearer, while maintaining acceptable quality. An analysis of variance (ANOVA) showed no significant effects of azimuth ($F = 0.85$, $p = 0.43$) or repeat ($F = 0.98$, $p = 0.38$) on the acceptability threshold.

3.3 Objective evaluation

Objective evaluation was conducted to support the listening test analysis. The objective evaluation employed two metrics: short-time objective intelligibility (STOI) [19], in the range [0,1], which predicts speech intelligibility; and perceptual evaluation of speech quality (PESQ) [15], in the range [−0.5,4.5], which

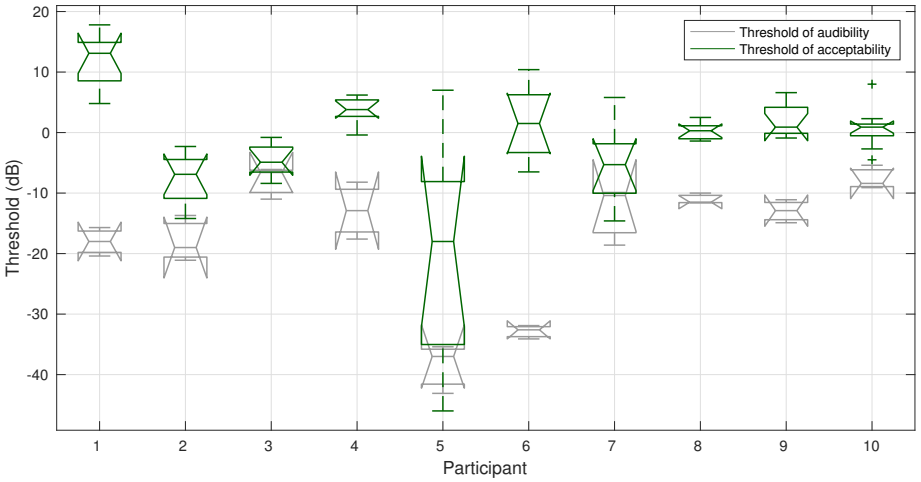

Fig. 1. Box plots of perceptual thresholds of audibility and acceptability, for remixing Talker 2 (estimated by Mandel's method). Notches show 95% confidence intervals around the median [14].

predicts speech quality. The mono sum of the stereo reference, mixed with the extracted speech object at relative levels in the range ±20 dB, was presented to the models. Prior to processing, all signals were downsampled to 16 kHz and each test mixture was loudness-matched to the reference lapel microphone signal using a Matlab implementation of [9]. Objective scores were calculated as the average of scores obtained individually for each sentence in the recording (four clips with average duration 3.2 s for Talker 2 as target; five clips with average duration 2.7 s for Talker 1 as target). Relative STOI and PESQ scores (target talker score − interfering talker score) were calculated for Mandel (corresponding to the subjective experiment described above), Alinaghi, and Liu.

The STOI scores are plotted in Fig. 2 for Talker 1 (left) and Talker 2 (right). The −0.1 relative STOI score for the target talker in the original stereo recording (relative SIR −4.6 dB) confirms that the interfering talker is more intelligible than the target talker before mixing the extracted object into the scene. By increasing the object's level in the mixture, the relative STOI scores increase. At the mean mixing level determined in the subjective tests using Mandel's method, the relative scores are both positive, implying that introducing the separated speech into the mix has resulted in an enhancement in speech intelligibility. Moreover, Mandel's method, as tested subjectively, performed worst among the three methods tested. For both talkers, Alinaghi was predicted to give the greatest improvement in speech intelligibility; Liu was ranked second while improving upon Mandel.

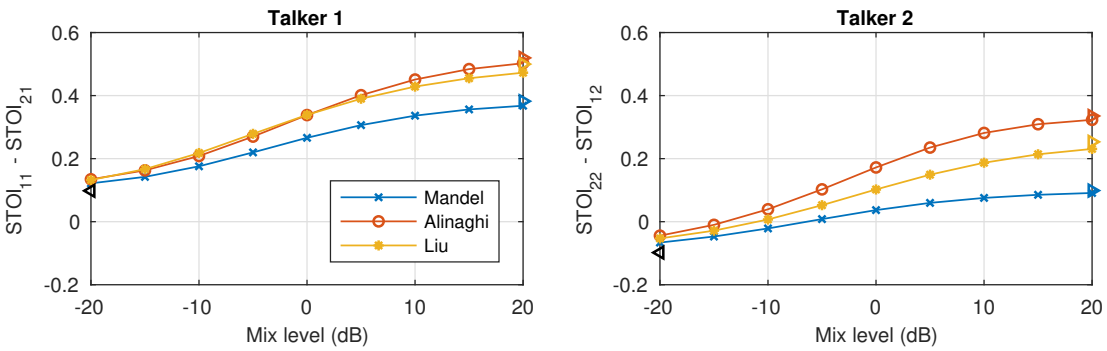

Fig. 2. Relative STOI scores, where STOI_{AB} denotes the score for target Talker A when adjusting the level of Talker B. The mixture score (◁) and object-only scores for each method (▷) are also marked.

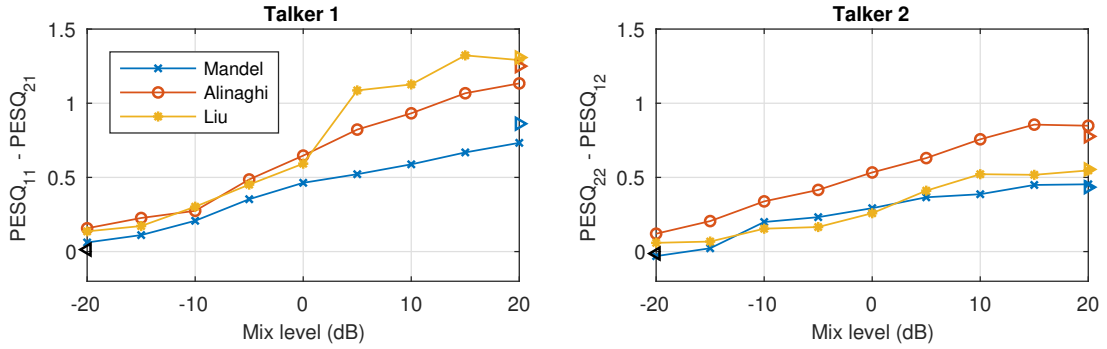

Fig. 3. Relative PESQ scores, where PESQ_{AB} denotes the score for target Talker A when adjusting the level of Talker B. The mixture score (◁) and object-only scores for each method (▷) are also marked.

The PESQ scores are plotted in Fig. 3 for Talker 1 (left) and Talker 2 (right). For all methods, and both Talkers, the relative PESQ scores increase with mixing level, implying that the separated speech is closer to the reference lapel microphone signal than the mixture. However, the subjective results indicate that the relative PESQ score does not fully convey the listening experience of the remixed speech, because the listeners identified a threshold of acceptability above which the target quality was not acceptable. For the PESQ scores, Mandel also performs worst among the methods tested. Alinaghi performs best for Talker 2, and well for Talker 1, although the relative scores for Liu are best for Talker 1 above a mix level of 0 dB. This performance is analysed further in terms of the signal-based metrics discussed below.

The objective evaluation was extended by obtaining the signal-to-interference, -artefact, and -distortion ratios (SIR, SAR, and SDR respectively) for each method, at each remix level, for both talkers. These results are plotted in Fig. 4

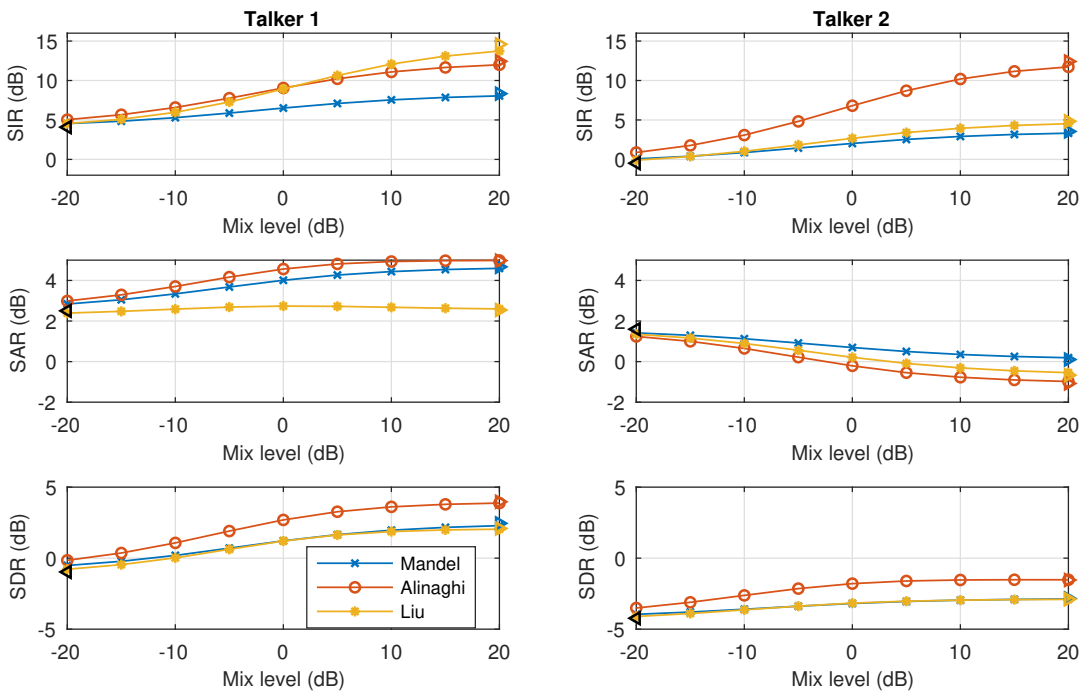

Fig. 4. Signal-based evaluations of SIR (top row), SAR (middle), and SDR (middle), adjusting the level for Talker 1 (left) and Talker 2 (right). The mixture score (◁) and object-only scores for each method (▷) are also marked.

for Talker 1 (left) and Talker 2 (right). Scores are absolute (i.e., only the target talker is taken into account). The SIR scores show that Mandel performs worst among the methods tested. For Talker 2, Liu is close to Mandel but slightly better, while Alinaghi performs over twice as well. The scores for Talker 1 are higher overall. Liu and Alinaghi give similar performance, but Liu exceeds Alinaghi for mix levels above 0 dB. These trends closely mirror the relative PESQ scores shown in Fig. 3, suggesting that SIR is the dominant signal property contributing to the relative PESQ scores.

The SAR scores have different trends for each talker. For Talker 2 (quieter in original mix), the SAR decreases with mix level, which may explain why listeners found there to a trade off between speech intelligibility and target quality. On the other hand, SAR actually increases with mix level for Talker 1 (apart from Liu, which remains approximately stable with mix level). Thus, if Mandel or Alinaghi were applied to remix Talker 1, the thresholds of acceptable quality would likely be higher than those reported in the subjective tests described above. Finally, the SDR scores for each method and talker increase with mix level, with Alinaghi outperforming Mandel and Liu in each case.

4 Conclusions

Subjective and objective results were presented to evaluate the performance of speech remixing, enabled by BSS. Such remixing has applications in object-based audio, where a producer may wish to make adjustments to a mix not facilitated by the available microphone signals, or an object-based renderer may adjust a mix based on a listener's personal preference or accessibility settings. The subjective scores showed that, in a challenging scenario with two interfering talkers, the quieter talker could be made clearer by mixing in an object estimated by BSS, while retaining acceptable audio quality. STOI, an objective perceptual model, was used to verify that the relative speech intelligibility increased with mix level. The SAR for Talker 2 for Mandel's method (corresponding to the subjective test scenario) reduced with mix level, which could explain why listeners felt that the quality degraded after the mean acceptability threshold at a mix level of 0.2 dB.

Further predictions of speech intelligibility, quality, and signal-based metrics of SIR, SAR and SDR suggested that the scenario considered for the subjective tests was the worst case among the two talkers and the three tested BSS algorithms (Mandel, Alinaghi, and Liu). In particular, the objective metrics suggested that Alinaghi may perform well compared to Mandel. Furthermore, as the DNN in Liu was trained on binaural features (including ILD), yet omnidirectional microphones were used here, the method would likely perform better if the training conditions were closer to the application example studied.

Further work should investigate whether the perceptual acceptability thresholds increase for the other methods tested. Other aspects not tested here that could be developed in future include respatialisation of BSS-estimated sources, and the applications to other sound sources, e.g. musical instruments. Finally, the possibility of creating an object-based scene with only BSS-extracted sources (i.e., no underlying channel-based recording) could be investigated. In [2, Sec. III.C], we made some informal comments about this scenario; in general the BSS-extraction allows for respatialization and some level control of the mixed sources, but degradations in the target quality due to the BSS are more exposed.

5 Acknowledgements

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1). Relevant data can be accessed via https://doi.org/10.15126/surreydata.00845514.

References

1. Alinaghi, A., Jackson, P.J., Liu, Q., Wang, W.: Joint mixing vector and binaural model based stereo source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22(9), 1434–1448 (Sept 2014)

2. Coleman, P., Franck, A., Francombe, J., Liu, Q., de Campos, T., Hughes, R., Menzies, D., Galvez, S., Tang, Y., Woodcock, J., et al.: An audio-visual system for object-based audio: from recording to listening. IEEE Transactions on Multimedia [in press], https://doi.org/10.1109/TMM.2018.2794780 (2018)
3. Emiya, V., Vincent, E., Harlander, N., Hohmann, V.: Subjective and objective quality assessment of audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 19(7), 2046–2057 (2011)
4. European Telecommunications Standards Institute: Digital audio compression (AC-4) standard part 2: immersive and personalized audio, ETSI-TS-103-190-2. European Telecommunications Standards Institute (2015)
5. Francombe, J., Brookes, T., Mason, R., Flindt, R., Coleman, P., Liu, Q., Jackson, P.: Production and reproduction of program material for a variety of spatial audio formats. In: 138 Conv. Audio Eng. Soc. Warsaw, Poland (2015)
6. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Philadelphia: Linguistic Data Consortium (1993)
7. Herre, J., Hilpert, J., Kuntz, A., Plogsties, J.: MPEG-H 3D audio — The new standard for coding of immersive spatial audio. IEEE J. Sel. Topics Signal Process. 9(5), 770–779 (2015)
8. ITU-R: Recommendation ITU-R BS.2051-0: Advanced sound system for programme reproduction. International Telecommunication Union (2014)
9. ITU-R: Recommendation BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level. International Telecommunication Union (2015)
10. ITU-R: Recommendation ITU-R BS.1116-3: Methods for the subjective assessment of small impairments in audio systems. International Telecommunication Union (2015)
11. Liu, Q., Xu, Y., Coleman, P., Jackson, P.J.B., Wang, W.: Iterative deep neural networks for speaker-independent binaural blind speech separation. In: IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP). [Accepted] (2018)
12. Liu, Q., Wang, W., Jackson, P.J., Cox, T.J.: A source separation evaluation method in object-based spatial audio. In: Signal Processing Conference (EUSIPCO), 2015 23rd European. pp. 1088–1092. IEEE (2015)
13. Mandel, M.I., Weiss, R.J., Ellis, D.P.W.: Model-based expectation-maximization source separation and localization. IEEE Transactions on Audio, Speech, and Language Processing 18(2), 382–394 (Feb 2010)
14. McGill, R., Tukey, J.W., Larsen, W.A.: Variations of box plots. The American Statistician 32(1), 12–16 (1978)
15. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In: IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP). vol. 2, pp. 749–752. IEEE, Salt Lake City, UT, USA (May 2001)
16. Roma, G., Grais, E.M., Simpson, A.J., Plumbley, M.D.: Music remixing and upmixing using source separation. In: Proceedings of the 2nd AES Workshop on Intelligent Music Production (2016)
17. Simpson, A.J., Roma, G., Grais, E.M., Mason, R.D., Hummersone, C., Plumbley, M.D.: Psychophysical evaluation of audio source separation methods. In: International Conference on Latent Variable Analysis and Signal Separation. pp. 211–221. Springer (2017)
18. Simpson, A.J., Roma, G., Plumbley, M.D.: Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In: International Con-

ference on Latent Variable Analysis and Signal Separation. pp. 429–436. Springer (2015)
19. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing 19(7), 2125–2136 (2011)
20. Torcoli, M., Herre, J., Paulus, J., Uhle, C., Fuchs, H., Hellmuth, O.: The adjustment/satisfaction test (a/st) for the subjective evaluation of dialogue enhancement. In: Audio Engineering Society Convention 143. Audio Engineering Society (2017)
21. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE transactions on audio, speech, and language processing 14(4), 1462–1469 (2006)
22. Wierstorf, H., Ward, D., Mason, R., Grais, E.M., Hummersone, C., Plumbley, M.D.: Perceptual evaluation of source separation for remixing music. In: Audio Engineering Society Convention 143 (2017)