This paper was presented at the 143th Convention of the Audio Engineering Society, as paper number 9880. The full published version can be found at http://www.aes.org/e-lib/browse.cfm?elib=19277.

Perceptual Evaluation of Source Separation for Remixing Music

Hagen Wierstorf[1], Dominic Ward[1], Russell Mason[2], Emad M. Grais[1], Chris Hummersone[2], and Mark D. Plumbley[1]

[1]*Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, U.K.*
[2]*Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, U.K.*

Correspondence should be addressed to Hagen Wierstorf (h.wierstorf@surrey.ac.uk)

ABSTRACT

Music remixing is difficult when the original multitrack recording is not available. One solution is to estimate the elements of a mixture using source separation. However, existing techniques suffer from imperfect separation and perceptible artifacts on single separated sources. To investigate their influence on a remix, five state-of-the-art source separation algorithms were used to remix six songs by increasing the level of the vocals. A listening test was conducted to assess the remixes in terms of loudness balance and sound quality. The results show that some source separation algorithms are able to increase the level of the vocals by up to 6 dB at the cost of introducing a small but perceptible degradation in sound quality.

Introduction

There is often a desire to remix existing audio content: combining elements from existing songs to create new music, or modifying the component levels, spatial positions and frequency content of a mix to optimise the listening experience over different reproduction systems. Remixing is easily possible when the original multitrack recording is readily accessible, but this is not always the case, especially for consumers of film and music. In such cases, it is possible to use source separation techniques to estimate the component sources of an existing audio mixture, thereby facilitate remixing [1]. Unfortunately, source separation often introduces perceptible artifacts, distorts the signal and suffers from imperfect separation, so it is important to evaluate the perceived quality of the resulting remix, ideally using some form of predictive metric. Most existing evaluation metrics for audio source separation are based on mathematical predictions of the artifacts, distortions, and interference introduced to the separated sources. The most prominent one is a toolbox evaluating blind source separation methods, called BSS Eval [2], which calculates metrics like the signal-to-artifacts ratio (SAR), signal-to-distortion ratio (SDR), and signal-to-interference ratio (SIR). A more recent approach is 'The Perceptual Evaluation methods for Audio Source Separation Toolkit' (PEASS) [3], which attempts to achieve better predictions by combining auditory-based metrics [4] with the signal decomposition approach of BSS Eval. Despite the establishment and widespread use of those metrics by the source separation community [5], some studies question their perceptual relevance [6, 7]. Additionally, complete separation is not always necessary for remixing because this depends on the nature of the remix. It is therefore not yet clear whether such measures can be used to predict the sound quality of mixtures generated after recombining the separated sources. As a first step towards developing a predictive model, the goal of this work is to collect ground truth data in which listeners judge the success of a remix generated using state-of-the-art source separation algorithms.

For the evaluation of remixing based on source separation there exist no well-established paradigm, but different approaches have been reported in the literature. Source separation techniques based on knowledge of the music score have been applied to change the level of single instruments [8, 9], or add audio effects such as reverb [9]. In both cases, evaluation was only performed on the separated sources and not on the result-

ing remix. Simpson et al. [10] evaluated vocal stems extracted from a music mix in terms of 'vocal similarity' and 'loudness balance' compared to the clean vocal and the original mix, but they did not introduce changes to the mix. Gillet and Richard [11] remixed music by extracting the drum track from the mixture and adjusting its level. In their experiment, listeners were asked to compare the remixes to the original mix and rate the perceived naturalness. However, the perceived difference in naturalness was caused not only by the artifacts introduced by the algorithm, but also by differences in the level of the drums introduced by the experimenters. Yoshii et al. [12] enabled listeners to change the volume or timbre for bass and snare sounds. Their evaluation showed that the audibility of the artifacts increased with level. Similarly, Pons et al. [13] demonstrated that an increase in level of the vocals in a music mix is desirable for cochlear implant users, but that a trade-off exists between achieving a high increase in level and avoiding the audibility of artifacts introduced by the algorithm. In their evaluation, listeners were able to adjust the level of the remix by themselves and in a second experiment rate the preferred remix in a paired comparison test.

The present study introduces a procedure for the perceptual evaluation of source separation methods within the context of music remixing. A listening test was performed using an adaptation of the multiple stimuli with hidden reference and anchor (MUSHRA) protocol [14], in which listeners compared a reference remix, generated from the original sources, with the remixes generated from the sources extracted by the different algorithms. Target remixes were created for a series of vocal level adjustments. In one run listeners rated the loudness balance between the vocals and the accompanying instruments, and in another they rated the sound quality of the mix. It was expected that the accuracy of achieving the target loudness balance would be largely affected by the amount of vocal separation achieved by the source separation algorithm. In contrast, the perceived sound quality was expected to be primarily influenced by the audibility of the time-varying distortions and artifacts introduced by the source separation algorithms.

Methodology

This section details the design of the listening experiment used to assess the performance of the source separation algorithms when used for remixing music. All source code used to generate the experiment and the stimuli are available for download from zenodo [15, 16].

Stimuli

The stimuli used for the experiment were derived from the Demixing Secret Database (DSD100), a set of 100 songs each comprising four stereo sources (bass, drums, vocals, and other) that sum to realistic mixtures [5]. Song genres range from hip-hop to heavy metal, though the majority of songs fall into rock and pop classes. The DSD100 was developed for the MUS task of the 2016 Signal Separation Evaluation Campaign (SiSEC)[1] [5], where 23 different source separation algorithms were evaluated using the BSS Eval performance measures. One half of the dataset was used for model training (the development set), and the other for evaluation (the test set). We did not apply the source separation algorithms ourself for this study, but used the SiSEC MUS submission data as kindly provided by Fabian-Robert Ströter. All stimuli were converted to monaural signals by averaging the two input channels.

Selection of songs and algorithms

From the 23 source separation algorithms that participated in the SiSEC MUS task, we selected the 12 that were able to extract all four instrument groups to allow for later investigation of changes to other instruments in each mix. The algorithms were sorted based on their average SDR for the vocal stem across all songs using the BSS Eval measures from SiSEC. From this list five different algorithms were picked in order to include a wide range of different SDR values as presented in Table 1.

Table 1: Average SDR values for vocals as target of the selected source separation algorithms as calculated by BSS Eval as part of SiSEC [5].

Algorithm	Reference	SDR
UHL3	Uhlich et al. [17]	5.3 dB
NUG3	Nugraha et al. [18]	4.1 dB
OZE	Ozerov et al. [19]	1.3 dB
GRA3	Grais et al. [20]	−2.2 dB
KON	Kong [based on 21]	−4.3 dB

[1] https://www.sisec17.audiolabs-erlangen.de

UHL3 uses a spatial covariance matrix, a blending of deep neural networks and bidirectional long-short term memory neural networks together with data augmentation [17]. NUG3 applies a spatial covariance matrix and deep neural networks in an iterative expectation-maximisation fashion [18]. OZE is a non-negative matrix factorization-based algorithm applying an iterative expectation-maximisation approach [19]. GRA3 concatenates the two input channels and uses a deep neural network to predict a soft mask [20]. KON averages the two input channels and applies a recurrent neural network [based on 21].

Song numbers 57 and 35 of DSD100 were used for the familiarisation and training stages, respectively, of the experiment. For the main experiment, song numbers 6 (pop/rock), 17 (electronic), 30 (jazz), 31 (pop/rock), 42 (pop/rock) and 48 (heavy metal) were chosen by comparing the between-algorithm distributions of the SDR, SIR and SAR BSS Eval measures across the 46 songs from the test set of DSD100. For each BSS Eval metric, the 23 distributions with the largest interquartile ranges were retained, from which two songs were selected based on the maximum and minimum of the medians. This procedure resulted in two songs for each SDR, SIR and SAR statistic, giving a total of six songs.

Remixes

For each of the six selected songs, we generated three reference mixes by adjusting the level of the vocals, relative to the level as set by the mixing engineer, by 0 dB, 6 dB or 12 dB before summing all four sources. This procedure was repeated for the sources estimated by the five selected source separation algorithms. In addition, we created four different anchor stimuli from those stems. Three were generated by changing the level of the vocals by -14 dB, -8 dB or -2 dB, i.e. 14 dB down from the level offsets used to create the reference mixes. We call these the 'loudness balance anchors'. The fourth anchor was generated by removing 20% of the time frames from the spectrogram of the reference mixture and lowpass filtering it with a hard cutoff frequency of 3.5 kHz. Musical noise was then created by randomly removing 99% of the time-frequency bins from a second (unfiltered) spectrogram before applying the same lowpass filter. The inverse of these two spectrogram were then summed at equal loudness according the ITU-R BS.1770 loudness algorithm [22]. We call this anchor the 'sound quality anchor', which is similar to the one used by [6], which itself is based on the anchors defined in [3].

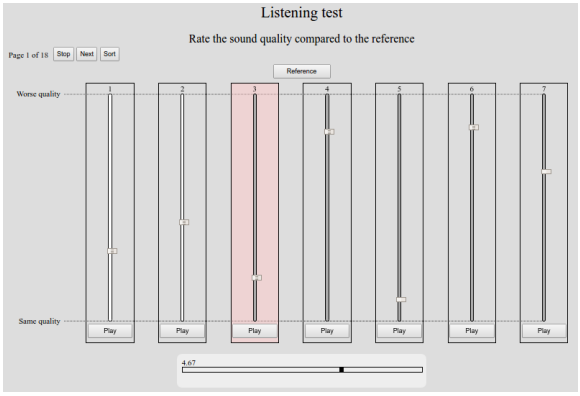

Fig. 1: MUHSRA interface for the sound quality task.

Procedure

Two MUSHRA listening assessments were conducted. In the first, participants were asked to judge the perceived sound quality of the test stimuli (approximated remixes) compared to a reference stimulus (target remix) on a scale ranging from 'same quality' to 'worse quality' with no intermediated labels shown in between, see Figure 1. The MUSHRA interface was a visually slightly modified version from the 'Web Audio Evaluation Tool' [23]. The second assessment required participants to judge the similarity of the loudness balance of the test stimuli compared to a reference on a scale ranging from 'same balance' to 'different balance'. The presentation order of these two assessments was balanced across the listeners.

In both assessments, seven test stimuli were presented on every page: the five remixes generated by the source separation algorithms under test, the appropriate anchor stimulus and the hidden reference. For the sound-quality assessment the same anchor was used for all three remixes of a given song. For the loudness-balance assessment, the anchor depended on the level offset applied to the vocals (see Section 2.1.2). Each test started with a written introduction to the test, followed by the same familiarisation page, where the following definitions of sound quality and loudness balance were presented:

"*Sound quality* relates to the amount of artifacts or distortions affecting the perception of a reference sound.

These can be heard as tone-like additions, abrupt changes in loudness, or missing parts of the audio. It does not include changes in loudness balance, e.g. the loudness of the vocals relative to the loudness of the other instruments."

"*Loudness balance* describes the relation of the overall loudness of the vocals to the overall loudness of the remaining instruments. It does not include short and abrupt changes in loudness that you might experience for some test sounds. It is more considered with the general balance of the vocals and the accompanying instruments."

These definitions were accompanied by example stimuli in which only the loudness balance or the sound quality was altered. Each assessment lasted approximately 50 minutes.

Participants

15 participants, including four of the authors, completed the experiment. The majority were research students from the Centre for Vision, Speech and Signal Processing (CVSSP), and had some previous experience with listening tests. Informed written consent was obtained from each participant, and they received a financial compensation.

Apparatus

The listening test took place in an acoustically isolated room at CVSSP. Listeners sat in front of a flat screen placed on a small table and used a computer mouse to complete the assessment. In a separate room, a computer equipped with an RME Hammerfall DSP MADI soundcard was used to deliver the stimuli digitally to the listening room where the signal was converted to analogue (RME MADIface XT) for reproduction over Sennheiser HD600 headphones.

Results

For the evaluation, the medians over all listeners were calculated for every remix. Figure 2 shows the medians, with 95% confidence intervals, by algorithm for the quality and loudness balance assessments. The reference and anchor remixes have been labelled 'Ref' and 'Anchor', respectively. For the analysis the two rating scales were assigned values of 0 at the lower end and 1 at the upper end with continuous steps in between.

For most songs, the sound quality and loudness balance rating depends on the relative level of the vocals. Increasing the level of the vocals tends to lead to ratings of worse quality and less similar loudness balance. This is more pronounced for the highest adjustment of 12 dB.

The perceived impairments have a dependency on the selected song as well as the algorithm. The best rated song in terms of sound quality and loudness balance is song 30. For the 12 dB conditions it has an average rating across the algorithms of 0.81 for sound quality and 0.74 for loudness balance. Song 48 received the lowest rating, with 0.5 for sound quality and 0.31 for loudness balance.

All of the participants rated the sound quality anchor to have consistently worse quality than the reference, resulting in an average rating across all song medians of 0 for all level settings. The loudness balance anchors were less consistently rated at the bottom end of the scale, resulting in average medians of 0 for a mixing level of 0 dB, 0.06 for 6 dB, and 0.1 for 12 dB.

To investigate the performance of the single algorithms, the average of the medians across songs was calculated and is presented in Figure 3 in direct dependency of the mixing levels. The error bars in this figure are now the standard deviation, indicating the variations for the single conditions across songs.

For a mixing level of 0 dB almost all of the algorithms achieved an average sound quality and loudness balance rating of above 0.98, only KON achieved only 0.75 for sound quality and 0.89 for loudness balance. For a mixing level of 6 dB UHL3, OZE, and GRA3 achieved an average sound quality rating of 0.95 or better, whereas KON was again on the lower end with a rating of 0.53. For the loudness balance ratings, UHL3, NUG3, and GRA3 achieved ratings above 0.9, with KON at the lower end with a rating of 0.7. In the 12 dB conditions across all songs UHL3 received the highest ratings with 0.83 for sound quality and 0.77 for loudness balance. The lowest rating for sound quality achieved KON with an average of 0.34. The lowest rating for loudness balance was achieved by OZE with an average of 0.46.

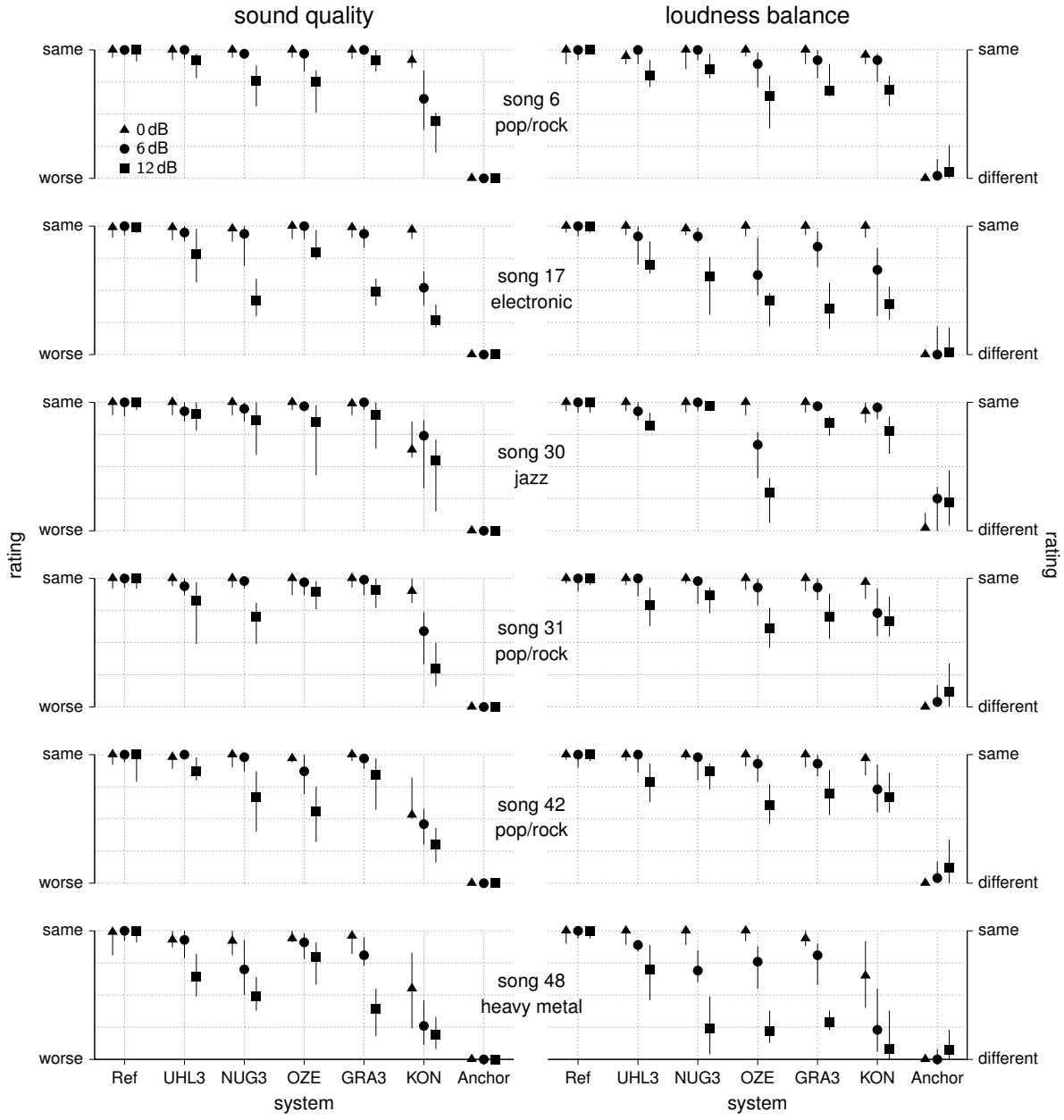

Fig. 2: Median results and the corresponding 95% confidence interval for sound quality and loudness balance ratings across listeners for six different songs. The ratings are shown for the three different level settings of 0 dB, 6 dB, 12 dB indicated by the different symbols. The conditions always start with the reference mixes labelled 'Ref', the five algorithms under test (compare Table 1) and end with the corresponding anchors. The grid is displayed only for visual enhancement, as the listening test facilitated a continuous scale. Data and code to generate this figure are available at [24].

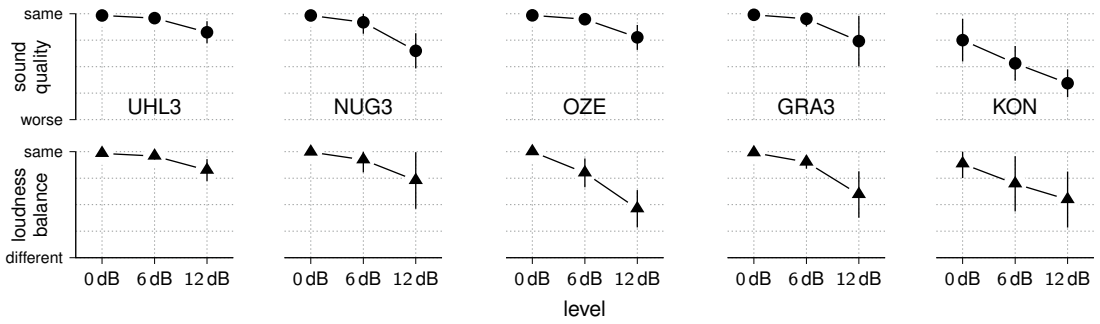

Fig. 3: Sound quality and loudness balance ratings for the five different source separation algorithms dependent on the mixing levels. The average across the medians of every song is presented together with the standard deviation, indicating the variation of the results with different songs. Data and code to generate this figure are available at [24].

Discussion

The results show that source separation algorithms can be successfully applied to the task of increasing the level of the vocals in a music mixture under certain conditions. For example, UHL3, NUG3 and GRA3 were able to achieve the desired change in relative loudness between the vocals and accompaniment for the 0 dB and 6 dB with little effect on sound quality in all six songs. However, when increasing the levels to 12 dB, all five algorithms introduced perceptible artifacts and struggled to achieve the desired loudness balance. We may conclude, therefore, that the perceptual impact of the artifacts, distortions and interference introduced by source separation algorithms when remixing music is dependent on the amount of processing involved, in addition to the spectro-temporal relationships between the instruments that affect masking.

The ratings for both sound quality and loudness balance show a pronounced dependency on the song. The genres of the two lowest rated songs were heavy metal and electronic and differed from the other songs that were from the genres pop/rock and jazz. This might reflect that the first two genres are more challenging for source separation algorithms, either in terms of the instrumentation or in terms of the relative levels of the components in the mix.

To summarise the effect of level on the average sound quality and loudness balance ratings for each algorithm, the difference between the average ratings for 0 dB and 12 dB was calculated. The algorithms were then ordered, for each assessment, from the lowest to the highest difference, where lower differences reflect a smaller effect of level. The rank order for the effect of level on sound quality was UHL3, OZE, GRA3, NUG3, KON, and for loudness balance: UHL3, NUG3, KON, GRA3, OZE. This highlights that UHL3 achieves the most consistent ratings for all mixing levels. OZE shows a similar consistency for sound quality, but is the least consistent for loudness balance. This indicates that it might suffer from severe interference, and therefore is less suited for remixing.

The experiment was able to show different performances for different songs, different algorithms, and the two rating tasks. This indicates that the introduced method with the specific definition of sound quality and loudness balance as presented in Section 2.2 seems to be suited to investigate the performance of source separation algorithms for remixing tasks.

What might be improved is the loudness balance anchor, which was not rated as consistently as the sound quality anchor. In addition, it might be of interest to investigate if the perceptual difference between the reference mix and the loudness balance anchor was the same for all three mixing levels as was intended by the usage of a constant offset in vocal level. One might also introduce the 0 dB reference mix as an additional anchor for all mixing levels. In addition to the anchors with changing vocal levels, this could serve as a useful baseline for assessing the remix performance of the algorithms.

There exists one difference between sound quality and loudness balance that might not be well represented

by the rating procedure. It might be that a slight deviation in loudness balance is easily acceptable as it could be corrected for by increasing the mixing level further, whereas a deviation in sound quality might not be acceptable at all for some applications. This could be investigated by using a similar method as Pons et al. [13] where the listeners could adjust the mixing level to a point where the amount of artifacts would be still considered to be acceptable.

Conclusion

This paper introduced a modified MUSHRA method to access source separation algorithms for remixing music by increasing the level of the vocal. The assessment method was able to differentiate between changes in sound quality due to artifacts and distortions introduced by the algorithms and changes of the desired loudness balance caused by weak separation.

From the five different algorithms under test, an algorithm based on deep neural networks, data augmentation and network blending (UHL3) achieved the best ratings across songs and mixing levels. In nearly all cases, increasing the vocal level by 12 dB degraded sound quality, indicating that the amount of remixing should be constrained.

Acknowledgements

This research has been supported by EPSRC grant Musical Audio Repurposing using Source Separation, EP/L027119/2.

References

[1] Roma, G., Grais, E., Simpson, A., and Plumbley, M., "Music Remixing and Upmixing using Source Separation," in *2nd AES Workshop on Intelligent Music Production*, 2016.

[2] Vincent, E., Gribonval, R., and Fevotte, C., "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4), pp. 1462–1469, 2006, doi:10.1109/TSA.2005.858005.

[3] Emiya, V., Vincent, E., Harlander, N., and Hohmann, V., "Subjective and Objective Quality Assessment of Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), pp. 2046–2057, 2011, doi:10.1109/TASL.2011.2109381.

[4] Huber, R. and Kollmeier, B., "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6), pp. 1902–1911, 2006, doi:10.1109/TASL.2006.883259.

[5] Liutkus, A., Stöter, F.-R., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., and Fontecave, J., "The 2016 Signal Separation Evaluation Campaign," in P. Tichavský, M. Babaie-Zadeh, O. J. J. Michel, and N. Thirion-Moreau, editors, *13th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 323–332, Springer, 2017, doi:10.1007/978-3-319-53547-0_31.

[6] Cano, E., FitzGerald, D., and Brandenburg, K., "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *24th European Signal Processing Conference (EUSIPCO)*, pp. 1758–1762, 2016, doi:10.1109/EUSIPCO.2016.7760550.

[7] Gupta, U., Moore, E., and Lerch, A., "On the perceptual relevance of objective source separation measures for singing voice separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2015, doi:10.1109/WASPAA.2015.7336923.

[8] Itoyama, K., Goto, M., Komatani, K., Ogata, T., and Okuno, H. G., "Instrument equalizer for query-by-example retrieval: improving sound source separation based on integrated harmonic and inharmonic models," in *9th International Conference on Music Information Retrieval (ISMIR)*, pp. 133–138, 2008.

[9] Woodruff, J., Pardo, B., and Dannenberg, R., "Remixing Stereo Music with Score-Informed Source Separation," in *7th International Conference on Music Information Retrieval (ISMIR)*, pp. 314–319, 2006.

[10] Simpson, A. J. R., Roma, G., Grais, E. M., Mason, R. D., Hummersone, C., and Plumbley, M. D., "Psychophysical Evaluation of Audio Source Separation Methods," in P. Tichavský, M. Babaie-Zadeh, O. J. J. Michel, and N. Thirion-Moreau, editors, *13th International Conference*

on Latent Variable Analysis and Signal Separation (LVA/ICA), pp. 211–221, Springer, 2017, doi:10.1007/978-3-319-53547-0_21.

[11] Gillet, O. and Richard, G., "Extraction and remixing of drum tracks from polyphonic music signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 315–318, 2005, doi:10.1109/ASPAA.2005.1540232.

[12] Yoshii, K., Goto, M., and Okuno, H. G., "INTER:D: a drum sound equalizer for controlling volume and timbre of drums," in *The 2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT)*, pp. 205–212, 2005, doi:10.1049/ic.2005.0733.

[13] Pons, J., Janer, J., Rode, T., and Nogueira, W., "Remixing music using source separation algorithms to improve the musical experience of cochlear implant users," *The Journal of the Acoustical Society of America*, 140(6), pp. 4338–4349, 2016, doi:10.1121/1.4971424.

[14] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," Technical Report 10, International Telecommunication Union, 2015.

[15] Wierstorf, H. and Ward, D., "Stimuli for the paper Perceptual Evaluation of Source Separation for Remixing Music," 2017, doi:10.5281/zenodo.835182.

[16] Wierstorf, H. and Ward, D., "Experimental procedure for the paper Perceptual Evaluation of Source Separation for Remixing Music," 2017, doi:10.5281/zenodo.835191.

[17] Uhlich, S., Porcu, M., Giron, F., Enenkl, M., Kemp, T., Takahashi, N., and Mitsufuji, Y., "Improving music source separation based on deep neural networks through data augmentation and network blending," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 261–265, 2017, doi:10.1109/ICASSP.2017.7952158.

[18] Nugraha, A. A., Liutkus, A., and Vincent, E., "Multichannel Audio Source Separation With Deep Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9), pp. 1652–1664, 2016, doi:10.1109/TASLP.2016.2580946.

[19] Ozerov, A., Vincent, E., and Bimbot, F., "A General Flexible Framework for the Handling of Prior Information in Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4), pp. 1118–1133, 2012, doi:10.1109/TASL.2011.2172425.

[20] Grais, E. M., Roma, G., Simpson, A. J. R., and Plumbley, M. D., "Single-Channel Audio Source Separation Using Deep Neural Network Ensembles," in *140th Convention of the Audio Engineering Society*, p. Paper 9494, 2016.

[21] Huang, P. S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P., "Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12), pp. 2136–2147, 2015, doi:10.1109/TASLP.2015.2468583.

[22] ITU-R BS.1770, "Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level," Technical Report 4, International Telecommunication Union, 2015.

[23] Jillings, N., Moffat, D., Man, B. D., and Reiss, J. D., "Web Audio Evaluation Tool: A browser-based listening test environment," in *12th Sound and Music Computing Conference*, 2015.

[24] Wierstorf, H. and Ward, D., "Figures and data for the paper Perceptual Evaluation of Source Separation for Remixing Music," 2017, doi:10.5281/zenodo.835196.