Psychophysical Evaluation of Audio Source Separation Methods

Andrew J.R. Simpson [#1], Gerard Roma [#2], Emad M. Grais [#3], Russell D. Mason [#4], Christopher Hummersone [#5], Mark D. Plumbley [#6]

[#1-3,6] Centre for Vision, Speech and Signal Processing / [#4,5] Institute of Sound Recording
University of Surrey, Guildford, UK

[1]andrew.simpson@surrey.ac.uk, [6]m.plumbley@surrey.ac.uk

Abstract. Source separation evaluation is typically a *top-down* process, starting with perceptual measures which capture fitness-for-purpose and followed by attempts to find physical (objective) measures that are predictive of the perceptual measures. In this paper, we take a contrasting *bottom-up* approach. We begin with the physical measures provided by the Blind Source Separation Evaluation Toolkit (BSS Eval) and we then look for corresponding perceptual correlates. This approach is known as *psychophysics* and has the distinct advantage of leading to interpretable, *psychophysical* models. We obtained perceptual similarity judgments from listeners in two experiments featuring vocal sources within musical mixtures. In the first experiment, listeners compared the overall quality of vocal signals estimated from musical mixtures using a range of competing source separation methods. In a loudness experiment, listeners compared the loudness balance of the competing musical accompaniment and vocal. Our preliminary results provide provisional validation of the psychophysical approach.
Keywords: Deep learning, source separation, perceptual evaluation.

1 Introduction

Audio source separation methods typically attempt to recover or estimate signals, known as 'sources', that have been mixed. The success of this "unmixing" process is evaluated both objectively, using BSS Eval [1], and subjectively by asking listeners to rate the perceived quality using methods such as MUSHRA [2,3]. Unfortunately, observed correlation between the BSS Eval measures and the subjective evaluations has proved sufficiently poor [4-9] that the community has rejected the BSS Eval measures as invalid.

It is difficult to relate physical measures (such as BSS Eval) to subjective evaluations of audio quality, as the latter is affected by a wide range of perceptual attributes, dependencies on suitability for purpose, as well as individual opinion and preference. This approach could be considered as being *top-down*, where we begin with perceptual (subjective) ratings that we find are important, and then seek physical (objective) measures which are correlated. The downside of the top-down approach is that we may not find such physical measures because we do not fully understand the process of audition.

An alternative method would be to approach the problem *bottom-up*, where we instead begin with physical measures which are descriptive of the audio signals and we look for perceptual measures which correlate to the physical measures. This approach is known as *psychophysics* and has the advantage of being able to produce interpretable, psychophysical models relating the known physical measures to corresponding perceptual measures. This paper examines whether bottom-up psychophysical evaluation principles can be applied in source separation research to obtain perceptual measures with better correlation to the objective measures.

A. *Psychophysics*

The main principle of psychophysics is that there exist psychological correlates of physical parameters [10,11]. A psychological correlate is a measurable behaviour, relating to body or mind, that is a (usually monotonic) function of some controllable, physical parameter of the stimulus. These psychophysical correlates are typically identified or substantiated using perceptual data. Perhaps the most well established example of a psychophysical correlate is loudness [12]. Loudness theory defines the loudness of a given sound as the auditory perceptual correlate of acoustic intensity. Hence, for a given signal, variations in intensity correspond to perceptual variations in loudness. Subjective judgments of loudness are elicited from listeners for stimuli with varying degrees of intensity. Although acoustic intensity may be considered as an absolute physical measurement (e.g., sound pressure level), psychophysical paradigms for the study of loudness typically involve relative judgments relating the loudness of one sound to that of another sound [13]. The most common methods for studying loudness involve comparisons of pairs of signals. The most direct method for loudness comparison is known as magnitude ratio estimation [13], where listeners provide a numerical ratio estimate that captures the loudness ratio between a given pair of stimuli.

B. *Application of Psychophysics to Source Separation*

A psychophysical approach could be applied to source separation. When undertaking objective measurements of an arbitrary mixture of known signals, the separated signal from a given separation algorithm (the signal estimate) may be compared to the known separate signals for evaluation. Any difference between the signal estimate and the corresponding known signal is often described as ‘distortion’ [1]. By subtracting the known signal from the signal estimate, a difference (i.e. distortion) signal may be obtained. The ratio of the known signal energy to the energy of this difference signal is known as the signal-to-distortion ratio (SDR) [1]. The difference signal is decomposed into two parts: an interference signal, considered to be due to the influence of other sources on the target source estimate, and an artefacts signal, considered to be that part of the estimate that is not due to either the target signal or any of the other signals. The ratio of the known target signal energy to an estimate of the energy remaining from the interfering signals is known as the source-to-interference ratio (SIR), and the ratio of the known target signal to the artefacts signal energy is known as the source-to-artefacts ratio (SAR). If SAR and SIR are employed as the physical measures, a magnitude ratio estimation methodology of subjective attributes related to

these factors could result in a close correlation between the subjective and objective metrics. As for the loudness example above, the magnitude ratio estimation methodology would involve comparative judgements, in this case judgements of the similarity of certain aspects of the signals.

The subjective attribute selected to compare to SAR was the similarity of the vocal; this is a judgement of the effect of artefacts on the perceived similarity of the target signal. This can be analysed such that judgements that are similar to the known target signal are positioned at one end of the scale, and judgements that are dissimilar to the known target signal are positioned towards the opposite end of the scale. The subjective attribute selected to compare to SIR was the loudness-balance-similarity; this is a judgement of the similarity of the perceived loudness balance between the target and interferer signals. This can be analysed such that judgements that are similar to the known target signal are positioned at one end of the scale, and judgements that are dissimilar to the known target signal, or similar to the unseparated mixture signal, are positioned at the towards the opposite end of the scale.

C. *Overview*

In this paper, we introduce a psychophysical evaluation method based on magnitude ratio estimation corresponding as closely as possible to SAR and SIR. In the following section we describe two preliminary listening tests featuring real-world audio examples obtained from a range of state-of-the-art audio source separation methods (see [14]). Next, we describe the resulting perceptual data and correlations with the physical measures to evaluate the match between the perceptual and physical data. Then, we analyse the data from the point of view of model comparison in order to evaluate whether the subjective judgements produce meaningful results. Finally, we provide some brief discussion.

2 Method

The prevailing MUSHRA perceptual evaluation methods [2-9] are focused on obtaining interpretable perceptual measures and, hence, any attempt at modelling is a secondary consideration. In contrast, the sole aim of our psychophysical study is to establish perceptual correlates of the physical parameters. Focussing on separation of vocals from musical mixtures, we conducted two listening tests using stimuli generated using competing methods for source separation. Listeners were asked to locate each of the respective versions of each given vocal signal on a perceptual line such that the placement on the line captured the perceptual similarity relationships between the respective sounds. In the first experiment, as a perceptual correlate of the physical measure SAR, we evaluated *similarity* of the vocal. In the second experiment, to capture the loudness balance (between the vocal and accompaniments) as a perceptual correlate of the physical measure SIR, we evaluated *loudness-balance-similarity*. Critically, in both cases the stimuli placed on each perceptual line included the mixture and original (pre-mixture) vocal signal.

We consider five competing source separation methods [14] featuring deep neural networks (DNN). One of the five methods comprises a baseline DNN model (*M _baseline*) and the remaining four methods comprise multi-stage architectures which

extend the baseline DNN model with various parameterisations (*M_DNN*_1-3) and/or non-negative matrix factorisation (*M_NMF*) – see [14]. Critically, the DNN architectures in [14] are designed as augmentations of the baseline DNN architecture which attempt to improve the method.

In an experiment where multiple stimuli were compared directly, six listeners were asked to provide judgements on the perceptual similarity between 30-second musical excerpts. The excerpts, chosen as representative of typical musical mixtures featuring typical vocal and accompaniment, were taken from the mixes of 10 songs selected from the *test set* of the 'MUS' task of the SiSEC challenge [15]. The mixture for each song was a summation of the available stems, where the stems comprised vocals, bass, drums and other (accompaniment). Listeners reported normal hearing and were naïve to the purpose of the test. Most listeners had some prior experience of listening tests and all were familiar with music listening, audio production and/or recording technologies.

Stimuli. We consider the separation of the vocal (stem) signal from the accompaniment signal within a mixture. All mixtures were collapsed (summed) to mono (single channel) and the various, competing source separation methods were each independently applied to each of the 10 mixtures. The voice separation output of each of the five source separation methods was used. In addition, the mixture and original vocal stem signals were also used, providing 7 alternate stimuli for each song. From the point of view of the source separation methods employed, we note that the mixtures separated were not used in training the deep neural networks (i.e., we are concerned with evaluation of the models on the *test set*). Closed (isolating) headphones were used to present the stimuli. Presentation was monaural and diotic (same in both ears). Listeners were instructed to set the volume control on the amplifier for a comfortable listening level at the beginning of the test and did not adjust it further during the test. Listeners were unpaid volunteers.

Procedure. In the case of the first experiment, similarity judgements about the vocal sources were solicited. Listeners were instructed to compare only the vocal component of the mixture in this experiment. Even in the case of the full mixture, this means that the listeners had to isolate their perception of the vocal component for comparison. The listeners declared that they were able to do this to their satisfaction. In the second experiment, similarity judgements were solicited comparing the loudness balance (see [16]) between the vocal and accompaniment in each presented stimulus.

Listeners were presented with an interface featuring seven play buttons and seven respective sliders (on a computer screen). Each play button and slider represented either one of the five voice separation outputs from the respective models or the mixture or the original vocal source. Using each individual 'play' button, listeners were able to listen to each of the respective alternate versions of the vocal at will and could repeat an unlimited number of times. Listeners arranged the vertical placement of the seven sliders to capture the similarity relationships between the various stimuli, such that sliders for very similar stimuli were placed closely (on the vertical axis) and sliders for dissimilar stimuli were placed with greater distance. Note that the absolute placements of the sliders is not informative. Listeners were briefed to maximise their use of the scale for each song but were not briefed to attempt to make consistent judgements across songs.

Listeners evaluated the stimuli in sessions of 10 songs. There was no explicit time constraint on the sessions. Most sessions were completed in under 25 minutes. The presentation order (both the order of the songs and the order of stimuli for each song) was randomised so that each slider corresponded to a different stimulus each time. Sliders were reset before each new song. When the listener had completed the arrangement of the sliders for a given song a 'next' button was pushed for the next song.

Analysis I: Correlation. Listeners did not make comparisons between songs but only within each song, hence we must first test correlations within the context of each song, before summarising the correlation over the 10 songs. For the perceptual data resulting from each listening test, the slider placements corresponding to the original vocal were subtracted (by way of reference) from the other placements on a song-by-song basis. The slider placement data for the original vocal were subsequently discarded. For each of the stimuli used in the experiments (except the original vocal signals which are not suitable for analysis), the corresponding physical measures (SAR/SIR) were computed using the toolbox associated with [1].

For the data of each listening test, a separate correlation analysis was conducted on a song-by-song basis. The medians of the subjective data were calculated for each song and each stimulus (therefore averaging across the results of the listeners). This resulted in 6 perceptual measures per song, per listening test. Next, the corresponding SAR and SIR values for each song (6 measures per song) were used to compute linear (Pearson) correlation coefficients with the respective perceptual data for that song. The correlation was calculated between the data from the first (vocal similarity) experiment and SAR, and the data from the second (loudness balance similarity) experiment and SIR. This provided, for each listening test, a set of 10 (song-wise) correlation coefficients. To summarise (over songs) each of the two respective distributions we take the median correlation coefficient.

Next, in order to provide a measure of statistical significance for the respective median correlation coefficients, permutation tests [17] were conducted. The above procedure for computing the song-wise distribution of correlation coefficients was repeated 10,000 times. Each of these 10,000 times, prior to the correlation computation, the order of the data were randomly shuffled. The median of this distribution was then taken and, over the 10,000 replications, an empirical null distribution (of null across-song medians) was accumulated. Finally, the number of median correlation coefficients that was greater than or equal to (and with the same sign as) the actual median correlation coefficient was counted and the resulting count divided by 10,000. This provides an empirical estimate of the probability of the respective across-song median correlation coefficient occurring by chance (a *P* value).

Analysis II: Model comparison. In contrast to the correlation analysis described above, in this analysis we are interested in overall performance comparisons of the models in terms of perception. Across-song means were computed for each listener and collated. The resulting data were analysed using non-parametric statistical methods (see [18]). Initially, a main-effects analysis was conducted for the data of each test using a Friedman test. Next, post-hoc analyses were conducted on a pair-wise basis in order to determine which pairs of models showed evidence of being significantly different. The post-hoc pairwise analyses can be considered 'planned tests' and so contrasts were limited (in advance) to comparisons between the baseline model and the respective, competing multi-stage models. We do not provide correction for mul-

tiple comparisons, primarily because of the limited number of listeners involved and because of the minimal number of planned contrasts.

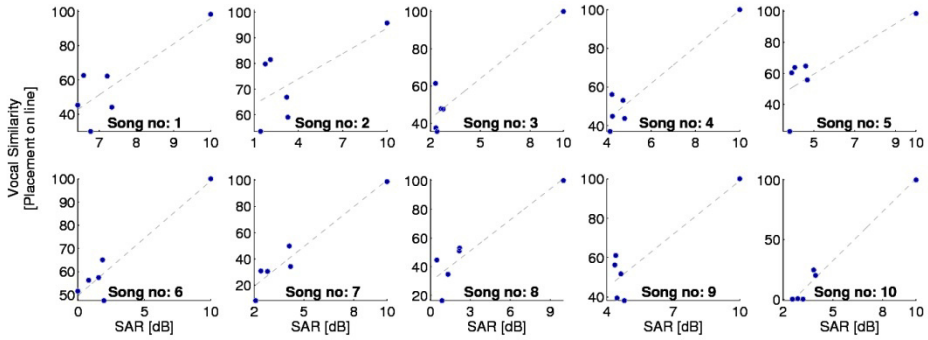

Fig. 1. Perceptual similarity versus SAR. Listeners organized sliders representing the seven respective stimuli along perceptual lines which depict the perceived similarity of the vocal component. These scatter plots show, on a song-by-song basis, across-listener median perceptual slider placement as a function of SAR. Dashed grey lines indicate linear regression lines shown for illustration only. Note: axis scale and range vary.

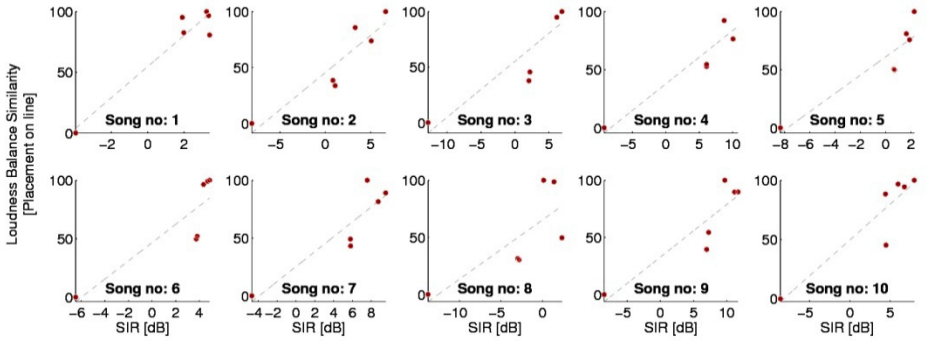

Fig. 2. Loudness balance similarity versus SIR. Listeners organized sliders representing the seven respective stimuli along perceptual lines which depict the perceived similarity of the loudness balance between the vocal and the accompaniment. These scatter plots show, on a song-by-song basis, across-listener median perceptual slider placement as a function of SIR. Dashed grey lines indicate linear regression lines shown for illustration only. Note: axis scale and range vary.

3 Results: Analysis I – *Psychophysical Correlation*

In the correlation analysis, we are not concerned with the question of which model is best, but rather we are concerned with the question of whether the physical measures correlate with the perceptual measures. For the perceptual data of the first listening test (*similarity*), Fig. 1 plots, on a song-by-song basis, the across-listener medians as a function of the respective SAR measures. Note that, for illustrative purposes only, in these plots we limited the upper SAR (for the original mixtures) to 10 dB (because these numbers would otherwise be at the limits of precision). Fig. 2 plots the equiva-

lent for the perceptual data of the second (*loudness balance similarity*) listening test. Linear regression lines are shown in grey for illustration. Qualitatively, the scatter plots of Fig. 1 show some evidence of monotonic trends but are somewhat noisy and appear to be dominated by the extremes of slider placement. The scatter plots of Fig. 2 show more obvious monotonic trends.

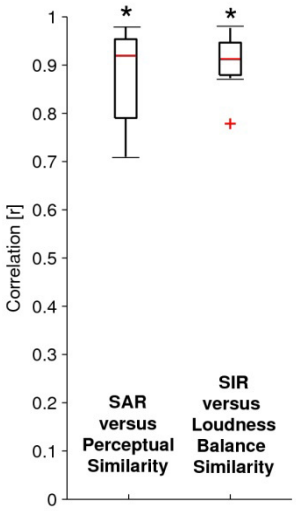

Fig. 3. Song-wise correlation coefficients. For each listening test, linear (Pearson) correlation coefficients were computed, on a song-by-song basis, for the listener-wise medians of the perceptual data with the respective BSS Eval measures. The above box-plots show median (in red), inter-quartile range (box) and 1.5x IQR respectively (whiskers). Outliers are given as red crosses. Asterisks (above box-plots) denote significant median correlation coefficients (P<0.01, *Permutation test*, n=10,000).

Figure 3 shows box-plots capturing the respective correlation coefficient distributions (each over the 10 songs) relating to the plots of Figs. 1 and 2. The median across-song correlations are both around 0.91 and both are significant (P<0.01, *Permutation test*, n=10,000). In other words, our measures correlate well, on a song-by-song basis, with the physical measures, suggesting that our psychophysical paradigm is reliable.

4 Results: Analysis II – Comparison of Separation Methods

Figure 4a shows box-plots of the data resulting from the first experiment (vocal similarity). The original vocal and the vocal-within-mixture are deemed to be very similar within the context of the experiment. There is a significant main effect among the different models (P < 0.05, Friedman Test, χ^2 = 10.93, df = 4). In the pairwise post-hoc analysis, we find no evidence that the results for multi-stage model M_DNN1 or the M_NMF model are significantly different from the baseline (P > 0.05, paired Wilcox tests, two tailed). However, multi-stage models M_DNN2-3 are significantly less similar to the original vocal than the baseline (P < 0.05, *paired Wilcox tests, two tailed*).

Figure 4b shows the respective box-plots of the data resulting from the second (loudness balance) experiment. In this case, the mixture and original voice are not located together (by the listeners) but are located at opposite ends of the perceptual space. This indicates that the listeners took these two as bounding the space; the mix-

ture providing the minimum ratio of vocal-to-accompaniment loudness and the original vocal providing the maximum ratio of vocal-to-accompaniment loudness (i.e., the ratio was theoretically infinite). Between the two extremes, there is a reasonable spread over the models. There is a significant main effect among the different models ($P < 0.05$, *Friedman Test*, $\chi^2 = 19.07$, $df = 4$). In post-hoc analysis, the baseline model is the worst performer and is not significantly different from the M_NMF model ($P > 0.05$, *paired Wilcox Test, two tailed*). All the multi-stage models suppress the accompaniment significantly better than the baseline model ($P < 0.05$, *paired Wilcox test, two-tailed*). We did not perform contrasts between the respective multi-stage models because we are chiefly interested in whether the multi-stage models offer an improvement over the baseline model.

Combining the evidence from the two respective listening tests, with respect to the performance of the baseline model, for multi-stage model M_DNN1 there is a significant perceptual improvement in accompaniment suppression and no associated evidence of a corresponding drop in vocal sound quality. However, for the alternative multi-stage models (M_DNN2, M_DNN3, M_NMF) although there is evidence of significantly better suppression than baseline, there is also evidence of correspondingly significantly worse distortion than baseline. Therefore, in these cases, it would appear that there has been a trade-off [19], with improved accompaniment suppression coming at the expense of vocal sound quality.

5 Conclusion and Discussion

In this paper, we have described and demonstrated a psychophysical evaluation method for audio source separation. Our method has been demonstrated in the context of vocal separation from musical mixtures. In contrast to the prevailing MUSHRA paradigms [4-9], our perceptual results are highly correlated with the physical measures SAR and SIR. Thus, our results tend to suggest that the previously reported failures of the physical measures to correlate with perceptual data [4-9] may be the inherent result of methods which do not hold to the necessary psychophysical principles. In addition, our psychophysical paradigm paves the way for the development of psychophysical models (e.g., see [19]) more suitable to act as bridge between the physical measures and the quality-of-experience measures which are more informed by the practical uses of and motivations for source separation. Future work is necessary to determine whether these preliminary results are generalizable to stimuli with a wider distribution of physical measurement values and a larger cohort of listeners.

We have also demonstrated that the psychophysical evaluation approach is suitable for comparison of competing audio source separation methods. For one of the multi-stage deep neural network separation methods, the combined results of the two experiments described here capture improved accompaniment suppression without any evidence of a corresponding penalty in the associated vocal quality. By contrast, the alternative multi-stage models appear to achieve their suppression at the cost of a trade-off [19] of improved suppression for added distortion. Future work should include generalisation of the psychophysical paradigm to a larger range of stimuli and a larger cohort of listeners. In addition, some means to obtain uniformly distributed physical measures would improve the interpretability of the results.

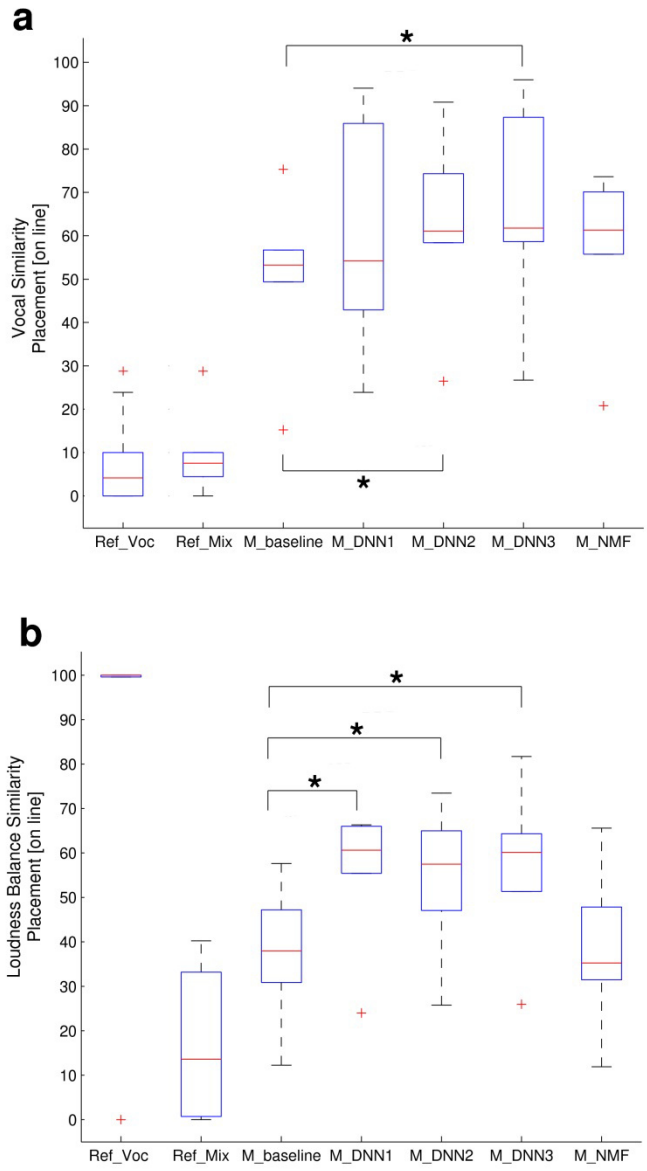

Fig. 4. Model Comparison: *Perceived Similarity.* Listeners organized sliders representing the seven respective stimuli along perceptual lines which depict the perceived similarity. **a** plots the data of experiment 1, capturing the similarity of the vocal component of the respective stimuli. **b** plots the respective data of experiment 2, capturing the loudness balance similarity. *Ref_Voc* refers to the original vocal signal, *Ref_Mix* refers to the mixture, *M_baseline* refers to the baseline DNN, *M_DNN1-3* refer to the respective multi-stage DNNs and *M_NMF* refers to the NMF model. Medians are shown in red. Boxes describe inter-quartile range and 'whiskers' indicate 95% confidence intervals. Bars with asterisks denote significant differences ($P < 0.05$, *paired Wilcox Test*). All other contrasts are not significant ($P > 0.05$, *paired Wilcox Test*).

6 Acknowledgment

This work was supported by grants EP/L027119/1 and EP/L027119/2 from the UK Engineering and Physical Sciences Research Council (EPSRC). The authors also wish to thank the reviewers for helpful comments on an earlier version of the paper.

7 References

1. Vincent E, Gribonval R, Févotte C (2006) "Performance measurement in blind audio source separation", IEEE Trans. on Audio, Speech and Language Processing 14:1462-1469.
2. Vincent E, Jafari MG, Plumbley MD (2006) "Preliminary guidelines for subjective evaluation of audio source separation algorithms", In: A K Nandi and X Zhu (eds.), Proc. ICA Research Network International Workshop, Liverpool, UK, pp. 93-96.
3. ITU (2014) "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems".
4. Emiya V, Vincent E, Harlander N, Hohmann V (2011) "Subjective and objective quality assessment of audio source separation," IEEE Transactions on Audio, Speech, and Language Processing 19: 2046-2057.
5. Cartwright M, Pardo B, Mysore GJ, Hoffman M (2016) "Fast and easy crowdsourced perceptual audio evaluation," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 619 - 623.
6. Kornycky J, Gunel B, Kondoz A (2008) "Comparison of subjective and objective evaluation methods for audio source separation," in Meetings on Acoustics, vol. 123, no. 5, Paris, France, p. 3569.
7. Langjahr P, Mowlaee P (2013) "Objective Quality Assessment of Target Speaker Separation Performance in Multisource Reverberant Environment," in 4th Int. Workshop on Perceptual Quality of Systems, Vienna, Austria, pp. 89-94.
8. Gupta U, Moore E, Lerch A (2015) "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 2015).
9. Cano E, FitzGerald D, Brandenburg K (2016) "Evaluation of Quality of Sound Source Separation Algorithms: Human Perception vs Quantitative Metrics", EUSIPCO 2016, pp. 1758-1762.
10. Fechner GT (1860) "Elemente der Psychophysik," Leipzig: Breitkopf und Härtel.
11. Gescheider G (1997) "Psychophysics: the fundamentals," (3rd ed.), Lawrence Erlbaum Associates.
12. Fletcher H, Munson WA (1933) "Loudness, its definition, measurement and calculation," The Journal of the Acoustical Society of America 5: 82-108.
13. Moore BCJ (2012) "An Introduction to the Psychology of Hearing," (6th ed.), Brill.
14. Grais EM, Roma G, Simpson AJR, Plumbley MD (2016) "Discriminative Enhancement for Single Channel Audio Source Separation using Deep Neural Networks", arXiv abs:1609.01678.
15. Ono N., Rafii Z, Kitamura D, Ito N, Liutkus A (2015) "The 2015 Signal Separation Evaluation Campaign." In Latent Variable Analysis and Signal Separation: 12th International Conference, E. Vincent et al. (Eds.): LVA/ICA 2015, LNCS 9237, pp. 387-395.
16. Terrell MJ, Simpson AJR, Sandler M (2014) "The Mathematics of Mixing", Journal of the Audio Engineering Society, 62(1/2), 4-13.
17. Dwass M (1957) "Modified Randomization Tests for Nonparametric Hypotheses". Annals of Mathematical Statistics 28: 181-187.
18. Simpson AJR, Roma G, Grais EM, Mason RD, Hummersone C, Liutkus A, Plumbley MD (2016) "Evaluation of Audio Source Separation Models Using Hypothesis-Driven Non-Parametric Statistical Methods", European Signal Processing Conference (EUSIPCO) 2016.
19. Simpson AJR, Roma G, Plumbley MD (2015) "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network", in proc. Int. Conf. Latent Variable Analysis and Signal Separation, pp. 429-436.