BSS EVAL OR PEASS? PREDICTING THE PERCEPTION OF SINGING-VOICE SEPARATION

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

[1] Centre for Vision, Speech and Signal Processing
[2] Institute of Sound Recording
University of Surrey, UK

ABSTRACT

There is some uncertainty as to whether objective metrics for predicting the perceived quality of audio source separation are sufficiently accurate. This issue was investigated by employing a revised experimental methodology to collect subjective ratings of sound quality and interference of singing-voice recordings that have been extracted from musical mixtures using state-of-the-art audio source separation. A correlation analysis between the experimental data and the measures of two objective evaluation toolkits, BSS Eval and PEASS, was performed to assess their performance. The artifacts-related perceptual score of the PEASS toolkit had the strongest correlation with the perception of artifacts and distortions caused by singing-voice separation. Both the source-to-interference ratio of BSS Eval and the interference-related perceptual score of PEASS showed comparable correlations with the human ratings of interference.

***Index Terms*—** Audio quality assessment, subjective evaluation, source separation

1. INTRODUCTION

High-quality separation of the singing voice from accompanying instruments is an important yet difficult task serving many applications, from remixing and upmixing music [1] to increasing vocal intelligibility for the hearing impaired [2]. Unfortunately, source separation introduces distortions and artifacts, consequently degrading the *sound quality* of the extracted source. A second issue is *interference*, whereby the unwanted sources remain present to some extent. It would therefore be useful to know how well source separation techniques are suited to a given application. This requires perceptual evaluation where experienced listeners judge real systems according to different perceptual attributes. An alternative is to employ objective metrics which have been developed to predict human perception. The purpose of this paper is to assess the performance of two predictors of audio source separation quality: BSS Eval [3] and PEASS [4].

This work is supported by grant EP/L027119/2 from the UK Engineering and Physical Sciences Research Council (EPSRC).

1.1. Previous work

The Blind Source Separation Evaluation (BSS Eval) toolkit [3, 5] decomposes the error between the target source and the extracted source into a target distortion component reflecting spatial or filtering errors, an artifacts component pertaining to artificial noise, and an interference component associated with the unwanted sources. The salience of these components is quantified using three energy ratios: source Image-to-Spatial distortion Ratio (ISR), Source-to-Artifacts Ratio (SAR), and Source-to-Interference Ratio (SIR). A fourth metric, the Source-to-Distortion Ratio (SDR), measures the global quality (all impairments combined).

A perceptually-motivated adaptation of this toolkit is PEASS (Perceptual Evaluation method for Audio Source Separation) [4, 6], which estimates the three distortion components from auditory representations of the reference and extracted sources, which are then input to the PEMO-Q auditory model [7] to measure their salience. In the final stage, a neural-network trained on human data combines the resulting component-wise salience features into four objective predictors: Target-related Perceptual Score (TPS), Artifacts-related Perceptual Score (APS), Interference-related Perceptual Score (IPS), and Overall Perceptual Score (OPS). The subjective data were obtained from a "MUlti-Stimulus test with Hidden Reference and Anchor" (MUSHRA) [8] listening assessment in which listeners were asked to rate target preservation, absence of artificial noises, suppression of other sources, and overall quality of 10 audio excerpts (primary speech/singing voice) estimated using 13 source separation algorithms. Vincent [6] later revised the model parameters to increase the correlation with the mean opinion scores of the same subjective data.

Despite the development of evaluation toolkits, there is some conflicting and inconclusive evidence as to their perceptual relevance. For example, Cano et al. [9] performed a correlation analysis to compare the measures of BSS Eval and PEASS with the mean opinion scores obtained from a MUSHRA experiment. They asked the same four questions as Emiya et al. [4], but used musical sounds as estimated by two harmonic-percussive separation algorithms. An across-

song correlation between the subjective scores and objective values showed that PEASS performed slightly better than BSS Eval, but that the correlations were weak and inconsistent across the two separation algorithms, indicating poor generalisation to other sources and types of algorithms.

Gupta et al. [10] conducted an experiment in which listeners were asked to rate the overall quality, interference, and intelligibility of vocal and accompaniment excerpts extracted by four singing-voice separation algorithms from nine songs. A correlation analysis was performed between each participant's rating and the BSS Eval measures, but the effect sizes were not consistently high, with wide confidence intervals. The authors concluded that SIR and SAR provided some indication of the perceived vocal isolation and intelligibility, respectively, and that overall quality correlations were generally poor. Cartwright et al. [11] repeated the original PEASS experiment [4], and found consistent positive correlations for all four BSS Eval statistics (PEASS was not assessed), with the highest being around 0.75 for SIR (interference) and 0.55 for SAR (artificial noise). Finally, Simpson et al. [12] asked listeners to rate the overall similarity of 10 vocal segments, whilst ignoring the accompaniment, extracted by five algorithms against the original source. They carried out a second experiment in which listeners judged the amount of interference indirectly by rating the similarity of the vocal-to-accompaniment loudness ratio to that of the original mixture. Simpson et al. reported high within-song Pearson correlations of around 0.91 for both SAR (similarity) and SIR (interference). Their correlations were, however, likely inflated by including the original mixture in the objective measurement, as this stimulus is often an outlier, especially in terms of SAR.

1.2. This work

The previous studies suggest that there is an association between the BSS Eval measures and perceptual characteristics of source separation algorithms when applied to speech/singing voice, but that the strength of these relationships depends on the perceptual attribute that listeners are asked to judge when rating different systems. Furthermore, the predictive success of PEASS when applied to new subjective data remains unknown. The present work investigates these issues by assessing both toolkits in terms of predicting *sound quality* and *interference* of singing voices extracted from musical mixtures. A revised experiment design is presented whereby the sound-quality rating scale is modified to better assess the influence of distortions and artifacts on perceived quality, independently of interference. In contrast to previous work, a broader sample of source separation algorithms (21) and mixtures (16) have been collated to better assess the generalisation of these metrics.

2. SUBJECTIVE ASSESSMENT

In previous MUSHRA assessments [4, 9, 11], listeners rated the quality of each test sound compared to a reference sound (the original isolated source) in terms of global quality (all impairments combined), preservation of the target source, suppression of other sources, and absence of additional artificial noise. However, in our previous experiment [13], listeners found it difficult to separate specific distortions when auditioning the output of real systems, which agrees with the post-hoc observations of Emiya et al. [4] and Cartwright et al. [11]. We therefore simplified the task by asking listeners to assess stimuli according to two criteria: **Sound quality** relates to the amount of artifacts and distortions that you can perceive, ranging from *worse quality* to *same quality*, with respect to the reference sound; **Interference** describes the loudness of the instruments compared to the loudness of the vocals, ranging from *strong interference* to *no interference*. Training examples were used to emphasize that sound quality focuses on general distortions and not the presence of accompanying instruments. Similar examples were presented to explain that interference should be judged independently of such distortions. The perception of global quality [4], and thus the evaluation of all-encompassing performance metrics like SDR and OPS, is the subject of future work.

2.1. Procedure

Our test interface was based on MUSHRA [8]. The listener clicked a 'reference' button to audition a reference singing voice, and clicked and dragged sliders to play and rate eight test sounds on a scale from 0–100, respectively.[1] Unlike MUSHRA, the scores were hidden from the listener and only the end points of the scale were labelled. These modifications were made to reduce potential bias effects introduced by verbal labels [14].

Participants were asked to rate the sound quality and perceived interference of eight test sounds in comparison to a reference. Sixteen vocals, each from a different song, were used, with one excerpt randomly selected (for each listener and task) as a replicate, resulting in 17 trials per task. The replication allowed for the measurement of intra-rater agreement and facilitated post-screening of participants. Both the order of the trials and the order of the test sounds within trial was random, and task order was counterbalanced across participants. Project resources can be found on the GitHub repository associated with the online assessment.[2]

2.2. Stimuli

Eight test sounds were used per trial: a hidden reference (the original vocal excerpt), two hidden anchors, and five vocals

[1] The interface was developed using https://github.com/deeuu/listen

[2] https://cvssp.github.io/perceptual-study-source-separation/

extracted from the mixture by five different source separation algorithms. The reference vocals were taken from the Demixing Secret Database [15], a set of 100 rock and pop songs each comprising four sources: bass, drums, vocals (lead and backing), and 'other'. This database was compiled to assess 23 source separation algorithms competing in the 2016 Signal Separation Evaluation Campaign (SiSEC16) [15], from which the submitted audio files were kindly provided by Fabian-Robert Stöter. We selected 16 songs and five different algorithms per song, using a sampling procedure which achieves a range of distortions and interference levels according to SAR and SIR [13]. The resulting stimuli comprised vocals estimated by 21 source separation algorithms.

In the MUSHRA protocol, low-quality anchors are test sounds that have been included (unbeknownst to the listener) to represent large impairments. In previous work [13] we found the artifacts and target distortion anchors defined by Emiya et al. [4] to be of higher quality than the worst performing algorithms of SiSEC16. We therefore modified their specifications to establish a more appropriate anchor for the sound-quality task. The sound-quality anchor was generated by removing 20% of the time frames from the spectrogram of the reference and lowpass filtering it with a cutoff frequency of 3.5 kHz. Musical noise was then created by randomly removing 99% of the time-frequency bins from a second spectrogram before applying the same lowpass filter. The inverse of these two spectrograms were loudness normalized according to ITU-R BS.1770 [16] and then summed. The original mixture associated with each reference vocal was used as the interference anchor. All stimuli were shortened to seven seconds, converted to mono, and then loudness normalized [16].

2.3. Participants

The listening assessment involved 24 listeners, 18 of which were assessed in an audio booth at CVSSP, and six experienced listeners completed the test online. Of the 24 participants, three were female and 21 were male, and all were aged between 21 and 41, with no known hearing impairments. Stimuli were reproduced over headphones.

3. ANALYSIS AND RESULTS

Each participant's per-trial ratings were min-max scaled such that the sound with the lowest rating was equal to zero and the sound with the highest rating was equal to 100 [8]. In what follows, median (second quartile) values are supplemented with measures of spread using the interquartile range (IQR) which is the difference between the third and first quartile.

3.1. Descriptive analysis

Intra-rater agreement was evaluated using the concordance correlation coefficient [17], which ranges between -1 (perfect negative agreement) and 1 (perfect agreement), applied

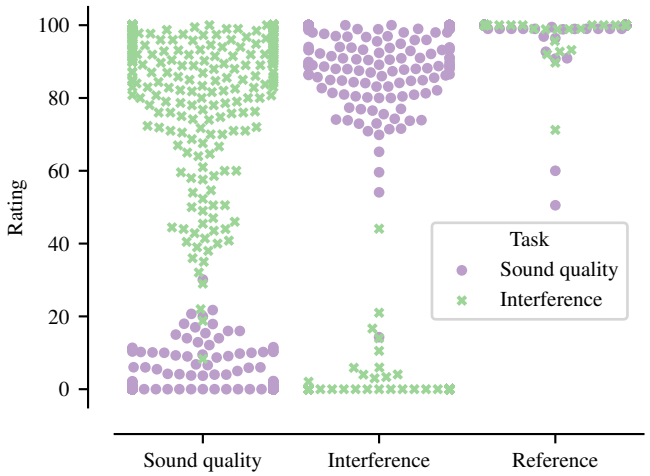

Fig. 1. Bee swarm plot of *all* ratings assigned to the hidden sound-quality and interference anchors, and the hidden reference in each task.

to the paired scores obtained from the replicated trials. The three hidden stimuli were removed to measure agreement on the systems under test only. The median of the 24 participant correlations was 0.79 (IQR = 0.39) for the sound-quality task, and 0.82 (IQR = 0.21) for the interference task. Although the magnitude of the two coefficients are comparable, the between-listener spread is roughly twice as large for the sound-quality task.

Fig. 1 shows all ratings assigned to the two hidden anchors and the hidden reference in each task. Ratings close to zero were expected for the sound-quality anchor in the quality task and for the interference anchor in interference task as these anchors were designed to emphasize low-quality degradations beyond those exhibited by real systems. It can be seen that listener agreement is highest when judging each anchor in its associated task. However, listeners were less certain when judging the interference present in the sound-quality anchor, which suggests that they were uncomfortable assigning 'no interference' to artificial musical noise. The figure also highlights that the majority of listeners were able to identify the hidden references.

Following Gupta et al. [10], inter-rater agreement was measured using Krippendorff's α [18], which ranges from 0 (absence of reliability) to 1 (perfect reliability). With the three hidden stimuli excluded, the across-song median α was 0.34 (IQR = 0.12) for the sound-quality task and 0.40 (IQR = 0.07) for the interference task. We repeated the analysis using the rank transformed rating data, i.e. treating the data as ordinal, and obtained higher medians of 0.77 (IQR = 0.17) for the sound-quality task and 0.81 (IQR = 0.19) for the interference task. This suggests that listeners were consistent with one another as to the relative ordering of the algorithms, and that the lower absolute agreement can be attributed to between-listener differences in the use of the rating scale.

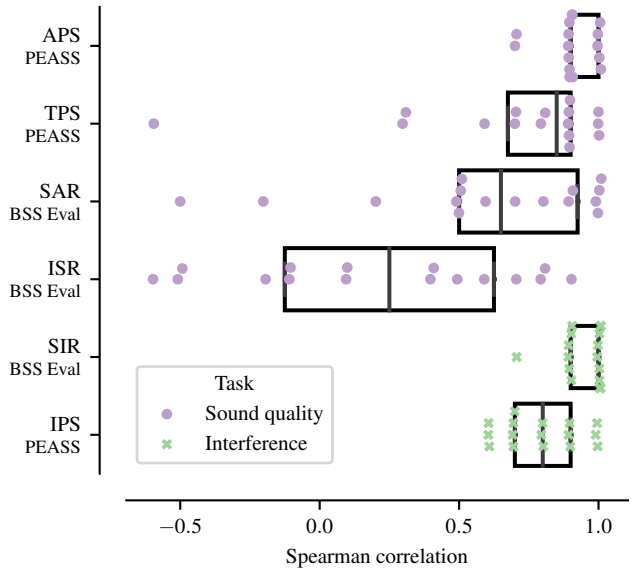

Fig. 2. Bee swarm plots (with boxplots underlaid) of the within-song Spearman correlations. Markers have been perturbed by ± 0.01 to facilitate visual separation.

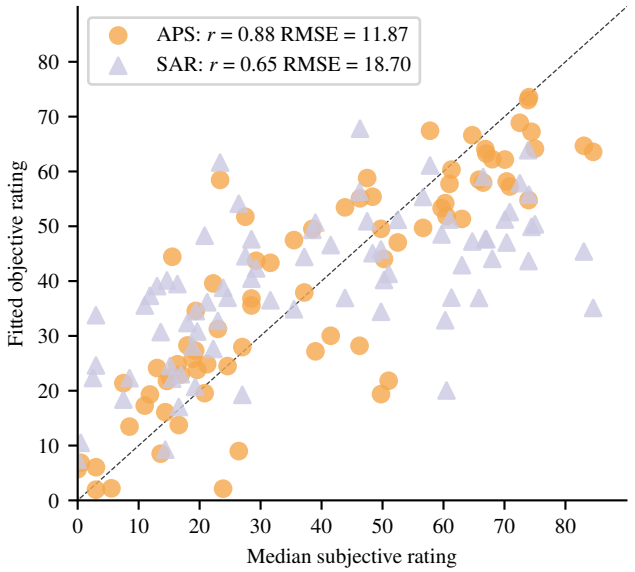

Fig. 3. Linear-regression fitted APS and SAR ratings versus medians of the subjective ratings for all test sounds.

3.2. Objective metrics

A Spearman correlation analysis, which assesses rank agreement between two variables, was first used to assess the performance of two objective toolkits: BSS Eval (SAR/ISR/SIR) and PEASS (APS/TPS/IPS). Objective measurements were made using the same (loudness normalized) stimuli as used in the experiment, where the reference vocal and accompaniment (mixture minus vocal) signals served as ground truth. Correlations were performed for each of the 16 songs with the reference and anchors excluded. Fig. 2 shows the correlations measured using each predictor for the appropriate listening task. APS performed best for the sound-quality task (median = 0.90), and SIR performed best for the interference task (median = 1.00). Although TPS and, to a lesser extent, SAR show high agreement as to the ordering of the algorithms for a few songs, the correlations are scattered over a wider region compared to those of APS. Such inconsistencies are even more pronounced for ISR. IPS correlations were generally strong (median: 0.80) for the interference task, but SIR performed more consistently.

Following Cartwright et al. [11], a Pearson correlation coefficient r was calculated using the across-participant medians of all 80 test sounds (16 songs x 5 systems) and the measures of each metric. The correlations obtained using the four sound-quality metrics were $r_{\text{APS}} = 0.88$, $r_{\text{TPS}} = 0.79$, $r_{\text{SAR}} = 0.65$, and $r_{\text{ISR}} = 0.28$. These effect sizes indicate that APS yields the strongest relationship with the subjective sound-quality ratings, with ISR performing the worst. Given that previous studies have found associations between SAR and sound-quality perception [11, 12], it is interesting to compare this metric with APS. Fig. 3 shows the regression-fitted measures of both metrics versus the median subjective ratings, and indicates a stronger monotonic upward trend when using APS, though marked deviations are observable for both metrics. Indeed, the root-mean-square error (RMSE; computing with 2 degrees of freedom) between the fitted objective and subjective ratings was 11.9% for APS and 18.7% for SAR, both of practical significance when judged in the context of the 100-point rating scale. The correlations measured using the two interference-based metrics were: $r_{\text{SIR}} = 0.81$ and $r_{\text{IPS}} = 0.81$. After fitting their measures to the subjective ratings, both metrics had an RMSE of 15%, and so we may infer comparable predictive capability.

4. CONCLUSIONS

The perception of sound quality and interference of 16 singing voices extracted by a range of source separation algorithms was measured. By redefining the sound-quality scale, these two perceptual attributes were measured independently of one another. A correlation analysis was used to assess the predictive capability of two objective toolkits for source separation performance evaluation. The results show that the APS metric of the PEASS toolkit had the strongest correlation with the subjective judgements of artifacts and distortions and is therefore a useful metric for performance evaluation. Both SIR of BSS Eval and IPS of PEASS showed comparable correlations with the interference ratings, with the former predicting well the rank order of the algorithms within song. In summary, we encourage researchers to make use of the PEASS toolkit in their evaluations, rather than relying solely on energy-based metrics. Further refinement is, however, warranted to reduce prediction errors to within tolerable limits. Additional experiments are needed to assess these metrics on different types of stimuli and also assess across-song prediction given specific separation algorithms.

5. REFERENCES

[1] G. Roma, E. M. Grais, A. J. R. Simpson, and M. D. Plumbley, "Music remixing and upmixing using source separation," in *2nd Audio Engineering Society Workshop on Intelligent Music Production*, Sep. 2016.

[2] J. Pons, J. Janer, T. Rode, and W. Nogueira, "Remixing music using source separation algorithms to improve the musical experience of cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 140, no. 6, pp. 4338–4349, Dec. 2016.

[3] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[4] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

[5] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Independent Component Analysis and Signal Separation*, Springer Berlin Heidelberg, Sep. 2007, pp. 552–559.

[6] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *Latent Variable Analysis and Signal Separation*, F. Theis, A. Cichocki, A. Yeredor, and M. Zibulevsky, Eds., Springer Berlin Heidelberg, 2012, pp. 430–437.

[7] R. Huber and B. Kollmeier, "PEMO-Q: a new method for objective audio quality assessment using a model of auditory perception," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[8] ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Tech. Rep., 2015.

[9] E. Cano, D. FitzGerald, and K. Brandenburg, "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *24th European Signal Processing Conference (EUSIPCO)*, IEEE, Aug. 2016.

[10] U. Gupta, E. Moore, and A. Lerch, "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)*, IEEE, Oct. 2015.

[11] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Mar. 2016.

[12] A. J. R. Simpson, G. Roma, E. M. Grais, R. D. Mason, C. Hummersone, and M. D. Plumbley, "Psychophysical evaluation of audio source separation methods," in *Latent Variable Analysis and Signal Separation*, P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, Eds., Springer International Publishing, 2017, pp. 211–221.

[13] H. Wierstorf, D. Ward, R. Mason, E. M. Grais, C. Hummersone, and M. D. Plumbley, "Perceptual evaluation of source separation for remixing music," in *143rd Convention of the Audio Engineering Society*, Oct. 2017.

[14] S. Zielinski, "On some biases encountered in modern audio quality listening tests (part 2): Selected graphical examples and discussion," *Journal of the Audio Engineering Society*, vol. 64, no. 1/2, pp. 55–74, Feb. 2016.

[15] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*, P. Tichavský, M. Babaie-Zadeh, O. J. Michel, and N. Thirion-Moreau, Eds., Springer International Publishing, 2017, pp. 323–332.

[16] ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level," International Telecommunication Union, Tech. Rep. 4, 2015.

[17] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, p. 255, Mar. 1989.

[18] K. Krippendorff, "Computing krippendorff's alpha-reliability," University of Pennsylvania, Tech. Rep., 2011. [Online]. Available: http://repository.upenn.edu/asc_papers/43.