ON THE PERCEPTUAL RELEVANCE OF OBJECTIVE SOURCE SEPARATION MEASURES FOR SINGING VOICE SEPARATION

Udit Gupta[1], Elliot Moore II[1], Alexander Lerch[2]

[1]School of Electrical and Comp. Eng., [2]Center for Music Technology
Georgia Institute of Technology, Atlanta, GA, USA
{uditgupta, em80, alexander.lerch}@gatech.edu

ABSTRACT

Singing Voice Separation (SVS) is a task which uses audio source separation methods to isolate the vocal component from the background accompaniment for a song mix. This paper discusses the methods of evaluating SVS algorithms, and determines how the current state of the art measures correlate to human perception. A modified ITU-R BS.1543 MUSHRA test is used to get the human perceptual ratings for the outputs of various SVS algorithms, which are correlated with widely used objective measures for source separation quality. The results show that while the objective measures provide a moderate correlation with perceived intelligibility and isolation, they may not adequately assess the overall perceptual quality.

Index Terms— Singing Voice Separation, Source Separation, Music Information Retrieval, MUSHRA

1. INTRODUCTION

Singing Voice Separation (SVS) has gained prominence as a Music Information Retrieval (MIR) task in the recent years. The goal for this task is to separate the lead vocals from the accompaniment for professionally produced songs. Various algorithms have been proposed which perform this task using diverse approaches. SVS is often used as a pre-processing step in other MIR tasks such as automatic lyrics recognition [1], singer identification [2–4], query by singing/humming [5], etc. It may also be useful in the context of applications such as karaoke, musical education, and audio remixing.

In MIREX, the Music Information Retrieval Evaluation eXchange, the performance of eleven submissions for the SVS task was tested [6]. The quality of the output produced by these algorithms was evaluated with objective measures such as NSDR (Normalized Signal to Distortion Ratio), SIR (Signal to Interference Ratio) and SAR (Signal to Artifacts Ratio) [6–9]; although these measures are widely used, it is not well understood how they compare to the perceived quality of the source separation as assessed by a human. In this paper, we aim to bridge this gap, provide an evaluation of these objective measures in the context of SVS, and investigate their correlation to the subjective quality as reported by human test subjects.

In Section 2 we discuss the current state-of-the-art measures used to evaluate source separation. Section 3 describes the experimental methodology used for performing perceptual evaluation. The analysis of the experiment as well as its results are presented in Section 4. Section 5 concludes the discussion.

2. EVALUATION OF SOURCE SEPARATION SYSTEMS

In order to consistently evaluate and compare the performances of different SVS algorithms, the use of a common scoring system is essential. There are many examples where subjective evaluation has been performed for comparing general audio source separation systems [10–12]. Many of these methods are geared towards evaluating source separation in speech-only mixtures and do not transfer elegantly to SVS evaluation where vocals are mixed with instrumental accompaniment. Last but not least, listening tests are time-consuming, have to be carefully planned, and are usually restricted to a relatively small subset of audio files. To counteract this issue some objective methods for performance evaluation have been suggested.

Emiya et al. [7, 8] have suggested objective measures based on the presence of target spatial distortion (Image to Spatial Distortion Ratio, ISR), interference (SIR), and artifacts (SAR) in the separated signals as compared to the clean source signals. The total distortion in the output signal compared to the source is measured by Signal to Distortion Ratio (SDR) [13]. SIR and SAR were used to evaluate the submissions in the MIREX 2014 - SVS task, along with the normalized version of SDR designated as NSDR [14]. Moderate correlation, in the range of 0.3 to 0.7, has been reported for these measures when compared to judgment by human evaluators for general audio source separation tasks [8, 15], but how they fare in the context of evaluating separation between vocal and instrumental components from a song remains unknown.

It is the goal of this paper to provide a better insight into the performance of these measures in the context of singing voice separation, and analyze how they compare with the various perceptual qualities human listeners look for while listening to the separated audio samples.

3. EXPERIMENT DESIGN

An experiment was designed to determine how well the objective measures NSDR, SIR and SAR used in the MIREX 2014 competition for the SVS task [6, 14] correlate to perceptual evaluation. Five to ten second long excerpts from pop music songs were processed with existing SVS algorithms and subjects were asked to rate them in a series of listening tasks evaluating factors related to the overall quality, degree of isolation and intelligibility. Four better performing algorithms [16–19] from MIREX 2014 were used to process the excerpts. SIR, SAR, SDR and NSDR measures were calculated for both the estimated vocal component and the estimated instrumental component for each of these using the BSS Eval toolkit [8]. For the

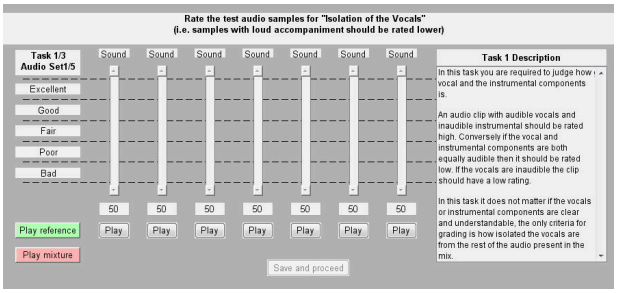


Figure 1: Experiment MUSHRA Interface

purpose of labeling the algorithms have been coded as:

1. II: Ikemiya et al. [16]
2. JL: Jeong and Lee [17]
3. RN: Rao et al. [18]
4. RP: Rafii and Pardo [19]

All the excerpts were processed with the original implementation (provided by authors) of these algorithms, resulting in the audio clips containing the estimated vocals and accompaniment. These processed audio clips, along with the clean audio samples were used to conduct listening tests where the subjects were asked to judge the performance of the algorithms. The listening experiment was conducted as a variation of the ITU-R BS.1543 MUSHRA standard [20]. While the original MUSHRA standard has been proposed for use in evaluation of medium impairments in signal quality by audio codecs, it was modified in the design of anchor audio signals to better conform to SVS system evaluation. PEASS Listening Test GUI [8] was modified and used to conduct the listening test. As shown in Figure 1 the subjects were provided with a graphical interface on a computer screen which allowed them to listen to the test audio samples, the target clip, and the original song excerpt. The test samples comprised of the output produced by the SVS algorithms along with the a hidden reference (same as target clip) and artificially degraded anchor signals as described in 3.4. Along with listening to the audio, slider controls were included in the interface that allowed the subjects to indicate their rating for each of the test audio clips. Each subject was required to rate the clips on a scale of 0 to 100, with equidistant markings providing labels *Bad, Poor, Fair, Good* and *Excellent* going in increasing order.

Two experiments were conducted. In experiment one, the subjects were asked to rate the performance based on the separated vocal component, while in experiment two, the instrumental component was judged. The first required the evaluation to be performed in three separate tasks and the second experiment required two. The tasks were chosen to emulate specific evaluation criteria as described in 3.3.

All the tests were conducted in a quiet environment with identical Dell Optiplex 980 computers and using audio-technica ATH-M30x professional monitor headphones.

3.1. Test Data

Excerpts varying in length from five to ten seconds were extracted from a random selection of nine songs from the MedleyDB database [21]. It was ensured that the selected songs had no cross-talk across the raw tracks. Of the ten selected songs, five were of *pop* or *singer/song writer* genre and the remaining four belonged to the *rock* genre. The test cases generated by mixing the vocals with the accompaniment with equal loudness (sones), along with the clean vocal and accompaniment audio signals, were used as baseline audio clips for these excerpts. The subjects were asked to rate five randomly chosen excerpts which provided good balance between consistency and statistical significance.

3.2. Subjects

Subjects were gathered from a normal hearing population of graduate and undergraduate students, with ages varying from nineteen to thirty-six, to participate in the experiment. Out of thirty subjects who participated, eleven had experience in a music related field and six were professionally trained in music and/or had studio recording experience. The others were not trained in music. The number of male participants was twenty-five, while five were female. Since the purpose of the experiment is to determine the performance of SVS algorithms which are expected to have moderate to severe impairments, no pre-experiment screening of the subjects was performed.

3.3. Evaluation Tasks

The listening test was divided into several tasks in order to obtain consistent perceptual rating across subjects. It was conducted in two sessions. In the first session (experiment one), which lasted generally from fifteen to twenty minutes, the subjects were asked to evaluate the vocal component. This evaluation was separated into three tasks.

T1.1 Vocal Isolation: The subjects were asked to judge how well-isolated the vocals were from the accompaniment. They were instructed to disregard all other factors.

T1.2 Vocal Intelligibility: The subjects were asked to judge how understandable the lyrics were, disregarding all other factors.

T1.3 Vocal Overall Quality: The subjects were asked to judge their overall perception of the quality of the algorithms, taking all impairments into account.

The second session (experiment two) lasted for ten to fifteen minutes and required the subjects to evaluate the separated instrumental component. This session had two tasks.

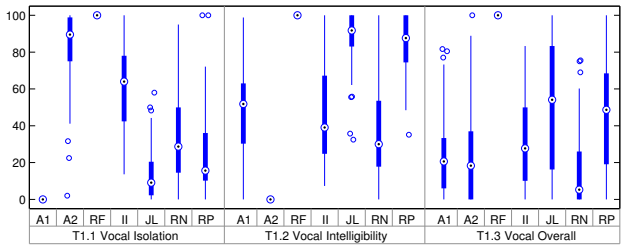
T2.1 Instrumental Isolation: The subjects were asked to judge how well-isolated the accompaniment was from the vocals. They were instructed to disregard all other factors.

T2.2 Instrumental Overall Quality: The subjects were asked to judge their overall perception of the quality of the algorithms taking all impairments into account.

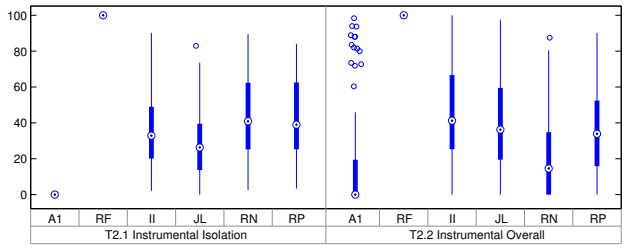
Each session was prefaced by a short demonstration video where the tasks were explained to the subjects. Written instructions for the tasks were also provided to the subjects for the duration of each session.

3.4. Anchors

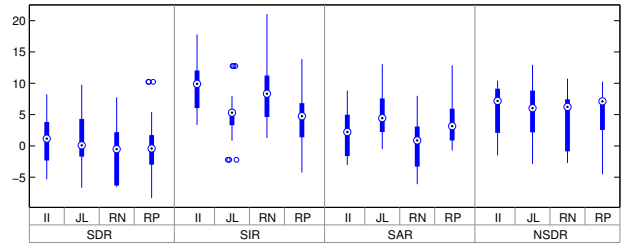
The purpose of the anchors in the experiment is to provide artificially degraded audio samples which should provide the experiment, along with the hidden reference, some control values to perform a post-rating subject screening. The anchors are designed such that depending on the assessment task, the subject will provide a very low score to the anchor sample. In this experiment we use two anchors.



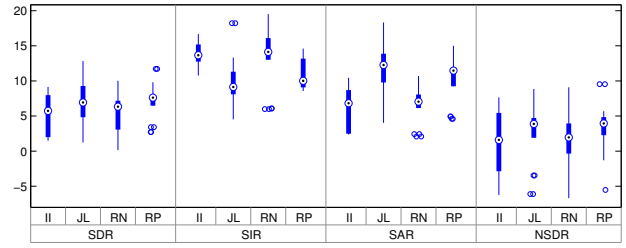
(a) Experiment one Results (Vocals)



(b) Experiment two Results (Accompaniment)



(c) Objective Measures (Vocals)



(d) Objective Measures (Accompaniment)

Figure 2: Results from the experiment (A1 and A2 are the two anchors, RF is the reference and the rest are the test algorithms)

A1 Isolation Anchor: The isolation anchor is produced by passing the original excerpt (mix of vocal and instrumental) to a 4 KHz low-pass filter, and amplifying the result to match the original loudness. This anchor helps in validating the subject's ratings for the two isolation tasks.

A2 Intelligibility Anchor: The intelligibility anchor is generated by passing the clean vocal audio to a 500 Hz low-pass filter. The result is then amplified such that the average Zwicker Loudness (ISO 532B) [22,23] in sone, for the result, is equal to the loudness of the original audio sample. This anchor is useful for determining the validity in the vocal intelligibility task.

4. ANALYSIS AND RESULTS

4.1. Post Evaluation Screening

The experiment described in Section 3 above involves a subjective study of human perception. This necessitates that a post evaluation screening be performed where the ratings for the subjects who may not have understood the task, or who may be outliers in the group is removed. To remove the ratings for subjects who may not have understood the task, the ratings which don't have the hidden reference marked as hundred are removed. Also it is expected that for the isolation tasks (T1.1 and T2.1), the isolation anchor A1 will have the poorest performance and similarly for the intelligibility task (T1.2), anchor A2 is expected to have the worst score. The subject ratings which do not conform to this are also removed from further consideration. On an average three to four subjects' ratings were removed for each task. The perception of the overall quality of the anchors can not be predicted as it is subject dependent, hence the anchors play no role in the quality assessment tasks (T1.3 and T2.2).

For each task the Spearman's Correlation Coefficient (ρ) of the individual subjective ratings v. the average rating of remaining subjects is found. The distribution of the ρ values for each of the tasks is modeled as a truncated t-distribution, and the subject ratings which have ρ values less than the five percent outlier limit on the lower tail of the distribution are removed for that task [24]. Using this method, one to three subject ratings were removed from each task. The parameters for the distribution of the estimated distributions, their chi-squared statistic from Pearson's chi-squared goodness of fit test, and the outlier limits for each task are listed in Table 1. The lower the value of the chi-squared statistic, the better is the approximation for the outlier limits. It can be inferred by the mean and the standard deviations listed in the table that the subjects agree more with each other in the matters of isolation and intelligibility than they do for the overall quality assessment. This may suggest that the perceived annoyance of different artifacts varies between subjects.

Table 1: Distribution of Spearman's correlation of subject ratings and associated outlier limits

Task	μ	σ	χ^2	Outlier Limit
T1.1 Vocal Isolation	0.856	0.050	0.211	0.713
T1.2 Vocal Intelligibility	0.881	0.049	1.206	0.800
T1.3 Vocal Overall	0.637	0.254	4.571	0.209
T2.1 Instrumental Isolation	0.804	0.068	0.925	0.668
T2.2 Instrumental Overall	0.652	0.147	1.265	0.373

4.2. Descriptive Analysis

Exploratory statistical analysis is performed for the results obtained from the experiment. In Figure 2 the trends for subjective and objective ratings for all the different test sets are shown. Figures 2a and 2b show the distribution of the subjective ratings according to the tasks. The objective evaluation metrics for the various algorithms under test are visualized in Figures 2c and 2d for the estimated vocals and accompaniment respectively. The whiskers in Figures 2a and 2b indicate that the variances in the subjective ratings for the intelligibility (T1.2) and isolation tasks (T1.1, T2.1), are smaller than for the overall quality tasks (T1.3, T2.2). This is in agreement

with the inference from Table 1 in Section 4.1.

In Figures 2a and 2b a large variance is seen in some of the results. This is due to the intra-algorithmic performance differences across different audio excerpts as well as the subjects using the MUSHRA scale differently. To ensure that the subject ratings are concordant with each other and the results are reliable, an inter-rater reliability test is performed for each task by calculating Krippendorff's alpha coefficient [25] for the subject ratings of all the clips processed from the same excerpt. The median values for Krippendorff's alpha coefficient, along with the maximum and minimum, are displayed in Figure 3 for each of the tasks. An alpha value of one means perfect reliability, while alpha of zero indicates the absence of reliability. The inter-rater reliability is high in the case of the vocal intelligibility (T1.2) and the two isolation tasks (T1.1, T2.1) and the inferences derived from them can be considered conclusive. However, in the case of the overall quality assessment tasks (T1.3, T2.2) the reliability is not very high and therefore they have a lesser inferential capability as compared to the other three tasks.

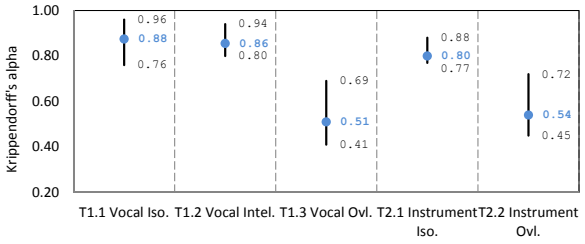


Figure 3: Per-excerpt statistics of Krippendorff's alpha for each task

Since the subjects who participated in the experiment were from various musical backgrounds, a similar analysis of pairwise ρ values was performed individually for non-musically trained subjects and the musically trained ones. The distributions obtained were compared with each other using a two-sample Kolmogorov-Smirnov test. The test indicated that the two sample distributions may be originating from identical population distributions and failed to reject the null hypothesis (that the samples are from identical distributions) with a p-value of 0.30.

4.3. Comparison of Objective Measures

In order to compare how well the objective measures generally used for SVS quality evaluation (compare MIREX 2014) correspond to the subjective evaluation performed in the experiment, two results have been compiled. In the first case the objective measures NSDR, SIR, and SAR for the estimated vocals are correlated with each subject's rating of the same excerpt using the Pearson's Product Moment Correlation Coefficient. This analysis is performed for each task of the first experiment. Similarly the corresponding objective measures for the estimated accompaniment are correlated with the ratings for tasks of the second experiment. The mean and ninety-five percent confidence interval for the Pearson's correlation is listed in the left half of Table 2. The confidence interval is found using a non-parametric estimate of the probability distribution function [26] for the correlation coefficient for each task, and estimating the region which corresponds to the central ninety-five percent area under the curve.

The same analysis is repeated with the Spearman's Rank Correlation Coefficient instead of the Pearson's Correlation Coefficient. The results for this analysis are shown in the right half of the table. Pearson's correlation coefficient provides a measure of linear dependence of the objective measures being tested against the subjective ratings for the various tasks while Spearman's correlation coefficient estimates if the objective measures and the subjective ratings have a monotonic relationship.

Table 2: Average Pearson's and Spearman's correlation coefficients for objective v. subjective ratings with ninety-five percent confidence intervals. The instances which show a significant positive or negative correlation by not having zero within the confidence interval have been highlighted.

Task	Pearson's Correlation Coeff.			Spearman's Correlation Coeff.		
	NSDR	SIR	SAR	NSDR	SIR	SAR
T1.1 Vocal Isolation	0.122 [-.95,+.97]	0.567 [-.65,+1.00]	-0.332 [-1.00,+.82]	0.409 [-.53,+.78]	0.701 [-.05,+.88]	0.069 [-.72,+.61]
T1.2 Vocal Intel.	0.089 [-.83,+.95]	**-0.726 [-1.00,-.13]**	**0.739 [+.07,+1.00]**	0.270 [-.57,+.67]	-0.463 [-.86,+.09]	**0.816 [+.32,+.93]**
T1.3 Vocal Overall	0.119 [-.90,+.96]	-0.278 [-.98,+.94]	0.370 [-.94,+1.00]	0.348 [-.55,+.74]	0.116 [-.71,+.66]	0.589 [-.38,+.84]
T2.1 Instr. Isolation	0.149 [-.74,+.88]	0.262 [-.88,+.99]	-0.145 [-.97,+.85]	0.399 [-.44,+.74]	0.469 [-.41,+.79]	0.197 [-.63,+.63]
T2.2 Instr. Overall	0.136 [-.86,+.95]	-0.020 [-.94,+.92]	0.092 [-.90,+.92]	0.384 [-.54,+.76]	0.364 [-.53,+.74]	0.282 [-.57,+.69]

From the highlighted results in Table 2 it can be observed that SIR and SAR have comparably high correlations (absolute value) for the vocal intelligibility task (T1.2) which provides evidence that these measures might be related to perceptual intelligibility. Since the confidence intervals are large in all the cases, and span from positive to negative it is not certain if these values are a chance occurrence. For NSDR the correlation varies greatly and has average values near zero for all the tasks. This may indicate that NSDR is not an important measure for SVS evaluation in a perceptual sense. None of the objective measures for the remaining tasks (T1.3, T2.x) show a high correlation. While this in itself is not a conclusive proof that they may not have a good predictive value for these perceptual tasks, the evidence does support the likelihood for it to be true to be high. In order to verify the conclusions drawn here, scatter plots for each objective measure vs. task ratings were plotted and no observable dependence was found. These results are in agreement with Emiya et al. [8], although in their case the evaluation was performed for source separation in general audio mixtures.

5. CONCLUSION

The purpose of this paper was to evaluate the performance of general source separation measures (as employed by MIREX 2014) in the context of separation of the vocal and instrumental components from songs. The analysis used human perceptual quality ratings in terms of isolation, intelligibility, fidelity, etc. and correlated the these ratings with the state of the art as used in MIREX. While the SIR and SAR measures provide some indication of the evaluation capability for vocal isolation and intelligibility perception, the correlation measured was not very high and the confidence intervals were too large. It is also likely that NSDR may not be perceptually significant as indicated by its lack of agreement with any of the perceptual tasks.

It may be concluded from this paper that the current objective measures may be insufficient for proper evaluation of SVS systems, and it is recommended that new measures be developed.

6. REFERENCES

[1] C. K. Wang, R. Y. Lyu, and Y. C. Chiang, "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker," in *Proceedings of Eighth European Conference on Speech Communication and Technology*, 2003.

[2] A. L. Berenzweig, D. P. W. Ellis, and S. Lawrence, "Using voice segments to improve artist classification of music," in *Proceeedings of Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Jun 2002.

[3] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection." in *Proceeedings of 6th International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 329–336.

[4] A. Mesaros, T. Virtanen, and A. Klapuri, "Singer identification in polyphonic music using vocal separation and pattern recognition methods." in *Proceedings of the 8th International Conference on Music Information Retrieval*, 2007, pp. 375–378.

[5] Y. Hung-Ming, W. H. Tsai, and W. Hsin-Min, "A query-by-singing system for retrieving karaoke music," *Multimedia, IEEE Transactions on*, vol. 10, no. 8, pp. 1626–1637, Dec 2008.

[6] "Mirex 2014: Singing voice separation (results)," accessed: 03/31/2015. [Online]. Available: http://www.music-ir.org/mirex/w/index.php?title=2014:Singing_Voice_Separation_Results&oldid=10571

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

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

[9] ——, "Multi-criteria subjective and objective evaluation of audio source separation," in *Proceedings of Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*. Audio Engineering Society, 2010.

[10] J. Joseph, "Why only two ears? some indicators from the study of source separation using two sensors," Ph.D. dissertation, Indian Institute of Science, Bangalore, India, 2004.

[11] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of subjective and objective evaluation methods for audio source separation," in *Proceedings of Meetings on Acoustics*, vol. 4. Acoustical Society of America, 2008, p. 050001.

[12] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *Signal Processing, IEEE Transactions on*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[13] S. Araki, F. Nesta, E. Vincent, Z. Koldovsk, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011):- audio source separation," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 414–422.

[14] "Mirex 2014: Singing voice separation task," accessed: 3/31/2015. [Online]. Available: http://www.music-ir.org/mirex/w/index.php?title=2014:Singing_Voice_Separation&oldid=10488

[15] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *Independent Component Analysis and Signal Separation*. Springer, 2007, pp. 454–461.

[16] Y. Ikemiya, K. Yoshii, and K. Itoyama, "Singing voice analysis and editing based on mutually dependent f0 estimation and source separation," in *Proceedings of 2015 International Conference on Acoustics, Speech and Signal Processing*, 2015.

[17] I.-Y. Jeong and K. Lee, "Vocal separation from monaural music using temporal/spectral continuity and sparsity constraints," *Signal Processing Letters, IEEE*, vol. 21, no. 10, pp. 1197–1200, 2014.

[18] P. Rao, N. Nayak, and S. Adavanne, "Singing voice separation using adaptive window harmonic sinusoidal modeling," *The Music Information Retrieval Exchange MIREX 2014.*, 2014.

[19] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix." in *Proceeedings of 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 583–588.

[20] M. Schoeffler, F.-R. Stter, B. Edler, and J. Herre, "Towards the next generation of web-based experiments: A case study assessing basic audio quality following the ITU-R recommendation bs. 1534 (mushra)."

[21] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam, and J. Bello, "Medleydb: a multitrack dataset for annotation-intensive mir research," in *Proceedings of 15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[22] B. C. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica united with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[23] E. Zwicker, H. Fastl, U. Widmann, K. Kurakata, S. Kuwano, and S. Namba, "Program for calculating loudness according to din 45631 (iso 532b)." *Journal of the Acoustical Society of Japan (E)*, vol. 12, no. 1, pp. 39–42, 1991.

[24] T. Sporer, J. Liebetrau, and S. Schneider, "Statistics of mushra revisited," in *Proceedings of Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[25] K. Krippendorff, "Computing krippendorff's alpha reliability," *Departmental papers (ASC)*, p. 43, 2007.

[26] G. Wahba, "Bayesian confidence intervals for the cross-validated smoothing spline," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 133–150, 1983.