Evaluation of Quality of Sound Source Separation Algorithms: Human Perception vs Quantitative Metrics

Estefanía Cano
Fraunhofer IDMT
Semantic Music Technologies
Ilmenau, Germany
cano@idmt.fraunhofer.de

Derry FitzGerald
Cork Institute of Technology
Nimbus Centre
Cork, Ireland
derry.fitzgerald@cit.ie

Karlheinz Brandenburg
Fraunhofer IDMT
Semantic Music Technologies
Ilmenau, Germany
bdg@idmt.fraunhofer.de

***Abstract*—In this paper we look into the test methods to evaluate the quality of audio separation algorithms. Specifically we try to correlate the results of listening tests with state-of-the-art objective measures. To this end, the quality of the harmonic signals obtained with two harmonic-percussive separation algorithms was evaluated with BSS_Eval, PEASS and via listening tests. A correlation analysis was conducted and results show that for harmonic-percussive separation algorithms, neither BSS_Eval nor PEASS show strong correlation with the ratings obtained via listening tests and suggest that existing perceptual objective measures for quality assessment do not generalize well to different separation algorithms.**

I. Introduction

All scientific endeavours rely on four basic principles laid down by the *Scientific Method*: (1) characterizations, (2) hypotheses, (3) predictions, and (4) experiments. Sound source separation research is not the exception. Taking the harmonic-percussive separation (HP) problem as an example, an initial characterization could be: How can harmonic signals be separated from percussive ones in an audio mixture? A possible hypothesis could then be: If we apply median filtering over time in the magnitude spectrogram, percussive elements will be reduced resulting in a harmonically-enhanced magnitude spectrogram. During the prediction step, harmonic and percussive signals are extracted using the proposed median filtering approach. Finally, the experiments are meant to determine whether the observations from the real world (original harmonic and percussive signals) agree with or conflict with the predictions derived from the hypotheses (estimated harmonic and percussive signals). In this paper, we focus on the **experiment** step, particularly on analyzing methods for separation quality assessment that attempt to determine the degree of agreement between the observations and the predictions in a sound separation context. Two clear research questions are addressed in this paper: *(1) Are available metrics for separation quality robust methods to assess algorithm performance? (2) Can we systematically and truthfully test separation quality using available quality measures?*

II. Background

The topic of audio quality of processed sound has long been studied. In early work on audio coding [1] attempts were made to find methods to reduce the necessity for expensive large scale listening tests. In audio coding, it early became obvious that there are situations where Euclidian distance measures, like Signal-to-Noise Ratio (SNR) are completely misleading when we look for the best audio quality. Since then, no papers using SNR as a quality measure for audio coding algorithms are even accepted at major conferences. Nonetheless, the goal of reducing the burden of doing listening tests all the time remained. Several algorithms have been proposed which try to simulate the processing of sound in the human auditory system [2], [3]. In the mid 90s, the ITU undertook the work to unify different approaches and in the end passed a recommendation for such techniques, i.e., PEAQ [4]. To this end, a number of large listening tests were conducted to correlate the output of models to the output of listening tests. The final result carries a clear caveat: The PEAQ measurement method can only help to estimate the quality possible with a certain class of coding algorithms. Outside this class (e.g. when Spectral Band Replication was introduced), the correlation can no longer be found.

In the case of sound source separation, three types of evaluation methods have been used in the separation community: listening tests, tests using quadratic error measures, and tests using measures with auditory models built in (often called perceptual quality assessment)

A. Listening Test-based Assessment

Subjective evaluation of audio quality is usually achieved by means of listening tests. In the source separation community however, listening tests have not been very common so far [5]. It is mostly in the audio coding and in the audio systems evaluation communities where active research in this field has been conducted in the past years. Particularly relevant for the separation community is the standard: Method for the subjective assessment of intermediate quality level of coding systems (ITU-R BS.1534-1) [6]. In this standard, the Multiple

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

Stimulus with Hidden Reference and Anchors (MUSHRA) test is defined. The main goal of MUSHRA tests is to evaluate signals of intermediate quality by assessing the degradation of a test signal relative to a known reference. In the specific context of sound separation, the test signal represents the estimated source $\hat{s_j}(t)$ and the reference would be the original recording of the source $s_j(t)$. An adaptation of the MUSHRA test for sound separation evaluation is presented in [5] and in [7] similar listening tests have been conducted.

B. Quadratic Error-based Assessment Measures

BSS_Eval is a set of four performance metrics that evaluates the quality of the extracted source $\hat{s_j}$ by means of energy ratios between the different signal components [8]. These metrics first attempt to decompose the signal into different signal distortions: interference from unwanted sources, sensor noise, and burbling artifacts (musical noise). The extracted source is then decomposed as follows:

$$\hat{s_j} = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{1}$$

where $s_{target} = f(s_j)$ is a version of the original source s_j modified by an allowed distortion f. The terms e_{interf}, e_{noise}, and e_{artif} are the interference, noise, and artifacts error terms, respectively.

The numerical performance criteria are then computed as energy ratios expressed in dB. Namely, Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR), Source to Noise Ratio (SNR), and Source to Artifacts Ratio (SAR) [8]. No perceptual information is used in these measures.

$$SDR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{2}$$

$$SIR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \tag{3}$$

$$SNR = 10\log_{10}\frac{\|s_{target} + e_{interf}\|^2}{\|e_{noise}\|^2} \tag{4}$$

$$SAR = 10\log_{10}\frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2} \tag{5}$$

An important characteristic of these set of objective measures is that they assign equal weights to the different error terms. This assumes that in terms of quality, all types of distortions contribute equally to the overall quality of the extracted source. Another important characteristics of these sets of measures is that they do not take into consideration perceptual aspects of hearing for their calculations.

C. Perceptual Quality Assessment

The development of objective perceptual measures for source quality assessment came as an attempt to create a set of metrics that could function as a proxy for listening tests, in a manner similar to that of PEAQ for audio coding, but in this case focusing on the quality of signals obtained via sound source separation techniques.

The PEASS Toolkit –Perceptual Evaluation Methods for Audio Source Separation– was developed as a set of four objective perceptual measures that attempt to predict the Mean Opinion Scores (MOS) of human listeners by decomposing the signal into different types of distortions; namely, interference, artifacts, and target distortions [7]. MOS scores were obtained by means of a listening test protocol designed to address the perceptual characteristics of the distortions components: target, interference, and artifacts. Objective scores were obtained by calculating the perceptual salience of each specific distortion and of the overall distortion using the PEMO-Q auditory model [7]. Subjective and objective results were joined using non-linear mappings which aimed to combine the salience features obtained with PEMO-Q into a single scalar value, and to adapt the feature scale to the subjective scale from the listening test.

A family of four objective perceptual measures was proposed: Overall Perceptual Score (OPS), the Target-related Perceptual Score (TPS), the Interference-related Perceptual Score (IPS) and the Artifacts-related Perceptual Score (APS).

III. Previous Work

In [9], a comparison between perceptual ratings from a listening test and objective measures obtained with BSS_Eval is presented. Results show a significant correlation between SIR and the *Distortion* rating from the listening test. SDR also showed significant correlation with the *Intrusiveness* and *Separation* ratings from the listening test. In [10] BSS_Eval, PEASS and PESQ where compared to listening test results in the context of speaker separation in multisource reveberant environment. The study showed that none of the metrics were able to reliably predict human quality ratings. More recently, a comparison between perceptual ratings from listening tests and BSS_Eval metrics for singing voice separation was investigated in [11], where it was found that overall separation quality was very poorly correlated with these metrics, and the correlations observed were not statistically significant.

IV. Problem Description & Experiment

The use of objective and perceptual objective measures for quality evaluation has become standard in sound separation research in the last years. The considerably less time-consuming procedure of calculating a set of measures in comparison to conducting listening tests, is clearly an advantage of these evaluation methods. However, the effectiveness of such measures to truly represent perceptual ratings as the ones obtained with human listeners has been a matter of discussion in the separation community for some years now. The experiments described in this section evaluate the degree of correlation between existing objective measures for quality of separation and MOS ratings from human listeners in the context of HP separation: The quality of two separation algorithms, **alg1** [12] and **alg2** [13], was evaluated under a common **dataset**. **alg1** performs separation based on a phase expectation analysis that discriminates between harmonic and percussive components, while **alg2** is based on median filtering of spectrograms

across frames to emphasise harmonic components, and median filtering of spectrogram across frequency bins to emphasise percussive components. For the task at hand, the quality of the separated harmonic signals extracted with the two algorithms was evaluated using two sets of measures: **(i) BSS_Eval, (ii) PEASS**. Additionally, the quality of the separated signals extracted with the two algorithms was evaluated by human listeners through a **(iii) listening test** (see Section IV-C). To evaluate how much objective measures correlate with MOS perceptual ratings obtained with human listeners, the quality ratings obtained from **(i) BSS_Eval, (ii) PEASS**, and **(iii) listening tests**, were compared and a correlation analysis performed (see Section V).

A. Dataset & Algorithms

A dataset composed of 10 music signals was used for this experiment. Audio mixtures for all signals were obtained using the original multi-track recordings and were processed with the two HP algorithms. In this study, only the extracted harmonic signals were used in the evaluation. The dataset is available for download on the project website [1].

B. Objective Quality Evaluation: BSS_Eval & PEASS

The quality of the harmonic signals extracted with both algorithms were evaluated using BSS_Eval [8] objective measures and objective perceptual measures from PEASS v2 [14].

C. Subjective Quality Evaluation: Listening test

A listening test procedure was conducted to evaluate quality of the signals extracted with the two separation algorithms. A "Multi-stimulus test with hidden reference and anchor (MUSHRA)" was used following the standard described in [6]. To allow comparison with the ratings obtained with the objective measures, the listening test was divided into four sections, each evaluating one of the following criteria: (i) Interference, (ii) Target Distortion, (iii) Artifacts, and (iv) Overall Quality. Each part of the test consisted of a *training* and an *evaluation phase*. During each *training phase*, the participants were given the opportunity to familiarize themselves with the test content. During each *evaluation phase* the users were presented with the following signals: (1) signal obtained with **alg1**, (2) signal obtained with **alg2**, (3) the original signal (reference), and (4) an anchor signal. For each part of the listening test, special anchor signals were created, based on those used in [7]: For the (i) interference section, anchor signals were created by taking a weighted sum of the original harmonic signal from the multi-track recording and the original percussive signal (interference). For the (ii) target distortion section, anchor signals were created by filtering the original harmonic source with a low-pass filter with cutoff frequency of 3500 Hz. Additionally, 10% of the time-frequency bins (randomly selected) were set to zero. For the (iii) artifacts section, the anchor signal was created by adding an artifacts signal to the original harmonic signal. The artifacts signal was created by randomly selecting 1% of the time-frequency bins of the harmonic signal and setting the remaining 99% coefficients to zero. For the (iv) Overall Quality Evaluation, the anchor signal was created as a weighted sum of the original harmonic signal low-pass filtered (3.5KHz cutoff), an artifacts signal, and an interference signal. This choice of anchor signal was made considering that current state-of-the-art HP separation algorithms tend to produce signals with the three types of distortions.

A total of 16 subjects conducted the listening test. All the subjects had at least one year of experience in audio processing. The subjects were asked to rate the quality of the signals based on the four evaluation criteria. All ratings in the listening test were performed in a continuous scale from 0 to 100. Additional descriptive hints were given as follows: Bad (0 to 20), Poor (20 to 40), Fair (40 to 60), Good (60 to 80), and Excellent (80 to 100). A full description of the results for all the evaluation criteria as well as the perceptual ratings obtained in the listening tests are available on the project website[2].

V. Results

As stated in the introduction, the power of any objective quality measure directly depends on its capability to correlate with human listener MOS scores obtained through listening tests. The results presented in this section precisely attempt to assess such capability for both BSS_Eval and PEASS. A correlation analysis was conducted between the MOS results from the listening tests and the objective measures obtained with the described dataset. The MOS was obtained as the mean value of the scores obtained from the 16 participants of the listening test. For each of the four evaluated criteria, linear correlations were calculated using Pearson's linear correlation coefficient. Statistical significance was evaluated given a significance level of 0.05. The analysis was conducted for the two separation algorithms: (i) **alg1** and (ii) **alg2**. Results from the correlation analysis are presented in Table I. The correlation coefficient ρ and the p-value P obtained are presented in the table.

VI. Discussion

With respect to the results obtained via BSS_Eval, it can be observed in general that there is very little correlation between the MOS results obtained from the listening tests and the metrics calculated by BSS_Eval, and further, that what correlation is observed is not statistically significant, i.e., the observed p-values are larger than the significance level 0.05. The only exception to this is with the SIR results for **alg2**, where there is a negative correlation of -0.56 that has a p-value of 0.0889, which is still above the level of 0.05. Further, a very different correlation is observed for SIR with **alg1**, which suggests that SIR does not represent a globally useful metric for determining the perception of interference in separated signals, even if it offers some level of prediction for **alg2**. The overall lack of correlation was to be expected as it

[1] http://www.idmt.fraunhofer.de/en/business_units/smt/perceptual_quality_sound_separation.html

[2] http://www.idmt.fraunhofer.de/en/business_units/smt/perceptual_quality_sound_separation.html

TABLE I
THE CORRELATION COEFFICIENT ρ AND THE *p-Value* P ARE PRESENTED FOR THE FOUR CRITERIA: (I) OVERALL QUALITY, (II) ARTIFACTS DISTORTIONS, (III) INTERFERENCE FROM OTHER SOURCES, AND (IV) TARGET DISTORTIONS. PARAMETERS FOR BOTH BSS_EVAL AND PEASS MEASURES ARE PRESENTED.

		ρ	P	ρ	P
		SDR		**OPS**	
Overall	*Alg1*	0.2762	0.4398	0.2372	0.5094
	Alg2	-0.0905	0.8037	0.5847	0.0758
		SAR		**APS**	
Artifacts	*Alg1*	-0.1784	0.6220	-0.4361	0.2077
	Alg2	-0.082	0.8257	-0.1773	0.6242
		SIR		**IPS**	
Interference	*Alg1*	0.0321	0.9299	-0.3070	0.3880
	Alg2	-0.5648	0.0889	0.1293	0.7218
		ISR		**TPS**	
Target	*Alg1*	-0.0047	0.9898	-0.6326	0.0497
	Alg2	-0.1419	0.6959	0.4680	0.1725

had long been informally observed by researchers in the field that the BSS_Eval metrics were poor measures of perceptual separation quality. Indeed, it was this observation which led to the development of the PEASS metrics, which were designed to be more perceptually relevant.

The results obtained via PEASS do exhibit in general higher levels of correlation with the subjective listening scores than those obtained via BSS_Eval, but what is of interest is the fact that none of the correlations observed are very strong, with the largest (negative) correlation of -0.63 observed for TPS when tested on **alg1**. It is also interesting to note that this is the only correlation which is statistically significant with a p-value of less than 0.05. However, it should be noted that there is a positive and non-significant correlation of TPS with **alg2**. The large difference in correlation values obtained when using different algorithms to perform the separation should also be noted, which demonstrates that the performance of the PEASS metrics varies considerably depending on the algorithm used.

The correlation coefficients obtained stand in stark contrast to those presented in [14] describing the second version of PEASS, where a mean correlation value of 0.909 is presented. It is also interesting to note that the statistical significance of the correlations is not recorded in the associated paper which presents these correlation scores. The considerable difference in correlation results between our experiment and those used to develop PEASS suggests that the PEASS metrics do not generalize well to algorithms and/or test material outside those used in training PEASS. The results obtained clearly indicate the fact that PEASS is not useful as a means of predicting the performance of harmonic-percussive algorithms, and suggest that it may not generalize well in other settings, given the disparities in prediction performance between the two algorithms tested.

VII. CONCLUSIONS

This paper has focused on analyzing experiments related to sound source separation with regards to identifying if separation quality can be systematically and truthfully determined using existing evaluation metrics. To this end, a set of listening tests were conducted on the outputs of a number of sound source separation algorithms on a test-set of recordings. In particular, two harmonic-percussive separation algorithms were chosen as exemplars. The outcomes of the tests were then used to obtain mean opinion scores for overall quality, artifact distortions, interference from other sources and target distortions. These scores were then taken as ground truth measures of separation quality.

Also discussed in the paper were existing metrics commonly used to attempt to measure separation quality, namely BSS_Eval, a quantitative set of energy-based metrics and PEASS, a perceptually motivated set of quantitative measures. The motivations behind these metrics were discussed and both sets of metrics were calculated for the outputs of the harmonic-percussive separation algorithms on the test set. The resultant metric scores were then correlated against the MOS scores from the listening test to determine if these metrics were suitable for use in predicting sound source separation quality.

The correlations showed that the scores obtained via BSS_Eval and PEASS were not indicative of the MOS scores obtained via the listening tests, and that the performance of the metrics varied greatly depending on what algorithm was used. This is in contrast to the high correlations obtained with the material used to train PEASS, and in conjunction with the large algorithm-dependent difference in correlations, suggests that PEASS is not useful for determining the quality of harmonic-percussive separation algorithms, and further that PEASS does not generalize well.

The test results and correlation computations shown in this paper suggest that existing metrics such as BSS_Eval and PEASS might not be suitable for determining perceptual sound quality on source separation tasks. While PEASS may work very well on the kinds of algorithms it was designed with, it does not generalize well to other types of algorithms as the ones used in this work. While the use of such measures may be justified as a means of speeding up algorithm development, for the moment, only listening tests can really tell whether one source separation algorithm works better than another. For the future, not all is lost: just like PEAQ was developed to mimic the way we listen to the results of certain audio

codecs, a variation on PEASS could be developed which shows better correlation with listening tests. However, this is not easy: first, the class of algorithms to be tested needs to be clearly defined; second, large amounts of content data and perceptual scores are needed for reliable results. Finally, a great deal of work is needed to correlate the new measures with the listening tests. As an initial step towards this, we are making the test set and MOS scores for the tests in the paper available. We also encourage other researchers to do similarly when conducting listening tests in an effort to generate a sufficiently large pool of test material and MOS scores to allow design of improved metrics. Even then, it is clear that whenever researchers come up with completely new ideas for separation algorithms, this tweaking of the measurement algorithms to the way we listen will probably need to be repeated. At the very least we will need new listening tests to verify the measurement algorithms whenever there are new classes of source separation algorithms.

REFERENCES

[1] K. Brandenburg, "Ocf–a new coding algorithm for high quality sound signals," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, vol. 12, Apr 1987, pp. 141–144.

[2] K. Brandenburg and T. Sporer, "Nmr and masking flag: Evaluation of quality using perceptual criteria," in *Audio Engineering Society Conference: 11th International Conference: Test and Measurement*, May 1992.

[3] T. Thiede and et al., "Peaq - the itu standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc*, vol. 48, no. 1/2, pp. 3–29, 2000.

[4] ITU, "RECOMMENDATION ITU-R BS . 1387-1 Method for Objective Measurements of Perceived Audio Quality," Tech. Rep., 2001.

[5] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary Guidelines for Subjective Evaluation of Audio Source Separation Algorithms," in *ICA Research Network International Workshop*, Liverpool, UK, 2006, pp. 93–96.

[6] ITU, "RECOMMENDATION ITU-R BS . 1534-1 Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems," Tech. Rep., 2003.

[7] 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.

[8] C. Févotte, R. Gribonval, and E. Vincent, "BSS_Eval Toolbox User Guide Revision 2.0," IRISA, Rennes, Bretagne Atlantique, Tech. Rep., 2011.

[9] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of subjective and objective evaluation methods for audio source separation," in *Meetings on Acoustics*, vol. 123, no. 5, Paris, France, 2008, p. 3569.

[10] P. Langjahr and P. Mowlaee, "Objective Quality Assessment of Target Speaker Separation Performance in Multisource Reverberant Environment," in *4th International Workshop on Perceptual Quality of Systems PQS*, Vienna, Austria, 2013, pp. 89–94.

[11] 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 2015)*, 2015.

[12] E. Cano, M. Plumbley, and C. Dittmar, "Phase-based harmonic/percussive separation," in *15th Annual Conference of the International Speech Communication Association (2014). Interspeech.*, 2014.

[13] D. FitzGerald, A. Liutkus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/percussive separation using kernel additive modelling," in *Proceedings of the Irish Signals and Systems Conference*, 2014.

[14] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *10th International Conference on Latent Variable Analysis and Signal Separation*, Tel-Aviv, Israel, 2012, pp. 430–437.