THE DIMENSIONS OF PERCEPTUAL QUALITY OF SOUND SOURCE SEPARATION

Estefanía Cano[†], *Judith Liebetrau*[†], *Derry Fitzgerald* [‡], *Karlheinz Brandenburg*[†]

[†]Fraunhofer IDMT, Germany
[‡] Cork Institute of Technology, Ireland

ABSTRACT

Quality of sound source separation algorithms has traditionally been evaluated based on a set of established quality metrics: target distortion, interference from other sources, artifacts distortion, and a measure of overall separation quality. In our previous work, listening test results were presented where no significant correlation between these quality metrics and perceptual ratings from the listening test could be observed. Following these results, we now attempt to better understand perceptual quality in a sound source separation context. We focus on determining how separation quality is actually defined by listeners and propose the use of a descriptive methodology to reveal its most relevant dimensions. A combination of Free-Choice Profiling and Repertory Grid Technique is used with 10 human listeners in an attempt to verify if the main dimensions of separation quality truly correspond to those established by the quality metrics. The outcomes of this exploration bring light to the development of new methodologies for sound separation quality evaluation, and suggest a two-dimensional perceptual space for quality of sound source separation.

Index Terms— Sound Source Separation, Quality Perception, Quality Metrics, Listening Tests, Repertory Grid

1. INTRODUCTION

The first established procedure for sound source separation (SSS) quality evaluation was proposed in [1]. In this work, a set of performance metrics known in the literature as BSS, are obtained by decomposing the separated source into three types of distortions: interference from unwanted sources, additive noise, and algorithm artifacts. Based on this decomposition, energy ratios between the signal components are used to define a global quality measure and three quality measures related to the error terms; namely, Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR), Source to Noise Ratio (SNR), and Source to Artifacts Ratio (SAR), respectively. These metrics are expressed in dBs.

In an attempt to incorporate perceptual information into the evaluation process, the PEASS Toolkit –Perceptual Evaluation Methods for Audio Source Separation– was later proposed in [2]. In this work, three types of signal distortions were considered: target, interference, and artifact distortions. Perceptual ratings obtained via listening tests were then mapped into objective scores calculated using the PEMO-Q auditory model by means of a non-linear function. Similarly to BSS, four quality measures were proposed; namely, Overall Perceptual Score (OPS), Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), and Artifacts-related Perceptual Score (APS).

While the availability of quality measures such as BSS and PEASS reduced the efforts of performing listening tests to the calculation of quality metrics, the separation community soon realized that numerical results obtained with the metrics did not necessarily match perceptual quality ratings from human listeners. In an attempt to better understand this matter, we presented in [3] the results from a series of listening tests that quantified the quality perception of several SSS algorithms. In the listening tests, besides the overall quality, the interference, artifact and target distortion were assessed by applying a multi-stimulus comparison according to ITU-R BS.1534-3 [4]. A correlation analysis was then performed, showing that the scores obtained via BSS and PEASS were not indicative of the scores obtained via the listening tests. These results indicated that existing metrics do not generalize well to all separation algorithms and hence, might not be suitable for SSS quality evaluation. Similar results have been observed in the context of speaker separation in multi-source reverberant environment [5], and singing voice separation [6].

It has now become clear in the SSS community that the development of robust quality metrics and quality evaluation procedures is of great importance. Some initial explorations into the development of alternative procedures have been presented in [7, 8]. In this paper, we investigate the perceptual dimensions of SSS quality by looking into sound perception as a multi-dimensional problem that includes several individual attributes [9]. We aim at the identification and determination of these individual attributes and at a generalization of these attributes in the context of SSS.

2. DESCRIPTIVE ATTRIBUTE GENERATION

There are several methods for generating descriptive attributes (see [10] for a comprehensive overview). In general, two types of methods are distinguishable: consensus and in-

dividual vocabulary techniques. With consensus vocabulary, a group of assessors (participants) develops a descriptive language and rating scales which are used in grading sessions. The consensus vocabulary procedure requires training of the panel members (expert listeners), as well as group discussions and consensus. In contrast, individual vocabulary techniques enable assessors to develop their own vocabulary. No training or group discussion is required, and it is applicable to experts and non-expert listeners alike. However, the analysis and interpretation of the individual vocabulary and its transformation to a common perceptual description is more complex compared to the consensus vocabulary techniques [11, p.160].

Methods like *Free-Choice Profiling (FCP)* or *Repertory Grid Technique (RGT)* are examples of individual vocabulary techniques. FCP was originally developed in food research, and was later adopted for multi-modal assessment [12]. First, the assessors define their own attributes to describe their perception. The assessors then evaluate all stimuli with their own adjective scales which are labeled with the attributes developed by each assesor. Due to the usage of individual vocabulary for each assessor a common perceptual space has to be calculated. RGT was originally proposed for analysis of personal constructs in psychotherapy [13]. Assessors compare triads of stimuli and describe the way two of the stimuli are similar and different from the third one (triadic elicitation procedure). As long as the assessor is able to give new descriptions for commonalities and differences of stimuli, this process continues. Based on the collected descriptors, an individual grid is constructed out of opposing terms and used in the grading phase. In recent years, this method has been applied to audio assessment (e.g., [14]). A drawback of the triadic elicitation procedure is that differences between two sound stimuli might be obscured if they are always presented with a more dissimilar sound. In the attempt to reduce the effects of construct masking, a modification of RGT with pairwise presentation of stimuli was proposed in [15]. Despite of the benefits in using individual vocabulary, assessors sometimes have difficulties describing their perception in the FCP procedures. In [16], a structured free-choice procedure, based on the philosophy of the RGT was proposed to help the assessors concentrate and focus on the development of descriptive attributes. We use this technique for the analysis conducted in this work.

3. METHODOLOGY

Building on our previous work, we used a selection of three tracks (items) from the data set used in [3]. A total of four harmonic-percussive (HP) algorithms were considered: Alg1 [17] and Alg2 [18], as used in our previous work, and Alg3 [19] and Alg4 [20] which were included for the attribute elicitation stage to obtain descriptors general to a diversity of separation algorithms. Two separate stimuli sets were constructed using the three chosen items: one set for the attribute elicitation (AESet), and one set for the grading procedure (GPSet). The AESet included the original mix from the multi-track recordings (also referred to as *mix* in the remainder of the paper), the original harmonic track of the multi-track recording (also referred to as *harm* in the remainder of the paper), and the harmonic signals extracted with the four HP algorithms. The GPSet consists of the harmonic signals extracted with Alg1 and Alg2 (to allow direct comparison with our previous work), and the harmonic track of the multi-track recordings.

Following recommendations in [21], ten people (three female, seven male), all employees or students at Fraunhofer IDMT, participated in the test. Most had a musical background, while only two had previous knowledge concerning SSS and quality metrics.

The test procedure consisted of two parts, conducted back-to-back: (1) vocabulary generation and (2) grading. A graphical user interface (GUI) was implemented for all assessment tasks. Although the GUI guided the assessors through the evaluation, a test supervisor was always present to assist the participants. For the vocabulary generation, the assessors were first presented with the original mixes of the three selected items. Each assessor was asked to choose his/her favorite one, which was then used in the remainder of the vocabulary generation stage. The assessors were then presented with the *harm* version of the chosen item, and the four harmonic signals obtained with the four separation algorithms (presented in random order). Given that HP separation algorithms aim at producing *harm* as its outcome (target source), *harm* was always highlighted as the reference. The assessors were then asked to do pair-wise comparisons between the reference and each of the four separated versions (in any order). For each comparison, the assessors were asked to characterize the perception of differences and similarities between the two tracks. They were encouraged to use a noun or adjective to characterize their perception but were also allowed to use phrases. The comparisons were conducted until each assessor produced a minimum of six descriptors; however, freedom was given to produce more than six.

In the grading stage, the goal was to rate the GPSet with the attributes developed in the first stage. Each test stimulus (the harmonic signals obtained with Alg1 and Alg2 for the three test items and their original *harm* versions) was played back to the assessors, who then rated them in a 100-point scale with respect to each of the generated descriptors. The participants could listen each test stimulus as often as required.

4. RESULTS OF THE PERCEPTUAL EVALUATION

A total of 93 individual attributes were developed, varying between six and 12 attributes per assessor. The multi-dimensional rating spaces of each assessor were mapped to a common perceptual space using a Multiple Factor Analysis (MFA) [22]. Each dimension (Dim) in the common per-

Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8
38.7 %	16 %	13.6 %	10.2 %	6.8 %	5.8 %	4.7 %	4.2 %

Table 1. MFA: Explained variance per dimension (Dim)

ceptual space explains a certain amount of the variance in the data. In this case, an 8-dimensional common perceptual space was found as shown in Table 1. The significance of the MFA model was investigated by comparing the explained variance of the GPSet (using the first three dimensions), to the explained variance of random data (derived through Monte-Carlo simulation and permutation of the original data with 1000 repetitions) [23]. The explained variance of the investigated data set (68.3 %) is higher than the 95th percentile from the variance distribution of the simulated data sets (67.7 %).

To identify the number of meaningful dimensions in the common perceptual space, four methods, were applied; namely, (1) Acceleration Factor, (2) Optimal Coordinate, (3) Parallel Analysis, and the (4) Kaiser Criterion [24, 25]. As it is often the case, no common solution was found between methods: Acceleration Factor resulted in a 1-D space, Optimal Coordinate and Parallel Analysis suggested 3 dimensions, while the Kaiser criterion resulted in 6 dimensions.

Given that the number of retained dimensions cannot only rely on these indices, the interpretability of the dimensions must also be considered. To interpret the meaning of the dimensions, only well-projected attributes with $\cos^2 \geq 0.5$ and high correlation ($r \geq |0.7|$) were taken into account [26]. For Dim 1, 32 attributes fulfilled these criteria, for Dim 2 seven attributes, for Dim 3 three attributes, for Dim 4 one attribute, and for Dim 5 two attributes. The retained attributes suggest three categories of descriptors: a) describing general audio quality (e.g., pleasing, unpleasing, annoying), b) describing the degree of interference (e.g., filled up, reduced, immersive), and c) describing distortions (e.g., distorted instruments, disturbing because no continuous loudness, bad coding, metallic and unnatural). However, based on the attributes themselves, a unique descriptor that represents each dimension could not be defined.

To further analyze the descriptors, we calculated and visualized how the presented stimuli (GPSet) are placed in the perceptual space, specifically in 2-D representations of the perceptual space. As a general reference, stimuli placed closer together are perceived as similar. Figure 1 displays the position of the stimuli in three subspaces: Dim 1 - Dim 2, Dim 2 - Dim 3, and Dim 3 - Dim 4. We used different colors and symbols to highlight the affiliation of the stimuli to either a condition (Alg1 $\diamond$, Alg2 $\triangle$, harm $\circ$), or to a test item (i1, i2, i3).

In the Dim 1 - Dim 2 subspace (Fig.1 left), it is clear that the three stimuli processed with Alg1 are perceived as similar, as well as the items processed with Alg2, and the original *harm*. A clear distinction between the algorithms is visible regardless of the type of audio content (item): Alg1 is placed more to the left side of Dim 1, Alg2 in the middle, and *harm* to the right of the axis. There is no overlap between the stimuli groups along Dim 1, therefore it is assumed that these groups are perceived very differently with respect to this dimension. A slight overlap of Alg1 and *harm* can be observed along Dim 2, which indicates a similar perception of these groups with respect to this dimension.

The same analysis was performed for the Dim 2 - Dim 3, Dim 3 - Dim 4, and Dim 4 - Dim 5 subspaces. For the Dim 2 - Dim 3 subspace, it appears that both the algorithms and the item type play an important role. In Fig.1- center, no clear separation between the algorithms can be seen with respect to Dim 3. For the Dim 3 - Dim 4 subspace, no clear distinction between the algorithms could be observed; however, the item type seems to play an important role in this subspace. As can be seen Fig.1- right, stimuli derived from item 1 (Alg1, Alg2 and *harm*) are placed in the second quadrant, stimuli belonging to item 2 are placed in the third quadrant, whereas those associated with item 3 are placed in the first and fourth quadrants. This indicates that a characterization of the item type and not algorithm quality differences is obtained here. The Dim 4 - Dim 5 subspace was investigated in the same manner, leading to similar results as with Dim 3 - Dim 4. These observations lead to the conclusion that only Dim 1 and Dim 2 describe the quality perception caused by properties of the algorithms, while higher dimensions seem to characterize differences between test items.

After defining the perceptual space as a two-dimensional one, the interpretation of these dimensions needs to be further explored. The starting point of our investigation were listening tests in which besides the overall quality, the interference, artifact and target distortions were assessed by applying a multi-stimulus comparison according to ITU-R BS.1534-3 [3]. In theory, target, interferences, and artifacts are independent dimensions of quality in a sound source separation context. Consequently, the mean opinion scores (MOS) for each stimulus represent their spatial distribution along the respective dimension. This allows us to perform a correlation analysis between the MOS in [3], and the first two dimensions of the common perceptual space derived in this work. In Table 2 the Pearson correlation coefficient r is displayed.

For Dim 1 a high correlation value can be observed with the MOS of the interference. Dim 1 is only moderately correlated to overall quality (but not significant), and no strong correlations can be observed with artifacts and target. Dim 2 has a strong linear correlation with overall quality but an even higher r value for target distortions and artifacts distortions. These results suggest that Dim 1 in the common perceptual space describes the degree of separation or the amount of interference from other sources. These results also indicate that Dim 2 encompasses all types of distortions, resulting in a strong relation to the overall audio quality. Based on this interpretation, Figure 1-left could be analyzed in more detail. The *harm* stimuli have a high overall quality and low of interference from other sources. Alg1 seems to result in good

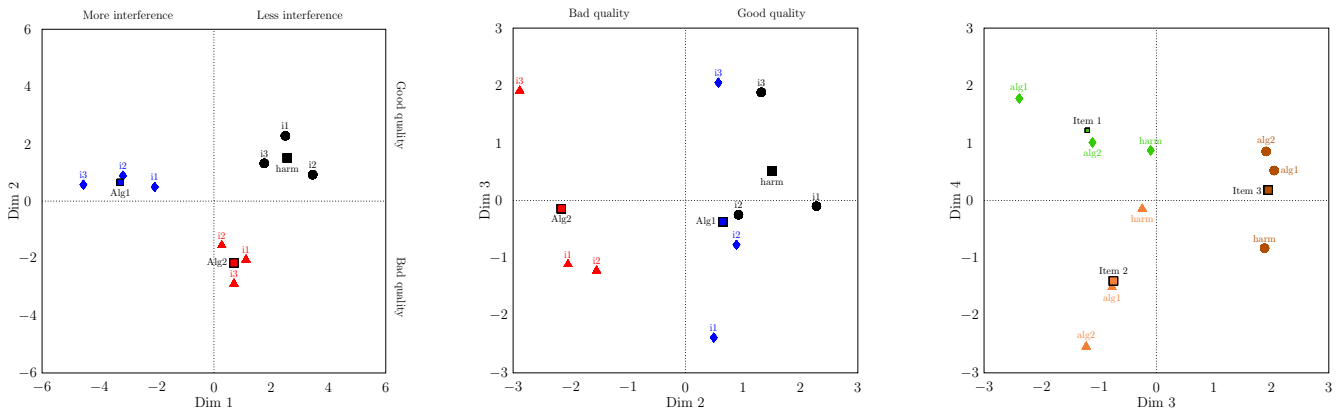

Fig. 1. Positioning of the stimuli in the Dim 1 - Dim 2 (left), Dim 2 - Dim 3 (center), and Dim 3 - Dim 4 (right) subspaces. In the left and center plots the following marker convention is used: ◇ Alg1, △ Alg2, and ○ harm. Item numbers are shown as i1, i2, and i3. In the plot in the right, the markers are used to indicate items as follows: ◇ item 1, △ item 2, and ○ item3, with the algorithms and *harm* displayed next to each marker. The mean value of each group is displayed with a rectangle.

overall quality but with higher amount of interference from other sources. Alg2 seems to result in lower overall quality but in less interference from other sources compared to Alg1.

	Dim 1	Dim 2
Overall	$r = 0.58, p = 0.1028$	$r = 0.77, p = 0.015$
Artifact	$r = 0.21, p = 0.5893$	$r = 0.93, p = 0.0003$
Interference	$r = 0.83, p = 0.0061$	$r = 0.34, p = 0.3775$
Target	$r = 0.21, p = 0.5818$	$r = 0.91, p = 0.0005$

Table 2. Pearson correlation coefficient r and p values between the MOS for overall quality, interference, artifact and target distortion from [3], and the spatial coordinates of Dim 1 and Dim 2 of the common perceptual space.

5. CONCLUSIONS

This study focused on how human listeners define SSS quality. Using a descriptive methodology combining FCP and RGT, the study showed that participants were successful in labelling and rating their perceptions of separation quality, allowing the creation of a common perceptual space which had high explained variances. The two most important dimensions relate to perceived separation quality, accounting for the majority of the variance in the space. The other dimensions appeared related to differences between the test items.

The first dimension was strongly correlated with the MOS for interference. The second dimension correlated strongly with the overall MOS score and was even more correlated with both the artifact and target MOS scores. The key point is that, at least with respect to the HP algorithms tested, interference from other sources was the dominant perceptual attribute for good quality separation, while both artifact and target distortions (and to a lesser extent, overall quality) were grouped together under one perceptual feature. This suggests that four quality metrics (as traditionally done in SSS) are not necessary, and that a properly defined set of two distortions is sufficient to capture the variance related to separation quality.

While these results highlight problems with existing procedures for assessing SSS quality, with respect to both existing metrics and distortion-based listening tests, there is a need for further study and research on this topic. In particular, these results were obtained using HP algorithms only, and there is a need to generalize this to other separation tasks. This will require larger scale tests, with increased numbers of participants and a larger choice of test items in the grading stage of the tests. It will also require MOS scores for distortions for these test items. Future work will therefore focus on extending the range and validity of the results obtained herein.

6. REFERENCES

[1] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[2] 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, Sept. 2011.

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

[4] ITU-R, "Recommendation BS.1534-3: Method for the

subjective assessment of indermediate quality levels of coding systems," 10/2015.

[5] 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.

[6] 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.

[7] 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 *13th International Conference on Latent Variable Analysis and Signal Separation*, Grenoble, France, 2017.

[8] M. Cartwright, B. Pardo, G. J. Mysore, and M. Hoffman, "Fast and easy crowdsourced perceptual audio evaluation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 619–623.

[9] J. Blauert and U. Jekosch, "Sound-quality evaluation – a multi-layered problem," *Acta Acustica united with Acustica*, vol. 83, no. 5, pp. 747–753, 1997.

[10] H.T. Lawless and H. Heymann, Eds., *Sensory Evaluation of Food*, Food Science Text Series. Springer New York, 2010.

[11] P. Varela and G. Ares, Eds., *Novel techniques in sensory characterization and consumer profiling*, CRC Press Inc, 2014.

[12] J.M Murray, C.M Delahunty, and I.A Baxter, "Descriptive sensory analysis: past, present and future," *Food Research International*, vol. 34, no. 6, pp. 461–471, 2001.

[13] G. Kelly, *The Psychology of Personal Constructs: Volume Two: Clinical Diagnosis and Psychotherapy*, Routledge, 2003.

[14] J. Berg and F. Rumsey, "Identification of perceived spatial attributes of recordings by repertory grid technique and other methods," in *Audio Engineering Society Convention 106*, 1999.

[15] S. Choisel and F. Wickelmaier, "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," *Journal of the Audio Engineering Society*, vol. 54, no. 9, pp. 815–826, 2006.

[16] J.A. McEwan, J.S. Colwill, and Thomson, D. M. H., "The application of two free-choise profile methods to investigate the sensory characteristics of chocolate," *Journal of Sensory Studies*, vol. 3, no. 4, pp. 271–286, 1989.

[17] 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.

[18] 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.

[19] D. FitzGerald, "Harmonic/percusssive separation using median filtering," in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx)*, Graz and Austria, 2010.

[20] H. Tachibana, N. Ono, H. Kameoka, and S. Sagayama, "Harmonic/percussive sound separation based on anisotropic smoothness of spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2059–2073, 2014.

[21] Bonnie M. King, Paul Arents, and Nathalie Moreau, "Cost/efficiency evaluation of descriptive analysis panels i. panel size," *Food Quality and Preference*, vol. 6, no. 4, pp. 245 – 261, 1995, Second Sensometrics Meeting.

[22] B. Escofier and J. Pagès, "Multiple factor analysis (afmult package)," *Computational Statistics & Data Analysis*, vol. 18, no. 1, pp. 121–140, 1994.

[23] I. Wakeling, M. M. Raats, and H.J.H. MacFie, "A new significance test for consensus in generalized procrusteres analysis," *Journal of Sensory Studies*, vol. 7, no. 2, pp. 91–96, 1992.

[24] A.B. Costello and J.W. Osborne, "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis," *Practical Assessment, Research & Evaluation*, vol. 10, no. 7, pp. 1–9, 2005.

[25] B. Thompson and L.G. Daniel, "Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines," *Educational and Psychological Measurement*, vol. 56, no. 2, pp. 197–208, 1996.

[26] H. Abdi and L.J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010.