HAL open science

Multi-criteria subjective and objective evaluation of audio source separation

Valentin Emiya, Emmanuel Vincent, Niklas Harlander, Volker Hohmann

To cite this version:

Valentin Emiya, Emmanuel Vincent, Niklas Harlander, Volker Hohmann. Multi-criteria subjective and objective evaluation of audio source separation. AES 38th International Conference on Sound Quality Evaluation, Jun 2010, Pitea, Sweden. inria-00545031

HAL Id: inria-00545031

https://inria.hal.science/inria-00545031

Submitted on 31 Jan 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multi-criteria subjective and objective evaluation of audio source separation

Valentin Emiya[1], Emmanuel Vincent[1], Niklas Harlander[2], and Volker Hohmann[2]

[1]*INRIA, Centre Inria Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France.*

[2]*Medizinische Physik, Carl von Ossietzky-Universität Oldenburg, Oldenburg, Germany.*

Correspondence should be addressed to Valentin Emiya (`firstname.lastname@inria.fr`)

ABSTRACT

In this paper, we address the problem of assessing the perceived quality of estimated source signals in the context of audio source separation. These signals involve different kinds of distortions depending on the considered separation algorithm, including distortion of the target source, interference from other sources or musical noise artifacts. A new MUSHRA-based subjective test protocol is proposed to assess the perceived quality with respect to each kind of distortion and collect the scores of 20 subjects over 80 sounds. Subsequently, the contribution of each type of distortion to the overall quality is analyzed. We propose a family of objective measures aiming to predict the subjective scores based on a decomposition of the estimation error into several distortion components. We conclude by discussing possible implications of this work in the field of 3D audio quality assessment.

1. INTRODUCTION

Audio source separation is the task of extracting the signal of a given source from a mixture signal involving concurrent sound sources. This is a core task of audio processing, with applications ranging from source extraction to content description and manipulation. In this paper, we focus on the range of applications where the separated sources are meant to be listened to. Such applications include for instance speech enhancement in noisy or multi-talker scenarios for hearing aids or phone devices, restoration of old recordings, and music de-soloing for automatic accompaniment or karaoke.

A variety of separation algorithms have been introduced in the last twenty years, based on either perceptually motivated or statistical models [1, 2], yet none has achieved perfect separation to date. Even the best algorithms result in heavy distortion compared to that observed in other fields such as audio coding or rendering. It is generally acknowledged that three kinds of distortion can be perceived together or alone depending on the algorithm [3]: distortion of the target source, interference from other sources, and musical noise or other artifacts introduced by the separation process. A multi-criteria approach is thus necessary to provide fine characterization of the pros and cons of each algorithm.

Few studies targeted to subjective or objective quality assessment of source separation have been performed so far. Most subjective studies aim to evaluate a single criterion: overall quality [4, 5, 6], preference [7, p. 138] or musical noise audibility [8]. Two multi-criteria studies dedicated to speech data have also been conducted, using either the standard ITU criteria for the evaluation of speech denoising algorithms [9], namely speech distortion, background noise intrusiveness and overall quality [10], or a different set of criteria called intelligibility, fidelity and suppression [11, p. 95]. With the exception of [5, 9], the above studies do not rely on general-purpose ITU standards for quality testing and are either not reproducible due to the lack of a precise test protocol [4, 8, 11] or inaccurate due to the use paired comparison tests designed for small degradations [6]. The validity of the resulting scores is also limited by the use of a small set of algorithms generating certain kinds of distortion only, *e.g.* Independent Component Analysis (ICA) in [7], time-frequency masking in [4] or simulated separation in [9], and a limited set of sounds categories, *e.g.* speech in [5] or isolated notes from an alto saxophone in [6]. A standardized multi-criteria test protocol applicable to any category of sounds would hence be highly desirable.

A few more studies have been made towards objective

evaluation (see [3] for a review). The energy ratio criteria in [3], termed Signal-to-Distortion ratio (SDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artifacts Ratio (SAR), are widely used and have been employed within evaluation campaigns [12]. Derived criteria seeking to predict the overall perceived quality by linear or nonlinear combination of these baseline criteria have also been proposed in [6, 9]. Nevertheless, none of these criteria takes auditory phenomena such as loudness weighting and spectral masking into account, such that their correlation to subjective quality remains limited.

In this paper, we propose a new subjective test protocol to address the multi-criteria evaluation of audio source separation, which we hope can serve as the basis for discussion towards a future standardized protocol. As opposed to the three criteria used for speech enhancement [10], we introduce a set of four criteria that are suitable for the source separation task. In addition, we collect subjective scores from 20 subjects over a large range of sounds obtained by several state-of-the-art source separation algorithms in various mixing configurations. We use these scores to train a family of objective measures with improved correlation with subjective ratings.

The structure of the rest of the paper is as follows. We present the multi-criteria subjective test protocol in Section 2 and analyze the resulting scores in Section 3. We then summarize the principles of the proposed objective measures in Section 4 and conclude by discussing possible implications of this work in the field of 3D audio quality assessment in Section 5.

2. MULTI-CRITERIA SUBJECTIVE TEST PROTOCOL

2.1. Instructions

As stated in the introduction, three specific kinds of impairment are generally distinguished in the field of source separation: distortion of the target source, interference and artifacts. The two latter terms are unclear for naive listeners and carry a distinct meaning in other fields. Assuming that the clean target source signal is available as a reference, we propose to rate the quality of an estimated source signal according to four subjective criteria, as specified by the following less ambiguous instructions:

(Q_1) rate the *global quality* compared to the reference for each test signal;

(Q_2) rate the quality in terms of *preservation of the target source* in each test signal;

(Q_3) rate the quality in terms of *suppression of other sources* in each test signal;

(Q_4) rate the quality in terms of *absence of additional artificial noise* in each test signal.

2.2. Protocol

The proposed test consists of four parts, each associated with one instruction, with breaks in between. The order of the instructions is fixed to the above, since informal preliminary tests suggested that global rating was harder to achieve after specific ratings had been given. Because of the medium to large impairments in the test material, we follow the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) protocol [13] for each part, which provides small confidence intervals with a reasonable number of subjects. This protocol involves a training phase, where all sounds are presented at the same time, and a rating phase, where the subjects undergo successive trials consisting of rating the quality of all sounds associated with a given reference on a scale from 0 to 100.

We here associate each trial with one sound mixture and one target source within that mixture. Several test items are presented for rating, including estimates of the target source produced by actual source separation algorithms, the reference clean target source and some anchor sounds. The loudness of the reference signals is assumed to be fixed for all trials. Other test items may be normalized to the same loudness or kept unchanged, depending whether erroneous scaling is considered as a distortion or not [3]. Subjects can listen to these sounds as many times as needed, as well as to the reference and the mixture. The trials and the test items are presented in random order to each subject.

2.3. Anchor sounds

We advocate the use of three anchor sounds designed to fit the three kinds of impairment, as specified by instructions Q_2 to Q_4. The first anchor (A_2) aims at reproducing the impairments related to the distortion of the target source, which typically includes the rejection of certain frequencies or time intervals. It is obtained by low-pass filtering the target source signal using a 3.5 kHz cut-off frequency and by randomly zeroing out 20% of the re-

maining time-frequency coefficients[1]. The second anchor (A_3) aims at reproducing the impairments related to the presence of concurrent sources and is defined as the sum of the target signal and an interference signal. This interference signal is created by summing all the other sources of the considered mixture and setting the loudness of their sum to that of the target. The third and last anchor (A_4) aims at reproducing the impairments related to the presence of additional artificial noise and is defined as the sum of the target signal and a musical noise signal. This musical noise signal is created by randomly zeroing out 99% of the time-frequency coefficients of the target and setting the loudness of the resulting signal to that of the target.

3. SUBJECTIVE TEST RESULTS AND STATISTICAL ANALYSIS

3.1. Test material and subjects

We applied the proposed test protocol to the results of various state-of-the-art source separation algorithms submitted to the 2008 Signal Separation Evaluation Campaign (SiSEC) [12] over 5 speech mixtures and 5 music mixtures. Each of the 10 trials involved the results of 4 actual algorithms, as well as the hidden reference and the 3 anchor sounds. Different algorithms were chosen for each trial. All sounds had a duration of 5 s. The sound material was hence composed of 80 sounds in total and covers a wide range of sounds encountered in practical source separation scenarios in terms of source categories (male and female speech, singing voice, pitched musical instruments and drums), of number of sources (two to ten) and of mixing techniques (panning, convolution with room impulse responses, professional software mixing or microphone array recording). Each reference was scaled to a fixed loudness using a Matlab toolbox[2] based on the ISO 532B standard [14].

The test was performed by 23 normal hearing subjects with general expertise in audio processing. Among these subjects, 13 were located in Rennes, France, and 10 in Oldenburg, Germany. All speech mixtures were in a foreign language for all participants in order to prevent any bias due to intelligibility issues. The guidelines were presented in a unique form written in English and the interface was implemented via a variant of the MUSHRAM toolbox[3].

3.2. Post-screening of subjects

A post-screening analysis was applied in order to detect and remove the subjects that did not perform consistent quality assessment. However, in the context of a multi-criteria evaluation involving multiple simultaneous kinds of impairment, subjects may have different rating strategies resulting in score disagreement. While such disagreements may happen with test sounds from actual source separation algorithms, they should not arise for the hidden references, which are supposed to be ranked as perfect, or for the anchor sounds, which involve a single kind of impairment. Consequently, post-screening analysis was performed on the latter subset of data only over which a consensus is expected. Among the 23 subjects, 3 were detected as outliers using the Mahalanobis distance [15] and a threshold set to the 0.975 quantile of the theoretical χ^2 distribution. As a consequence, only the remaining 20 subjects will be taken into account in the rest of this paper.

3.3. Effects of location

The statistical significance of the effects of subject's location (Oldenburg *vs.* Rennes) was examined by means of an Analysis of Variance (ANOVA) using SPSS Statistics 12.0[4]. Location was considered as the between factor, while instructions (4) and mixtures (10) were considered as within factors. The level of significance was set to $\alpha = 0.05$. The effects of instructions ($\eta^2 = 0.837$) and mixtures ($\eta^2 = 0.567$) are highly significant (all $p < 0.05$, corrected F-values from 92.3 to 23.6) but the effect of location is not significant with no effect size ($F(1,18) < 1$, $p = 0.597$, $\eta^2 = 0.01$). The minor strength of this effect compared to that of the other within variables indicates that location does not have a significant influence on the subject ratings. Also, most of the within interactions are significant ($p < 0.05$) as long as they are not combined with the between factor location.

3.4. General trends and consistency

As a preliminary statistical analysis, the means and 95%-confidence intervals of the subjective scores of the hidden reference and anchor sounds are presented in Fig. 1. This validates some expected trends such as:

- (almost) perfect score for the hidden reference;

[1] A short time Fourier transform with a 46ms-Hann window was used to generate this anchor sound, as well as the third one.

[2] http://www.auditory.org/mhonarc/2000/zip00001.zip

[3] http://www.elec.qmul.ac.uk/digitalmusic/downloads/#mushram

[4] http://www.spss.com/software/statistics/advanced-statistics/

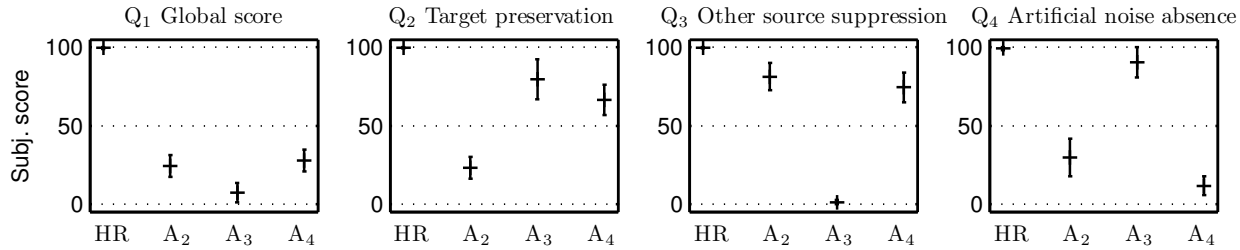

Fig. 1: Mean and 95%-confidence interval over all subjects and all trials of the subjective scores of the hidden reference (HR), the distorted target anchor (A_2), the interference anchor (A_3) and the artifact anchor (A_4). Each sub-figure corresponds to one instruction.

- (almost) null confidence interval for the hidden reference;
- low scores for the anchor sounds;
- consensus between the subjects over anchor sounds, with confidence intervals from ± 1.4 to ± 12.6;
- low score for the anchor sound A_i related to the instruction Q_i, $2 \leq i \leq 4$;
- high score for the two anchor sounds A_i related to the other specific instructions Q_j, $2 \leq i, j \leq 4$, $i \neq j$, with the exception of A_2 with instruction Q_4.

This exception means that the subjects reckoned that the distorted target anchors (A_2) were corrupted by artificial noise. This suggests that strong distortions of the target source may sound as artificial noise and be no more perceived as related to the target.

As illustrated in Fig. 2, the confidence intervals over the subjective scores of estimated sources from actual source separation algorithms are all narrower than ± 15. Narrower confidence intervals are obtained for instructions Q_1 and Q_3 than for Q_2 and Q_4, which suggests that the global quality and the level of interference may be easier to rate than the level of target distortion and artifacts. Again, this may be due to strong distortions of the target perceived as artificial noise.

3.5. Prediction of the global score from specific scores

In order to study how global quality (instruction Q_1) can be explained by one or more of the three specific quality criteria (instructions Q_2 to Q_4), we now investigate how the former can be predicted from the latter.

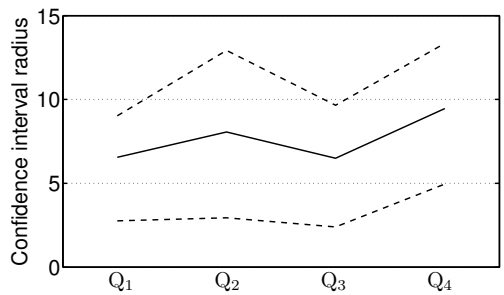

Fig. 2: Average (plain line), minimum (lower dashed line) and maximum (upper dashed line) over all trials of the 95%-confidence intervals over all subjects of the subjective scores of estimated sources from actual source separation algorithms.

3.5.1. Prediction model

We use a one-hidden layer feed forward neural network composed of K sigmoids to map the specific scores to the global score in a nonlinear way. The mapping function is defined as

$$f(I) \triangleq \sum_{k=1}^{K} v_k g(W_{k:} I + b_k) \tag{1}$$

where I is the input vector of length L, $g(x) \triangleq \frac{1}{1+e^{-x}}$ is the sigmoid function, v_k is the weight of sigmoid k, $W \triangleq [W_{1:}, \ldots, W_{K:}]$ is the $K \times L$ matrix of input weights and b_k is the bias of sigmoid k.

Since g is monotonous with values between 0 and 1, we ensure the monotonicity of f by constraining W and v to nonnegative values. The parameters $v \triangleq (v_1, \ldots, v_K)^{\mathrm{T}}$, W and $b \triangleq (b_1, \ldots, b_K)^{\mathrm{T}}$ can be estimated in a least-squares sense using Matlab's `fmincon` function.

3.5.2. Cross validation procedure

The parameters (v, W, b) of the mapping function are trained on a subset of the subjective scores and the prediction performance is evaluated on the remaining test set. Several cross validation procedures can be chosen:

- $\mathbb{X}_{\text{subj}}$: this 20-fold cross validation set-up aims at predicting the answer of a new subject. The training is performed on the data from 19 subjects while the data from the remaining subject is used for testing.
- $\mathbb{X}_{\text{mix}}$: this 10-fold cross validation set-up aims at predicting the quality for new mixtures. The training is performed on the data from 9 mixtures while the data from the remaining mixture is used for testing.
- $\mathbb{X}_{\text{subj\&mix}}$: this 200-fold cross validation set-up finally aims at predicting the quality for new mixtures and new subjects. The training is performed on the data from 19 subjects for 9 mixtures while the data from the remaining subject for the remaining mixture is used for testing.

3.5.3. Global score prediction results

We use three quality assessment metrics, as defined in [16], to study the prediction performance: the prediction *accuracy* given by Pearson's linear correlation between the predicted scores and the true global scores, the prediction *monotonicity* given by Spearman's rank correlation and the prediction *consistency* given by $1 - R_o$, where R_o is the amount of prediction outliers. Outliers are defined as values for which the prediction error is greater than twice the standard deviation among subjects.

The main results are presented in Fig. 3 using the $\mathbb{X}_{\text{subj\&mix}}$ cross-validation set-up. Several combinations of one to three of the specific criteria were tested to predict the global score. The best prediction is obtained when combining all three criteria, which confirms that these criteria provide distinct information and must all be tested. Criterion 3 (suppression of the other sources) appears to be the most important factor to predict the global score. Indeed prediction performance does not decrease much when combining this criterion with one of the other criteria, while poor results are obtained when combining criteria 2 and 4. The worst prediction is obtained when considering criterion 4 alone (absence of additional artificial noise). Finally, the optimal number of sigmoids is from $K = 2$ to 8.

In the subsequent study, we only consider the prediction of the global score from all the three specific criteria and the best value of K in each case.

The effect of the cross validation procedure is shown in Fig. 4. The $\mathbb{X}_{\text{subj\&mix}}$ procedure, for which there is a maximum independence between the training and testing data, gives the lowest prediction performance. It also can be seen that the prediction of the global score for an unknown sound is more difficult than the prediction of a new subject's anwers. However, the values obtained for the three cross-validation procedures are close to each other, suggesting that the amount of data is large enough to ensure good generalization, even though it was designed to provide a large variability in its contents.

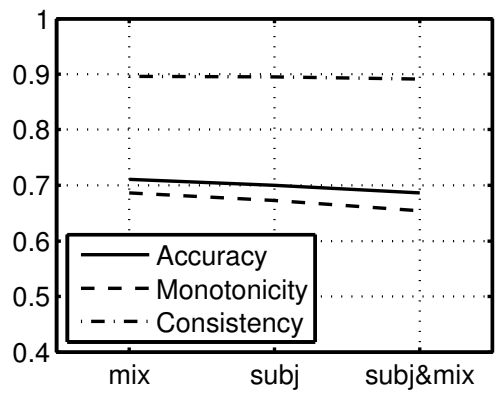

Fig. 4: Prediction of the global score from the three specific scores as a function of the cross validation procedure ($\mathbb{X}_{\text{mix}}$, $\mathbb{X}_{\text{subj}}$ and $\mathbb{X}_{\text{subj\&mix}}$).

Using the $\mathbb{X}_{\text{subj\&mix}}$ procedure, we present the results as a function of the type of data in Fig. 5. The prediction of the global quality of anchor sounds happens to be difficult to achieve. This may be due to the high level of distortion in these sounds, which may cause subjects to answer with very diverse values.

Finally, we investigate the possible existence of subject-dependent strategies: for each subject independently, we use a 10-fold cross validation by training the predictor on 9 mixtures and testing it on the remaining one. The results are presented in Fig. 6 and show a significant improvement of the performance when isolating a subject to train the predictor and to test it on a new sound. Thus, subjects may have their own criteria and strategies to assess the global quality and predicting it from the three

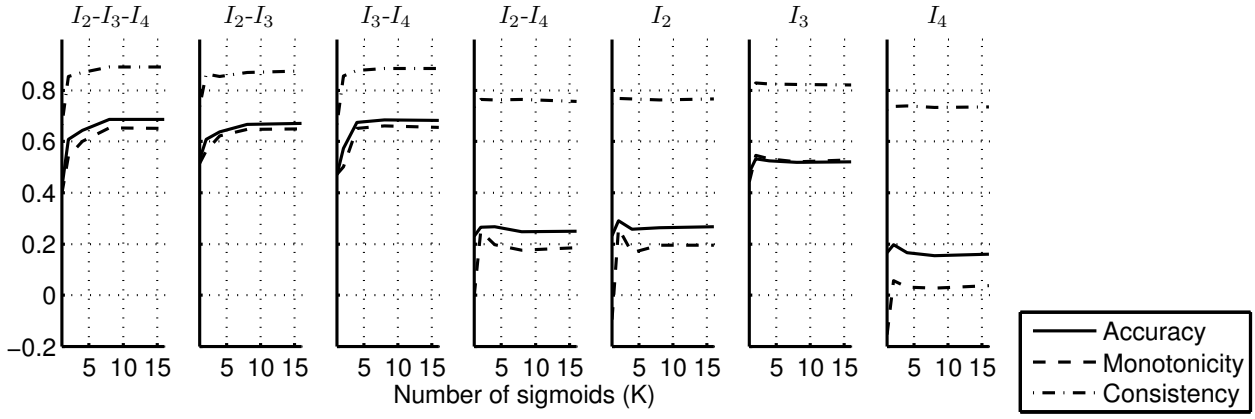

Fig. 3: Prediction of the global score from one to three specific scores, as a function of the number of sigmoids (x-axis) and of the inputs (sub-figures). Inputs I_2, I_3 and I_4 refer to the subjective scores according to instructions Q_2, Q_3 and Q_4 respectively.

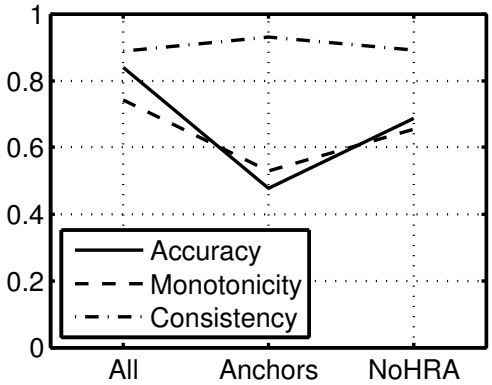

Fig. 5: Prediction of the global score from the three specific scores for all the data (All), the anchors only (Anchors), the actual source separation sounds (NoHRA, no hidden reference or anchors).

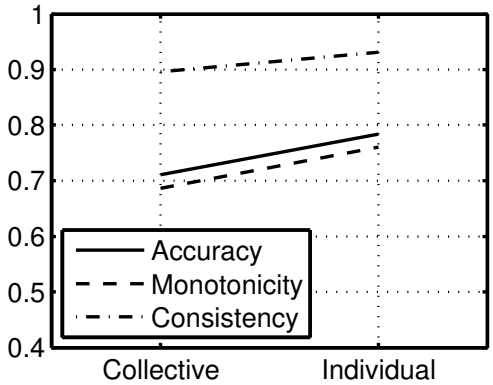

Fig. 6: Investigation of possible subject-dependent strategies: collective (*i.e.* $\mathbb{X}_{\text{mix}}$ procedure) vs individual training.

specific considered criteria may be valid to some extent only. Further subjective tests are needed to investigate these possible strategies.

4. OBJECTIVE MEASURES

4.1. Principle

Let us consider the audio source separation problem where we obtained an estimate $\widehat{\mathbf{s}}_{j_0}$ of a target source $\mathbf{s}_{j_0}$ from the mixture $\mathbf{x}(t) = \sum_{j=1}^{J} \mathbf{s}_j(t)$ of J sources, in a single or multichannel setting. In this scheme, we only consider the contributions of the sources at the location of the microphones – *i.e.* the so-called source images [3] – but our approach can be applied to the source signals at the source locations in a straigthforward way.

Previous works [3, 12] introduced the decomposition of the resulting distortion as:

$$\widehat{\mathbf{s}}_{j_0} - \mathbf{s}_{j_0} \triangleq \mathbf{e}_{j_0}^{\text{Target}} + \mathbf{e}_{j_0}^{\text{Interf}} + \mathbf{e}_{j_0}^{\text{Artif}} \qquad (2)$$

where $\mathbf{e}_{j_0}^{\text{Target}}$ is the distortion of the target, $\mathbf{e}_{j_0}^{\text{Interf}}$ is the interference distortion component due to the other sources

and $\mathbf{e}_{j_0}^{\text{Artif}}$ is the artifact distortion component, not related to the sources.

Extracting the three distortion components requires the definition and the extraction of the distortions one can perceive in an estimated source. In the most recent related state-of-the-art approach [17], a matched time-invariant multichannel filter is used to extract $\mathbf{e}_{j_0}^{\text{Target}}$ and $\mathbf{e}_{j_0}^{\text{Interf}}$ given reference signals of all sources. However, the resulting decomposition suffers from some limitations of this distortion model, which for instance is time-invariant and does not take any auditory model into account. As a consequence, the original sources can often be heard in the $\mathbf{e}_{j_0}^{\text{Artif}}$ component. We propose a new method to perform the decomposition described in Eq. (2) based on an auditory subband processing in which a time-varying filtering distortion is allowed.

Once the distortion components are extracted, the multi-criteria evaluation can be addressed by computing energy ratios [17]. As energy ratios do not well describe the perceived audio quality, we propose to use the PEMO-Q audio quality measure [18] provide a multi-criteria evaluation framework based on the distortion decomposition

4.2. Algorithm

The distortion decomposition is first obtained by:

- applying a gammatone filterbank [19] to the estimated signal and to the reference source signals;
- segmenting the subband signals into overlapping frames;
- decomposing each frame of the estimated source into distortion components using a matched FIR filter akin to that in [17], with a subband-dependent length;
- reconstructing the subband distortion components by an overlap-and-add method;
- reconstructing the 3 full-band distortion components using the inverse filter bank;

The objective measures are then obtained by:

- using the PEMO-Q measure [18] to compute the following features:
  - the salience of the target distortion, by comparing $\widehat{\mathbf{s}}_{j_0}$ with $\widehat{\mathbf{s}}_{j_0} - \mathbf{e}_{j_0}^{\text{Target}}$;
  - the salience of the interference distortion component, by comparing $\widehat{\mathbf{s}}_{j_0}$ with $\widehat{\mathbf{s}}_{j_0} - \mathbf{e}_{j_0}^{\text{Interf}}$;
  - the salience of the artifact distortion component, by comparing $\widehat{\mathbf{s}}_{j_0}$ with $\widehat{\mathbf{s}}_{j_0} - \mathbf{e}_{j_0}^{\text{Artif}}$;
  - the global distortion, by comparing $\widehat{\mathbf{s}}_{j_0}$ with $\mathbf{s}_{j_0}$.
- using the non-linear mapping function defined by Eq. (1) to predict the four subjective quality criteria from the above signal features.

	Accuracy	Monotonicity	Consistency
Q_1	0.68	0.64	0.92
Q_2	0.51	0.51	0.88
Q_3	0.75	0.76	0.91
Q_4	0.49	0.48	0.93

Table 1: Prediction performance using the proposed objective measures.

4.3. Prediction performance

We trained the parameters of the mapping function using the subjective measures in order to predict the answer to the four instructions Q_1 to Q_4. Using the $\mathbb{X}_{\text{subj\&mix}}$ cross-validation procedure, we obtained the prediction performance detailed in Table 1.

The performance for the prediction of the global score (Q_1) is very close to the figures obtained in Section 3, in particular in the left part of Fig. 3. This suggests that the proposed objective measure performs well in reproducing the multi-criteria subjective grading described in Section 2. A significant performance improvement is also observed when comparing the proposed approach to a single criterion evaluation. For instance, the accuracy, monotonicity and consistency are equal to 0.37, 0.37 and 0.79 respectively in the case of the SDR and 0.53, 0.41 and 0.83 respectively in the case of applying PEMO-Q to the estimated and original signals without any decomposition. The performance of prediction of the specific criteria (Q_2,Q_3,Q_4) can be related to the results shown in Section 3.4: in both cases, the interference criterion (Q_3) is easier to predict than the target distortion and the artifact criteria (Q_2 and Q_4).

5. CONCLUSIONS

We proposed a test protocol for the subjective evaluation of audio source separation that includes a four-criteria

evaluation and a MUSHRA protocol with several dedicated anchor sounds. Such a test was conducted and the statistical analysis showed the consistency of the results. Some noticeable conclusions were drawn up in terms of dependence between the target distortion and artificial noise criteria; of prediction of the global score from the specific criteria; and of subject-dependent strategies. We also briefly described a set of objective measures which provides a similar multi-criteria evaluation and offers good prediction performance.

We believe that the proposed objective measures could be adapted to evaluate the perceived quality in different application scenarios where the sources are not directly listened to, but subject to remixing or simultaneous 3D rendering. In this case, the target signal to be estimated is the remix or the rendering of the true sources. The proposed decomposition procedure could then be used to decompose the distortion into inteference resulting in spatial spreading of the rendered sources and artifacts which may or may not be heard depending on the presence of maskers. Separate distortion components could even be computed for each source, enabling the evaluation of advanced 3D attributes which cannot be accurately computed from the mixture today.

ACKNOWLEDGMENT

This work was supported by the French Ministry of Foreign and European Affairs and the German Academic Exchange Service under projet Procope N 20140NH, as well as by the Federeral Ministry of Education and Research within the scope of "Model-based hearing systems".

6. REFERENCES

[1] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Wiley/IEEE Press, 2006.

[2] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Probabilistic modeling paradigms for audio source separation," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. IGI Global, in press.

[3] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Lang. Proces.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

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

[5] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms," in *Proc. of UK ICA research network workshop*, Liverpool, UK, Sep. 2006.

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

[7] R. Prasad, "Fixed-point ICA based speech signal separation and enhancement with generalized Gaussian model," Ph.D. dissertation, Nara Insitute of Science and Technology, 2005.

[8] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 3, Mar. 2005, pp. 81–84.

[9] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of subjective and objective evaluation methods for audio source separation," in *Proc. Meetings on Acoustics*, vol. 4, no. 1. ASA, 2008, p. 050001.

[10] ITU, "ITU-T Recommendation P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," 2003.

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

[12] E. Vincent, S. Araki, and P. Bofill, "The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation*. Springer, Mar. 2009, pp. 734–741.

[13] ITU, "ITU-R Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems," 2003.

[14] ISO, "ISO 532: Acoustics – method for calculating loudness level," 1975.

[15] P. J. Rousseeuw and B. C. v. Zomeren, "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990.

[16] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. Wiley, 2005.

[17] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," in *Int. Conf. on Independent Component Analysis and Signal Separation*. London, UK: Springer-Verlag New York Inc, 2007.

[18] R. Huber and B. Kollmeier, "PEMO-Q – A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. Audio, Speech and Lang. Proces.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[19] T. Herzke and V. Hohmann, "Improved numerical methods for gammatone filterbank analysis and synthesis," *Acta Acustica*, vol. 93, no. 3, pp. 498–500, 2007.