Subjective and objective quality assessment of audio source separation

Valentin Emiya, Emmanuel Vincent, Niklas Harlander, Volker Hohmann

To cite this version:

Valentin Emiya, Emmanuel Vincent, Niklas Harlander, Volker Hohmann. Subjective and objective quality assessment of audio source separation. IEEE Transactions on Audio, Speech and Language Processing, Institute of Electrical and Electronics Engineers, 2011, 19 (7), pp.2046-2057. <http://ieeexplore.ieee.org/search/srchabstract.jsp?tp=&arnumber=5704564>. <10.1109/TASL.2011.2109381>. <inria-00567152>

HAL Id: inria-00567152

https://hal.inria.fr/inria-00567152

Submitted on 18 Feb 2011

Subjective and objective quality assessment of audio source separation

Valentin Emiya, *Member, IEEE*, Emmanuel Vincent, *Member, IEEE*, Niklas Harlander, Volker Hohmann

***Abstract*—We aim to assess the perceived quality of estimated source signals in the context of audio source separation. These signals may involve one or more kinds of distortions, including distortion of the target source, interference from the other sources or musical noise artifacts. We propose a subjective test protocol to assess the perceived quality with respect to each kind of distortion and collect the scores of 20 subjects over 80 sounds. We then propose a family of objective measures aiming to predict these subjective scores based on the decomposition of the estimation error into several distortion components and on the use of the PEMO-Q perceptual salience measure to provide multiple features that are then combined. These measures increase correlation with subjective scores up to 0.5 compared to nonlinear mapping of individual state-of-the-art source separation measures. Finally, we released the data and code presented in this paper in a freely-available toolkit called PEASS.**

***Index Terms*—Source separation, audio, quality assessment, objective measure, subjective test protocol**

I. Introduction and state of the art

Audio source separation is the task of extracting the signal of each sound source from a mixture of concurrent sources (see [1], [2], [3], [4] for a review). It underlies a wide range of applications from speech enhancement to content description and manipulation [5]. In this article, we consider applications where the estimated source signals are to be listened to, such as speech enhancement for hearing aids, denoising of old music recordings, and voice muting for karaoke. Separation performance then amounts to the subjective judgment of listeners. We focus on measuring and predicting the audio quality perceived by normal-hearing listeners for any input data and do not assess speech intelligibility or speech transcription, for which specific metrics were proposed in [6], [7], [8]. One or more kinds of distortions may be perceived depending on the separation algorithm, including *distortion of the target source*, *interference* from the other sources, and musical noise or other *artifacts* [9]. Multi-criteria evaluation is therefore necessary.

A number of studies have been performed to assess the subjective quality of certain source separation schemes [10], [11], [12], [13], [14], [15], [16], [17]. Most studies consider either a single criterion, such as overall quality [10], [14], [15], preference [13, p. 138] or musical noise salience [12], [18], or a set of criteria restricted to speech [11], [16]. Three such criteria called intelligibility, fidelity and suppression were proposed in [11, p. 95], while [16] employs the standard ITU criteria for noise suppression [19], namely speech signal distortion, background noise intrusiveness and overall quality. Dedicated multi-criteria protocols are a promising extension to established single-criterion protocols but have not been investigated in detail yet. Besides, most studies consider a single class of algorithms producing specific kinds and levels of distortion, *e.g.* Independent Component Analysis (ICA) in [13], time-frequency masking in [10] or simulated separation in [16], and a narrow range of sound material, *e.g.* male speech in [14], [17] or isolated notes from a single musical instrument in [15]. The resulting scores can hence not be compared due to the lack of a common absolute reference. Finally, some test protocols are inappropriate or insufficiently documented. Pairwise comparison tests are employed in [15], while joint presentation is known to be preferable with large degradations [20] such as those encountered in source separation. Also, the protocols in [10], [12], [11] are not fully described, *e.g.* in terms of sound normalization, sound presentation or subject training, so that they are not exactly reproducible.

In parallel to the subjective studies [10], [11], [12], [13], [14], [15], [16], [17], the objective evaluation of source separation algorithms has also received some attention. A common approach to evaluating the quality of an estimated source signal is to compute the Signal to Distortion Ratio (SDR) between the energy of the *reference*, *i.e.* the clean target signal, and that of the distortion [9]. Two directions have been investigated to derive additional objective measures. The first one consists of decomposing the distortion signal into several components [9], related to *e.g.* target distortion, interference, sensor noise and artifacts, and deriving a specific energy ratio from each of the distortion components [9], [21]. These energy ratios may further be combined using linear or nonlinear mapping to increase correlation with subjective ratings [15], [16]. However, the distortion decomposition algorithms proposed so far do not always yield the expected components and one may question the ability of energy ratios to fit subjective ratings since auditory phenomena such as loudness perception [22] and spectral masking are not taken into account. A second direction is to use auditory-motivated metrics to compare the target and the estimated source. Existing metrics designed for

V. Emiya and E. Vincent are with INRIA, Centre Inria Rennes - Bretagne Atlantique, 35042 Rennes Cedex, France (e-mail: valentin.emiya@inria.fr; emmanuel.vincent@inria.fr). N. Harlander and V. Hohmann are with Medizinische Physik, Carl von Ossietzky-Universität Oldenburg, Oldenburg, Germany (e-mail: niklas.harlander@uni-oldenburg.de; volker.hohmann@uni-oldenburg.de).

This work was supported by the French Ministry of Foreign and European Affairs, the French Ministry of Higher Education and Research and the German Academic Exchange Service under project Procope 20142UD, as well as by the German Federal Ministry of Education and Research under project number 01EZ0741 “Model-based hearing systems”.

Part of this study has been presented at the 38th International AES Conference on Sound Quality Evaluation in June 2010.

audio coding or speech enhancement remain however limited to the assessment of overall quality [23], [24], [25] and appear to perform poorer than decomposition-based measures in the context of source separation [15].

This article provides the following contributions for subjective and objective quality assessment of audio source separation: a principled multi-criteria subjective test protocol dedicated to the evaluation of source separation (Section II), a large database of 6400 subjective scores for a wide range of mixture signals and source separation schemes (Section III), a family of auditory-motivated objective measures based on improved distortion decomposition (Section IV) and a validation of the ability of these objective measures to predict these subjective scores (Section V). We provide in particular additional evidence compared to [15], [16] that decomposing the distortion into several components and combining the resulting objective measures improves the quality prediction. The sound material, the subjective data and the objective measures are released as a toolkit named PEASS (Section VI. We conclude in Section VII.

II. Multi-criteria subjective test protocol

The proposed subjective test protocol relies on the principle of multi-criteria evaluation in a similar way as the ITU standard for the evaluation of noise suppression algorithms [19]. Based on previous work on the objective evaluation of source separation [9], we propose a set of three specific criteria besides overall quality which are dedicated to source separation: preservation of the target source, suppression of other sources and absence of additional artificial noise. We formulated these criteria for experts in general audio applications so as to avoid reference to specific source separation terms such as interference and artifacts.

A. Protocol

We propose four separate listening tests, in which the subjects are asked to address the following four tasks respectively:

1) rate the *global quality* compared to the reference for each test signal;
2) rate the quality in terms of *preservation of the target source* in each test signal;
3) rate the quality in terms of *suppression of other sources* in each test signal;
4) rate the quality in terms of *absence of additional artificial* noise in each test signal.

The tests are performed in the above order, with a break at the end of each test. This is a major difference with respect to the ITU P.835 standard [19]. In the latter, the overall quality is rated after speech signal distortion and background noise intrusiveness and required to be a combination of these two subjective factors. In the proposed protocol, the global quality is assessed first for the opposite purpose: instead of aiming for a global score combining the three specific scores only, we want to relax their influence and allow global quality scores to possibly involve other subjective factors at the expense of a possibly larger variance.

The MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [20] protocol is employed for each test. This protocol is appropriate here since medium and large impairments are encountered [14]. For a given mixture and a given target source within that mixture, the subject is jointly presented with several test sounds in a random order, including the results of the source separation algorithms under test, the reference clean target source and some anchor sounds introduced below. The reference and the mixture are also available for comparison. The perceived loudness of the reference should be adjusted as much as possible to the same value for all mixtures. The other test sounds may be normalized to the same loudness or not, depending on whether erroneous scaling is considered as a distortion or not [9].

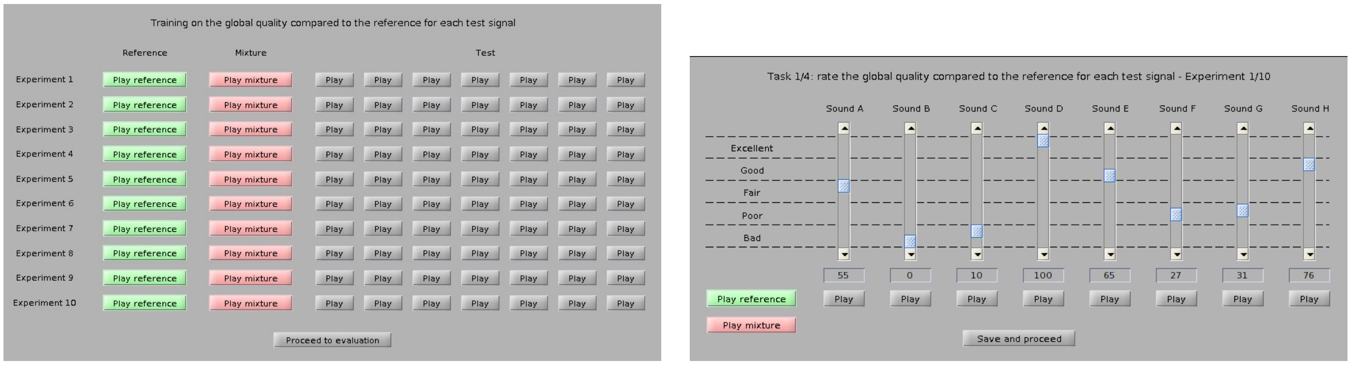
A training phase is first conducted where the subject listens to all sounds of all mixtures (see Fig. 1(a)). This aims to train the subject to address the required task, to learn the range of observed quality according to that task and to fix the volume of the headphones to a comfortable level. A grading phase is then performed for each mixture and target source where the subject rates the quality of each test sound compared to the reference on a scale from 0 to 100, where higher ratings indicate better quality (see Fig. 1(b)). Sounds may be listened to as many times as desired. The subject should make sure that the ratings are consistent across mixtures (*i.e.* if one sound has better quality than another, it should be rated higher) and that the whole rating scale is used (*i.e.* sounds with perfect quality should be rated 100 and the worst test sound over all mixtures should be rated 0).

The guidelines of the test are presented as a unique written document for all subjects in order to avoid any influence from the supervisor of the test.

B. Anchor sounds

An essential aspect of the MUSHRA protocol is the use of anchor sounds, *i.e.* artificial sounds presenting large impairments of the same kind as those generated by actual systems [20]. Precisely defined anchors act as absolute quality levels and allow the comparison of ratings obtained in different listening conditions or for different test sounds. In the context of audio coding, several anchors reproducing the distortions generated by audio coders were proposed in [20], [26]. Anchors for the evaluation of source separation were also introduced in [14]. We propose a new set of anchors inspired from [14] which better fit the target distortions and the artifacts produced by actual systems. Each anchor is associated with one of the three aforementioned kinds of distortion.

- The *distorted target anchor* is created by low-pass filtering the target source signal to a 3.5 kHz cut-off frequency and by randomly setting 20% of the remaining time-frequency coefficients to zero.
- The *interference anchor* is defined as the sum of the target source signal and an interfering signal. The latter is obtained by summing all interfering sources and by adjusting the loudness of the resulting signal to that of the target.
- The *artifacts anchor* is defined as the sum of the target source signal and an artifact signal. In order to generate



Training on the global quality compared to the reference for each test signal

Task 1/4: rate the global quality compared to the reference for each test signal - Experiment 1/10

(a) Training interface

(b) Grading interface

Fig. 1. Screenshots of the subjective test interfaces for the training and grading phases of task 1. The reference is not included in the test sounds of the training phase.

musical noise, which can be defined as "generated audible isolated spectral components" perceived as "harsch and artificial" [18], [27] or as "isolated [and] short ridges in the spectrogram" [28], [12], the latter artifact signal is created by randomly setting 99% of the time-frequency coefficients of the target to zero and by adjusting the loudness of the resulting signal to that of the target.

Fixed choice of the time-frequency transform and the loudness measure is needed for reproducibility. We consider the short time Fourier transform (STFT) with half-overlapping 46 ms sine windows (*i.e.* the square root of a Hann window) and the ISO 532B loudness measure [29] because of its availability as free Matlab software[1].

III. Database of subjective scores

We collected a set of 6400 subjective scores by implementing the above protocol via a dedicated interface. This interface is available together with the test sounds, the anchor sounds and the resulting scores within the PEASS toolkit (see Section VI).

A. Test material and subjects

1) Test material: We selected 8 stereo mixtures and 2 4-channel mixtures of 5 s duration from various datasets of the 2008 Signal Separation Evaluation Campaign (SiSEC) [30]. The target to be estimated was either the stereo spatial image of one source in the former case or one original single-channel source in the latter case [30]. These mixtures were chosen so as to cover a wide range of source separation settings as shown in Table I: two or more sources; instantaneous, anechoic, convolutive or professionally-produced mixtures; male speech, female speech, singing voice, pitched musical instrument or drums as the target source. The target-to-interference ratios of the mixtures ranged from -12 dB to 2 dB. For each mixture, the 8 test sounds consisted of four sounds generated by actual source separation algorithms, the reference and the three anchor sounds. All references were set to the same loudness using the ISO 532B standard. The sounds from actual source separation schemes were obtained by 13 different algorithms as described in [30]. From one mixture to another, different algorithms were chosen in order to favor a wide range of distortions and state-of-the-art separation methods.

2) Subjects: 23 normal-hearing subjects (excluding the authors) participated in the test, including 13 in Rennes, France, and 10 in Oldenburg, Germany. All subjects were experts in general audio applications, as required by the MUSHRA protocol [20]. They used the same AKG 271 headphones and performed the test in different offices, in a quiet environment. The guidelines were written in English.

#	Mixture	Type	Target	Interferences
1	Convolutive	Speech	Male	Male, female
2	Convolutive	Speech	Female	Male
3	Anechoic	Speech	2 males successively	Male & female successively
4	Professional mix	Music (rock)	Male singer	2 guitars, 2 keyboards, bass, drums
5	Instantaneous	Music (pop)	Piano	Male singer, bass
6	Instantaneous	Music (pop)	Electric guitar	Acoustic guitar, bass
7	Convolutive	Speech	Male	Female
8	Professional mix	Music (bossa nova)	Female singer	Acoustic guitar
9	Convolutive	Speech	Male	2 males
10	Convolutive	Music (rock)	Drums	Female singer, electric guitar

TABLE I
Mixing conditions, type of sounds and nature of the target and interfering sources for each of the ten test mixtures.

B. Statistical analysis of the results

1) Detection of outlier subjects: A post-screening was applied so as to discard outlier subjects that may have misunderstood the guidelines. This post-screening was performed on the scores related to the hidden reference and the anchor sounds for all mixtures only. Indeed, a consensus among subjects is expected over these sounds since they involve either no distortion or a single kind of distortion. By contrast, subjects may have individual rating strategies over the remaining sounds involving multiple distortions due their individual perceptual weighting of each kind of distortion.

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

We used the multivariate Mahalanobis distance-based outlier detection technique in [31]. The set of subjective scores of subject m is considered as a vector $\mathbf{y}_m$. Let us denote by $d_m^2 = (\mathbf{y}_m - \boldsymbol{\mu}_\mathbf{y})\boldsymbol{\Sigma}_\mathbf{y}^{-1}(\mathbf{y}_m - \boldsymbol{\mu}_\mathbf{y})^\mathrm{T}$ the squared Mahalanobis distance between $\mathbf{y}_m$ and the empirical data mean $\boldsymbol{\mu}_\mathbf{y}$, $\boldsymbol{\Sigma}_\mathbf{y}$ being the empirical data covariance. The distances d_m are assumed to be distributed according to a χ^2 law [31]. Hence, by matching the empirical and theoretical cumulative distributions, outliers are obtained as points of the empirical distribution above the 0.975 quantile of the theoretical χ^2 distribution [31]. In the current case, 3 outliers were detected among the 23 subjects and removed for subsequent use of the subjective scores.

2) Effect of location: To substantiate confidence in the results, an analysis of variance (ANOVA) was performed regarding the subject location (Oldenburg *vs.* Rennes). We used SPSS Statistics 12.0[2] with a significance level of $\alpha = 0.05$. The two locations were a "between" factor while the four tasks and the 10 mixtures were "within" factors. We obtained highly significant effects of tasks ($\eta^2 = 0.837$) and mixtures ($\eta^2 = 0.567$), with all $p < 0.05$ and corrected F-values from 92.3 to 23.6. No significant effect of locations was detected ($F(1, 18) < 1$, $p = 0.597$, $\eta^2 = 0.01$). As a result, location did not have a significant influence on the subjective scores.

3) Statistical analysis for hidden references and anchors: A separate statistical analysis of the subjective ratings is provided for the set of hidden references and anchors and for the set of sounds from actual source separation schemes. The statistical analysis related to the hidden references and anchor sounds is presented in Fig. 2. It shows that for each task, the hidden references were scored around 100, as expected, with a very narrow confidence interval (less than 1). The confidence intervals related to anchor sounds are wider, with half-widths from ±1.4 to ±12.6.

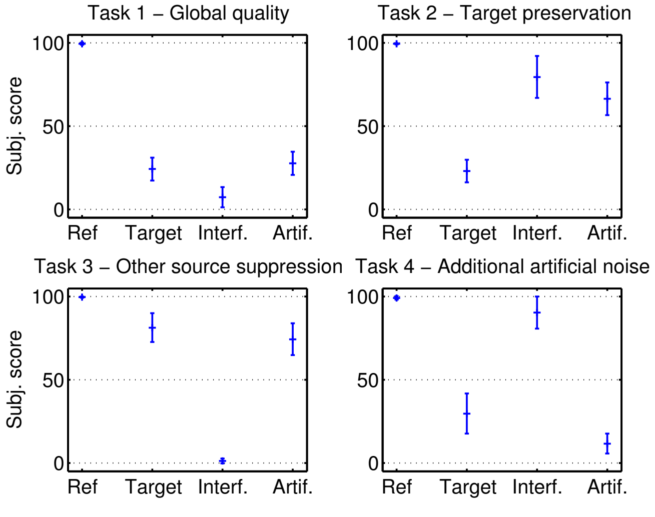


Fig. 2. Mean and 95% confidence intervals of the subjective scores for the hidden references and the three anchor sounds (abcissa) for each of the four tasks (subfigures).

The mean values in Fig. 2 indicate that all anchors have low scores for task 1 (global quality), as expected. For tasks 2 to 4, the anchor related to the considered task has a low score while the other ones have high scores, except for the distorted target anchor in task 4. Indeed, the distorted target anchor presents large distortions which do sound as artificial noise. Conversely, the artifacts anchor does not have a low score in task 2 since artifacts do not sound as target distortion. Thus, we see that the three anchors involve independent distortions to some extent only. Future investigations may be needed to identify the kinds of target distortions that are subjectively correlated with the target and design more independent anchors.

Tasks	Min.	Average	Max
Task 1: Global score	±2.8	±6.5	±9.0
Task 2: Target preservation	±2.9	±8.1	±12.9
Task 3: Other source suppression	±2.4	±6.5	±9.7
Task 4: Additional artificial noise	±5.0	±9.5	±13.3

TABLE II
MINIMUM, AVERAGE AND MAXIMUM WIDTH OF THE 95% CONFIDENCE INTERVALS (IN GRADING POINTS) OVER THE SUBJECTIVE SCORES OF THE SOURCES ESTIMATED BY ACTUAL SOURCE SEPARATION ALGORITHMS.

4) Statistical analysis for test sounds produced by separation schemes: The confidence intervals related to the sounds from actual source separation algorithms are summarized in Table II. All half-widths are lower than 15, which is satisfying given the width of the grading scale and of the same order as in [14], [15], [16], [17]. Note that narrower confidence intervals were obtained for tasks 1 (global quality) and 3 (suppression of other sources), which indicates a slightly higher agreement of the subjects on these tasks than on tasks 2 and 4.

IV. MULTI-CRITERIA OBJECTIVE MEASURES

We now design a family of four objective measures aiming to predict the subjective scores of the above test. The proposed approach consists of splitting the distortion signal into a sum of components related to target distortion, interference and artifacts, of assessing their perceptual salience using auditory-motivated metrics and of combining the resulting features via nonlinear mappings. The distortion components are extracted using a new approach described in Section IV-A and validated in Section IV-B, while the derived measures are detailed in Section IV-C.

In the following, we consider a mixture with I channels and J sources indexed by i and j respectively. The spatial image of source j sampled at time t, *i.e.* its contribution to each mixture channel i, is denoted by $s_{ij}(t)$. We assume that the true spatial images of all sources are known. For a given target source j, we evaluate the quality of *source spatial image estimation* [30] by comparing the multichannel spatial image $\widehat{s}_{ij}(t)$ estimated by some source separation algorithm to the target $s_{ij}(t)$. The following derivations can be applied in a straightforward way to the problem of *source signal estimation* [30] by replacing these signals by the estimated and target single-channel source signals $\widehat{s}_j(t)$ and $s_j(t)$ instead.

A. Distortion component model and estimation

Following [21], we split the distortion between the estimate $\widehat{s}_{ij}(t)$ and the target $s_{ij}(t)$ into the sum of a target distortion

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

component $e_{ij}^{\text{target}}(t)$, an interference component $e_{ij}^{\text{interf}}(t)$ and an artifacts component $e_{ij}^{\text{artif}}(t)$ such that[3]

$$\widehat{s}_{ij}(t) - s_{ij}(t) = e_{ij}^{\text{target}}(t) + e_{ij}^{\text{interf}}(t) + e_{ij}^{\text{artif}}(t). \qquad (1)$$

In order to perform this decomposition, one must specify how the target distortion and interference components relate to the true source signals. It remains unknown however how the auditory system segregates the streams associated to these components. One approach [21] is to assume that these components are linearly distorted versions of the true source signals, where distortion is modeled via multichannel time-invariant Finite Impulse Response (FIR) filters. The algorithm in [21] computes the coefficients of these filters by two nested least-square projections: first, the distortion signal is projected onto the subspace spanned by delayed versions of all source signals $s_{kl}(t-\tau)$, $1 \leq k \leq I$, $1 \leq l \leq J$, $0 \leq \tau \leq L-1$, so as to obtain $e_{ij}^{\text{target}}(t) + e_{ij}^{\text{interf}}(t)$; then it is further projected on the smaller subspace spanned by delayed versions of the target signal $s_{kj}(t-\tau)$, $1 \leq k \leq I$, $0 \leq \tau \leq L-1$, so as to obtain $e_{ij}^{\text{target}}(t)$ alone; finally, $e_{ij}^{\text{artif}}(t)$ is defined as the residual. The filter length L is typically set to 32 ms [21].

Despite their use in several evaluation campaigns [21], [30], [32], the resulting distortion components do not always fit those perceived by human listeners. This can be checked by listening to the audio examples accompanying [9] or the current article (see Section VI). For instance, one can often hear the original sources when listening to the artifacts component. This is due in particular to the time-invariant model which does not fit the time-varying nature of the encountered distortions and to the constant frequency resolution of the FIR filter which does not match that of the ear. A time-varying decomposition was proposed in [9]. However, due to its large computational cost, it was restricted in practice to filters with both low frequency resolution and low time resolution [9] and consequently did not improve the results. Another issue is that the target distortion component may be nonzero even when the target is not distorted. Indeed, due to the nested projection algorithm, the target distortion component includes part of the target source signal $s_{ij}(t)$ in addition to the interfering source signals $s_{il}(t)$, $l \neq j$, as soon as these signals are correlated.

The proposed decomposition algorithm aims to fix these issues and output more perceptually relevant distortion components by approximating the auditory time-frequency resolution. As illustrated in Fig. 3, it involves three successive steps: firstly, all signals are partitioned into time- and frequency-localized signals via a gammatone filterbank [33] followed by downsampling and windowing; secondly, a time-invariant FIR-based decomposition is performed in each subband and each time frame by joint least-squares projection; finally, time-domain signals are reconstructed via overlap-and-add (OLA) and filterbank inversion. Besides its desirable auditory-motivated resolution, the filterbank makes it possible to decrease the filter length and hence the computational cost of each decomposition. We now describe the details of each step.

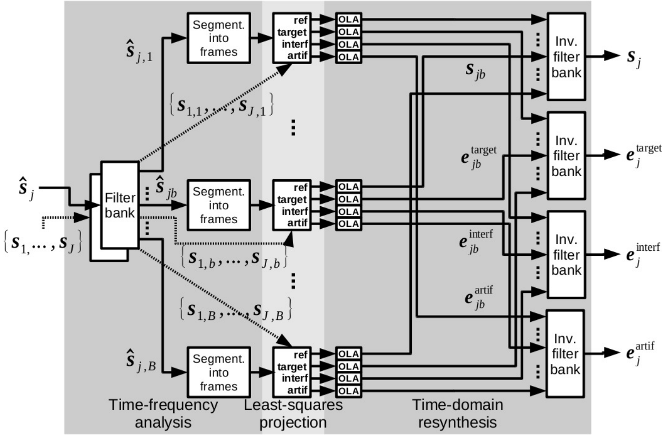


Fig. 3. Block diagram of the proposed algorithm for the decomposition of an estimated source into the sum of the target source and three distortion components corresponding to target distortion, interference and artifacts.

1) Time-frequency analysis: We split the estimated source signal $\widehat{s}_{ij}(t)$ and the true signals of all sources $s_{kl}(t)$, $1 \leq k \leq I$, $1 \leq l \leq J$, into subband signals $\widehat{s}_{ijb}(t)$ and $s_{klb}(t)$ indexed by b using a bank of 4th-order gammatone filters as implemented in [33], [34]. The center frequencies are linearly spaced on the auditory-motivated Equivalent Rectangular Bandwidth (ERB) scale from 20 Hz to the Nyquist frequency. This scale is approximately linear at low frequencies and logarithmic at high frequencies. To ensure good reconstruction properties, the number of filters per ERB is set to 3. All subband signals are downsampled by a factor equal to the ratio of the Nyquist frequency and the filter bandwidth *i.e.* 1 ERB[4].

In each subband, the estimated source signal $\widehat{s}_{ijb}(t)$ and the delayed true source signals $s_{klb}(t-\tau)$, $1 \leq k \leq I$, $1 \leq l \leq J$, $-L/2 \leq \tau \leq L/2$, are partitioned into overlapping time frames indexed by u via

$$\widehat{s}_{ijbu}(t) = w_a(t)\widehat{s}_{ijb}(t-uN) \qquad (2)$$
$$s_{klbu}^{\tau}(t) = w_a(t)s_{klb}(t-uN-\tau) \qquad (3)$$

where w_a denotes the analysis window and N the stepsize. We employ a sine window with fixed length T and stepsize $N = T/4$. Due to downsampling, this translates into variable time resolution in the original time domain: the time resolution in each subband is inversely proportional to its bandwidth. Several window lengths are considered in Section V-C and shown to be non critical.

2) Joint least-squares decomposition: Due to the wide bandwidth of gammatone filters, the distortion components are estimated by an additional filtering in each subband and each time frame. These components are defined by multichannel time-invariant FIR filtering of the target source signal and the interfering source signals, respectively, while the artifacts

[3] An additional residual noise component may be defined when considering noisy mixtures [9].

[4] For simplicity, we also denote by t the time index after downsampling.

component is given by the residual distortion:

$$e_{ijbu}^{\text{target}}(t) = \sum_{k=1}^{I} \sum_{\tau=-L/2}^{L/2} \alpha_{ijbu,kj}(\tau) s_{kjbu}^{\tau}(t) \quad (4)$$

$$e_{ijbu}^{\text{interf}}(t) = \sum_{k=1}^{I} \sum_{l \neq j} \sum_{\tau=-L/2}^{L/2} \alpha_{ijbu,kl}(\tau) s_{klbu}^{\tau}(t) \quad (5)$$

$$e_{ijbu}^{\text{artif}}(t) = \widehat{s}_{ijbu}(t) - s_{ijbu}^{0}(t) - e_{ijbu}^{\text{target}}(t) - e_{ijbu}^{\text{interf}}(t) \quad (6)$$

Note that, unlike [9], [21], centered FIR filters are used and the interference component explicitly excludes the target source j. The filter coefficients are computed by least-squares projection of the distortion $\widehat{s}_{ijbu}(t) - s_{ijbu}^{0}(t)$ onto the subspace spanned by the delayed source signals $s_{klbu}^{\tau}(t)$, $1 \leq k \leq I$, $1 \leq l \leq J$, $-L/2 \leq \tau \leq L/2$. Classically, the optimal $(L+1)IJ \times 1$ vector of coefficients is given by $\boldsymbol{\alpha}_{ijbu} = \mathbf{S}_{bu}^{+}(\widehat{\mathbf{s}}_{ijbu} - \mathbf{s}_{ijbu})$ where $\widehat{\mathbf{s}}_{ijbu}$ and $\mathbf{s}_{ijbu}$ are respectively the estimated and true $T \times 1$ vectors of target source samples, $\mathbf{S}_{bu}$ is the $T \times (L+1)IJ$ matrix of delayed true source samples and $^{+}$ denotes matrix pseudo-inversion.

The filter length L is set to a constant. Various lengths are considered in Section V-C. Again, due to downsampling, this translates into variable auditory-motivated frequency resolution in the original time domain.

3) Time-domain resynthesis: Full-duration distortion components are reconstructed from the time-localized components in each subband using OLA

$$e_{ijb}^{\text{target}}(t) = \sum_{u} w_s(t - uN) e_{ijbu}^{\text{target}}(t - uN) \quad (7)$$

$$e_{ijb}^{\text{interf}}(t) = \sum_{u} w_s(t - uN) e_{ijbu}^{\text{interf}}(t - uN) \quad (8)$$

$$e_{ijb}^{\text{artif}}(t) = \sum_{u} w_s(t - uN) e_{ijbu}^{\text{artif}}(t - uN) \quad (9)$$

where w_s is a sine synthesis window of length T such that $\sum_u w_s(t - uN) w_a(t - uN) = 1$. Finally, the fullband distortion components $e_{ij}^{\text{target}}(t)$, $e_{ij}^{\text{interf}}(t)$ and $e_{ij}^{\text{artif}}(t)$ are obtained using the synthesis filters [33] associated with the gammatone filterbank. In order to account for inaudible but measurable distortion due to filterbank inversion, the fullband estimated and true target signals $\widehat{s}_{ij}(t)$ and $s_{ij}(t)$ are also reconstructed from their subbands $\widehat{s}_{ijb}(t)$ and $s_{ijb}(t)$. These reconstructed versions are used in place of the original signals from now on.

B. Evaluation of the signal decomposition

An objective evaluation of the distortion decomposition is not obvious to design since reference signals for the distortion components are not available. Moreover, the creation of synthetic reference signals is not possible since it would imply some reductive *a priori* on the distortions, *e.g.* on the choice of time and frequency resolution or on the definition of artifacts. In order to validate the proposed method, the salience of the distortion components obtained via the state-of-the-art [21] (see Section IV-A) and the proposed decomposition are compared in Fig. 4 over the data of Section III. Two series of scatter plots are shown depending on the salience criteria defined hereafter in Section IV-C: either the energy ratios given by (11), (12) and (13) or the features q_j^{target}, q_j^{interf} and q_j^{artif} obtained from the auditory-based PEMO-Q metric [35] in (15), (16) and (17). Circled items can be listened to as part of the sound examples of the PEASS toolkit (see Section VI).

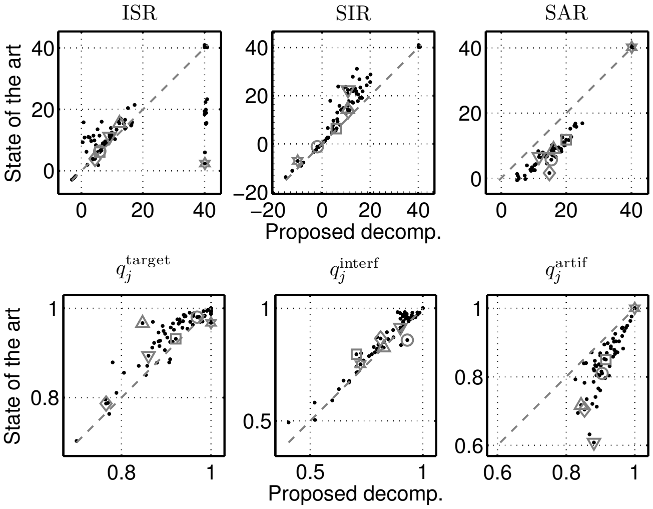


Fig. 4. Scatter plot of the energy ratios in dB (top) and the PEMO-Q-based features (bottom) for the state-of-the-art distortion decomposition (y-axis) vs. the proposed distortion decomposition (x-axis). The maximum value of energy ratios have been limited to 40 dB. Circled items can be listened to as part of the sound examples of the PEASS toolkit (see Section VI).

Many points are far from the diagonal dashed line, showing that the proposed decomposition differs from the state of the art for many of the tested sounds. In general, salience values are differently distributed for the PEMO-Q-based criteria and for the energy ratios. When listening to the artifacts components, one can realize that the sources are well removed with the proposed method whereas they can still be heard with the state-of-the-art one. This can be observed in the right plots of Fig. 4 where points are in the bottom part of the plots since the artifacts components obtained with the proposed method contains less energy. The proposed method also enhances the relevance of the target distortion and interference components, which results in scattered points in the left and center plots. Note that in the left plots, the vertically aligned points on the right side correspond to the interference anchor sounds for which an almost ideal decomposition is obtained with the proposed method thanks to joint projection onto all source signals, while the state-of-the-art one erroneously provides a nonzero target distortion due to nested projections.

C. Component-based objective measures

Given some decomposition of the distortion, like the ones presented in [21] or in Section IV-A, we now aim to assess the similarity between the estimated and the reference source signal according to each of the four subjective rating tasks of Section II. The state-of-the-art approach in the source separation community consists in measuring the salience of the overall distortion and of the target distortion, interference and artifacts components by means of energy ratios called respectively the Signal to Distortion Ratio (SDR), the source Image to Spatial distortion Ratio (ISR), the Signal to Interference Ratio (SIR) and the Signal to Artifacts Ratio (SAR) [21]:

$$\mathrm{SDR}_j = 10\log_{10}\frac{\sum_i\sum_t |s_{ij}(t)|^2}{\sum_i\sum_t |\widehat{s}_{ij}(t)-s_{ij}(t)|^2} \tag{10}$$

$$\mathrm{ISR}_j = 10\log_{10}\frac{\sum_i\sum_t |s_{ij}(t)|^2}{\sum_i\sum_t |e_{ij}^{\text{target}}(t)|^2} \tag{11}$$

$$\mathrm{SIR}_j = 10\log_{10}\frac{\sum_i\sum_t |s_{ij}(t)+e_{ij}^{\text{target}}(t)|^2}{\sum_i\sum_t |e_{ij}^{\text{interf}}(t)|^2} \tag{12}$$

$$\mathrm{SAR}_j = 10\log_{10}\frac{\sum_i\sum_t |s_{ij}(t)+e_{ij}^{\text{target}}(t)+e_{ij}^{\text{interf}}(t)|^2}{\sum_i\sum_t |e_{ij}^{\text{artif}}(t)|^2}. \tag{13}$$

These energy ratios do not always fit the perceptual salience of each component within the estimated source. For instance, low frequency components affect more energy ratios than perception. Also, the auditory masking of soft distortion components by the target signal or by louder distortion components is not taken into account.

In order to overcome these issues, we adopt the two-step approach in Fig. 5. First, we assess the salience of each distortion component using auditory model-based metrics. Note that this differs from the conventional use of such metrics, which are applied to the overall distortion instead of individual components [24], [25]. Then, we combine the resulting component-wise salience features by nonlinear mapping, yielding a family of four objective measures:

- the Overall Perceptual Score (OPS),
- the Target-related Perceptual Score (TPS),
- the Interference-related Perceptual Score (IPS),
- the Artifacts-related Perceptual Score (APS).

The details of each step are as follows.

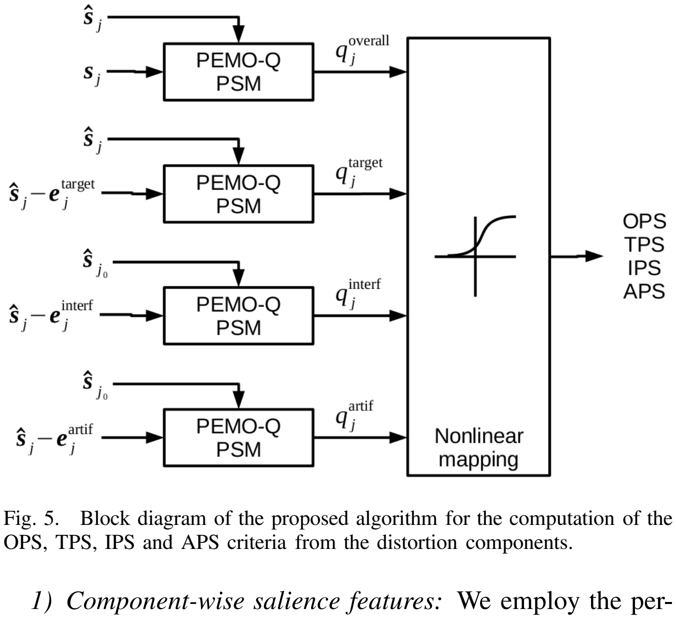


Fig. 5. Block diagram of the proposed algorithm for the computation of the OPS, TPS, IPS and APS criteria from the distortion components.

1) Component-wise salience features: We employ the perceptual similarity measure (PSM) provided by the PEMO-Q auditory model[5] [35]. The perceptual salience of the overall distortion and of each specific distortion component is assessed by comparing the estimated source signal with itself minus the considered distortion. This yields the following four features:

$$q_j^{\text{overall}} = \mathrm{PSM}(\widehat{\mathbf{s}}_j, \mathbf{s}_j) \tag{14}$$

$$q_j^{\text{target}} = \mathrm{PSM}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\text{target}}) \tag{15}$$

$$q_j^{\text{interf}} = \mathrm{PSM}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\text{interf}}) \tag{16}$$

$$q_j^{\text{artif}} = \mathrm{PSM}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\text{artif}}) \tag{17}$$

where bold letters denote the single-channel vectors[6] for all time indexes t.

2) Nonlinear mapping: Following other objective measures [23], [35], [15], a nonlinear mapping is applied to combine these features into a single scalar measure for each grading task and to adapt the feature scale to the subjective grading scale. We assume that the vector of features $\mathbf{q}_{jr}$ for a given task r involves either the four features $\mathbf{q}_{jr} = [q_j^{\text{overall}}, q_j^{\text{target}}, q_j^{\text{interf}}, q_j^{\text{artif}}]$ or a subset of these.

Complex shapes of nonlinear functions can be simulated by using several sigmoids. We employ a one hidden layer feedforward neural network composed of K sigmoids, the number of sigmoids being chosen empirically (see Sec. V-A). Each feature vector $\mathbf{q}_{jr}$ is mapped into an OPS, TPS, IPS or APS score $x_{jr} = f_r(\mathbf{q}_{jr})$ via the function

$$f_r(\mathbf{q}) = \sum_{k=1}^{K} v_{rk}\, g(\mathbf{w}_{rk}^{\mathrm{T}}\mathbf{q} + b_{rk}) \tag{18}$$

where $g(x) = 1/(1+e^{-x})$ is the sigmoid function and v_{rk}, $\mathbf{w}_{rk}$ and b_{rk} denote respectively the output weight, the vector of input weights and the bias of sigmoid k. Table III presents the various configurations of input vectors which are tested and discussed in Section V.

The neural network parameters are trained using Matlab's `fmincon` optimizer so as to minimize the mean square error between the predicted score x_{jr} and the subjective scores y_{jrm} of all subjects m. This is equivalent to minimizing the mean square error between x_{jr} and the mean subjective score $\bar{y}_{jr}$.

V. Evaluation of the objective measures

We assessed the ability of the family of objective measures proposed in Section IV to predict the subjective scores of Section II. In particular, the following factors were investigated: the use of the proposed distortion decomposition as opposed to that in [21], the choice of the window and filter lengths T and L, the use of PEMO-Q as opposed to energy ratios and the various configurations of the feature vector. In order to ensure a fair comparison, the nonlinear mapping defined in Eq. (18) is used in all cases to match the objective scores as well as possible. In the case of energy ratios, the same configurations of the feature vector are employed as in Table III with q_j^{overall}, q_j^{target}, q_j^{interf} and q_j^{artif} being replaced by SDR_j, ISR_j, SIR_j and SAR_j respectively.

[5] When using the PEMO-Q software, the options *delay compensation*, *pause cut* and *level alignment* are disabled since the signals to be compared are aligned and silence segments or gain distortion must be evaluated.

[6] PEMO-Q only handles single-channel signals. An extension to multichannel signals can be obtained by concatenating all channels into a single one.

Feature vector size	Task 1	Task 2	Task 3	Task 4
1	q_j^{overall}	q_j^{target}	q_j^{interf}	q_j^{artif}
2	-	q_j^{target} q_j^{artif}	-	q_j^{target} q_j^{artif}
3	q_j^{target} q_j^{interf} q_j^{artif}	q_j^{target} q_j^{interf} q_j^{artif}	q_j^{target} q_j^{interf} q_j^{artif}	q_j^{target} q_j^{interf} q_j^{artif}
4	q_j^{overall} q_j^{target} q_j^{interf} q_j^{artif}	q_j^{overall} q_j^{target} q_j^{interf} q_j^{artif}	q_j^{overall} q_j^{target} q_j^{interf} q_j^{artif}	q_j^{overall} q_j^{target} q_j^{interf} q_j^{artif}

TABLE III
FOR EACH TASK (COLUMN), SEVERAL CONFIGURATIONS OF FEATURE VECTORS ARE INVESTIGATED, DEPENDING ON THE NUMBER OF FEATURES (ROWS).

A. Training and test data and evaluation metrics

In order to account for performance bounds due to subject disagreement, we assess prediction performance with respect to the individual subjective scores. For each task r, let us denote by $\{y_{jrm}\}$ the set of subjective scores of all sounds j by all subjects m and by $\bar{y}_r$ its mean. For a given objective measure, we denote by x_{jrm} the prediction of y_{jrm}, which does not depend on m, and by $\bar{x}_r$ the mean of $\{x_{jrm}\}$. Each objective measure is evaluated via the following criteria, as defined in [36]:

- the prediction *accuracy* given by Pearson's linear correlation coefficient $\frac{\sum_{jm}(x_{jrm}-\bar{x}_r)(y_{jrm}-\bar{y}_r)}{\sqrt{\sum_{jm}(x_{jrm}-\bar{x}_r)^2}\sqrt{\sum_{jm}(y_{jrm}-\bar{y}_r)^2}}$,
- the prediction *monotonicity* given by Spearman's rank correlation, *i.e.* the linear correlation coefficient between n_{jrm}^x and n_{jrm}^y where n_{jrm}^x (resp. n_{jrm}^y) is the rank of x_{jrm} (resp. y_{jrm}) after sorting in ascending order,
- the prediction *consistency* given by $1 - R_o$, where the outlier ratio R_o is the proportion of sounds j and subjects m for which the prediction error $|x_{jrm} - y_{jrm}|$ is larger than twice the standard deviation of the subjective scores for that sound[7].

These criteria are expressed as real-valued figures between -1 and 1 or between 0 and 1.

The subjective scores collected in Section III were used both to train the neural network parameters and to test the resulting objective measures. We considered three cross-validation settings by splitting the data into a training set and a test set according to the 10 mixtures, to the 20 subjects or both. These settings did not significantly affect the trends of the results. In the following, we consider the most challenging setting aiming to predict the quality of a novel sound for an unknown subject. For each of 200 folds, the subjective scores of 19 subjects over 9 mixtures are used for training while testing is performed on the scores of the remaining subject over the remaining mixture. For each task and each feature vector, the number of sigmoids K was adjusted between 1 to 8 so as to maximize accuracy.

Note that a common way to evaluate the prediction performance consists in correlating objective measures with the mean opinion scores (MOS). We propose a more detailed analysis, involving MOS and individual ratings as in [24]: Table IV presents the main performance results when either the MOS or the individual ratings are used in the correlations, while only individual ratings are used in the subsequent detailed analysis. Table IV also shows the prediction performance depending whether the hidden reference and the anchors are used or not for evaluation – while training includes them in all cases. In the subsequent analysis, hidden reference and anchors are not taken into account in order to provide a realistic assessment over sounds from actual separation algorithms.

	Accuracy	Monotonicity
OPS	0.61 / 0.79 / 0.90	0.55 / 0.74 / 0.76
TPS	0.46 / 0.70 / 0.79	0.44 / 0.62 / 0.78
IPS	0.60 / 0.72 / 0.87	0.59 / 0.69 / 0.82
APS	0.43 / 0.71 / 0.87	0.43 / 0.71 / 0.85
SDR	0.37 / 0.50 / 0.85	0.36 / 0.48 / 0.63
ISR	−0.14 / −0.16 / 0.53	−0.07 / −0.07 / 0.35
SIR	0.72 / 0.85 / 0.94	0.67 / 0.79 / 0.88
SAR	0.31 / 0.52 / 0.88	0.31 / 0.55 / 0.84

TABLE IV
ACCURACY AND MONOTONICITY OF THE PROPOSED MEASURES VS. NONLINEARLY-MAPPED STATE-OF-THE-ART FEATURES COMPUTED WITH RESPECT TO: INDIVIDUAL SUBJECTIVE SCORES WITHOUT ANCHORS NOR REFERENCES (LEFT), MOS WITHOUT ANCHORS NOR REFERENCES (CENTER), OR MOS INCLUDING ANCHORS AND REFERENCES (RIGHT).

B. Choice of the decomposition parameters

As a preliminary experiment, we analyzed the performance of the OPS measure for the prediction of global quality as a function of the frame length T and the filter length L of the distortion decomposition algorithm. The results are reported in Table V for five different settings of T and L expressed in ms at 1 kHz. All performance figures exihibit very small variations on the order of ± 0.02. Thus, these parameters do not have a crucial influence. The optimal lengths corresponding to $T = 500$ ms and $L = 40$ ms at 1 kHz are used from now on.

T_{1kHz} (ms)	300	300	500	500	1000
L_{1kHz} (ms)	10	20	20	40	40
Accuracy	0.60	0.60	0.60	0.61	0.58
Monotonicity	0.57	0.56	0.56	0.55	0.55
Consistency	0.86	0.86	0.86	0.87	0.85

TABLE V
PERFORMANCE OF THE OPS MEASURE AS A FUNCTION OF THE FRAME LENGTH T AND THE FILTER LENGTH L OF THE DISTORTION DECOMPOSITION ALGORITHM EXPRESSED IN MS AT 1 KHZ.

With these settings, the minimum value of the component-wise salience features obtained by PEMO-Q was equal to 0.37, 0.76, 0.52 and 0.83 for q_j^{overall} ,q_j^{target} ,q_j^{interf} ,q_j^{artif}, respectively, while their maximum value was equal to 1.

C. Prediction of the global score with the OPS

Fig. 6 presents the main results regarding the assessment of global quality. Performance is analyzed as a function of the

[7] Note that outliers in the current section and in section III-B1 refer to different definitions.

chosen distortion decomposition algorithm, distortion salience metrics and feature vector configuration. The proposed OPS measure achieves the best performance in terms of accuracy, monotonicity and consistency and improves accuracy by more than 0.2 compared to nonlinear mapping of the SDR. The use of PEMO-Q instead of energy ratios results in a dramatic increase of accuracy of more than 0.1. This definitely validates the exploitation of auditory-based salience metrics. A smaller improvement on the order of 0.02 is observed when replacing the state-of-the-art decomposition [21] by the proposed one. Finally, accuracy improves by about 0.1 when using all four distortion salience features instead of a single feature corresponding to the overall distortion signal. This confirms that the decomposition of the distortion signal into several components is beneficial even for global quality assessment, given that listeners may associate a different weight to each kind of distortion. Nevertheless, the performance of the OPS remains somewhat below the upper performance bound corresponding to performance of the MOS, which suggests that room is left for future improvement.

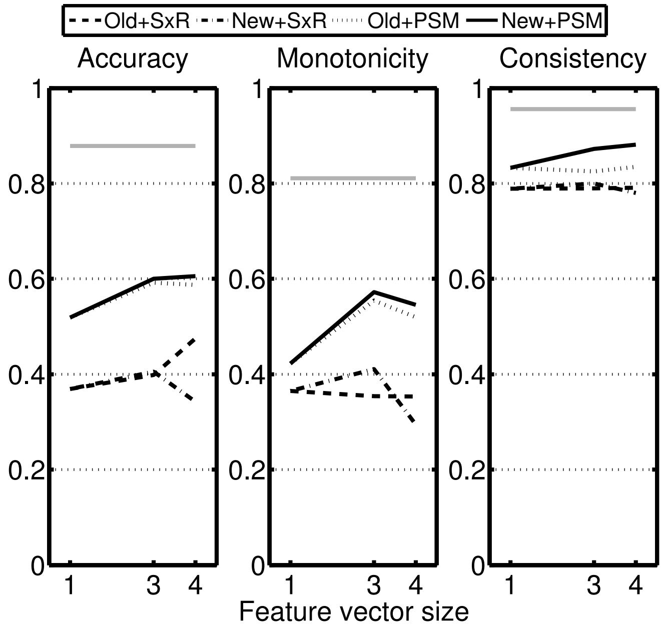


Fig. 6. Global score prediction performance as a function of the size of the feature vector, corresponding to different feature vector configurations shown in the first column of Table III. The four curves correspond to the use of the state-of-the-art [21] (Old) *vs.* the proposed (New) distortion decomposition algorithm and of energy ratio-based (SxR) *vs.* PEMO-Q-based (PSM) salience features. The solid curves corresponds to the proposed OPS measure. The gray lines indicates the upper performance bound corresponding to performance of the MOS compared to individual scores.

An insight into the neural network trained in the best configuration – *i.e.* four inputs and one sigmoid – is given in the first row of Table VI. Weights show that the influence of q_j^{interf} and q_j^{artif} is much larger than the influence of q_j^{overall} and q_j^{target} in the computation of OPS.

	v_{rk}	$\mathbf{w}_{rk}$				b_{rk}
OPS ($r = 1$)	1340.9	0.1	−0.2	6.3	5.7	−14.4
TPS ($r = 2$)	625.2	−0.7	8.4	0.8	−1.3	−9.1
	1210.1	34.2	−27.0	7.0	−10.2	−8.2
IPS ($r = 3$)	100.0	14.1				−11.9
APS ($r = 4$)	100.0	14.3				−14.5
	100.0	14.2				−14.3

TABLE VI
PARAMETERS OF THE NONLINEAR MAPPING FOR EACH OBJECTIVE MEASURE. EACH ROW OF A GIVEN CELL CORRESPONDS TO ONE SIGMOID k AND EACH COLUMN OF $\mathbf{w}_{rk}$ TO ONE INPUT FEATURE AS DEFINED IN TABLE III.

Objective measure	Accuracy	Monotonicity	Consistency
Energy ratio-based measures			
SNR [39]	0.41	0.42	0.72
Segmental SNR [39]	0.51	0.48	0.80
Freq.-wei. seg. SNR [39]	0.59	0.52	0.84
Spectral distance-based measures			
Itakura-Saito [39]	0.11	0.30	0.63
LLR [39]	0.39	0.46	0.76
Cepstrum dist. [39]	0.44	0.45	0.79
WSS [39]	0.46	0.46	0.78
Auditory-motivated measures			
PEAQ [23]	0.45	0.54	0.75
PESQ [37], [38]	0.54	0.48	0.83
Composite measures			
Fox et al. [15, p. 7]	0.23	0.20	0.34
Composite meas. [39]	0.61	0.56	0.83
OPS	**0.61**	**0.55**	**0.87**

TABLE VII
GLOBAL SCORE PREDICTION PERFORMANCE ACHIEVED BY VARIOUS STATE-OF-THE-ART AUDIO QUALITY MEASURES COMPARED TO THE PROPOSED OPS MEASURE.

Table VII further compares the proposed OPS measure with a number of objective measures for the evaluation of denoising or coding [37], [38], [23], [39] and source separation [15, p. 7] systems. These measures were scaled and shifted so as to ensure a fair evaluation of consistency. The OPS outperforms all concurrent measures. On average, conventional auditory-motivated measures do not perform better than energy ratio-based measures, while spectral distance-based measures perform worse despite their appropriateness for speech recognition [7]. The composite measure in [39] provides similar accuracy and monotonicity to the OPS but lower consistency, which indicates that the OPS generates fewer outlier values.

D. Prediction of specific scores with the TPS, IPS and APS

The results for the assessment of the preservation of the target source (task 2), the suppression of other sources (task 3) and the absence of additional artificial noise (task 4), are reported in Fig. 7. Performance is analyzed again as a function of the chosen distortion decomposition algorithm, distortion salience metrics and feature vector configuration.

As a general trend, quality assessment is improved by the proposed measures except for task 3, for which state-of-the-art measures performed already quite well. For all systems, the absolute values of accuracy and monotonicity for tasks 2 and 4 are about 0.1 to 0.2 lower than for tasks 1 and 3, which show that quality is more difficult to predict regarding the former. This is mostly due to the fact that a larger consensus between subjects was observed for the assessment of interference suppression than for that of target preservation and artificial noise.

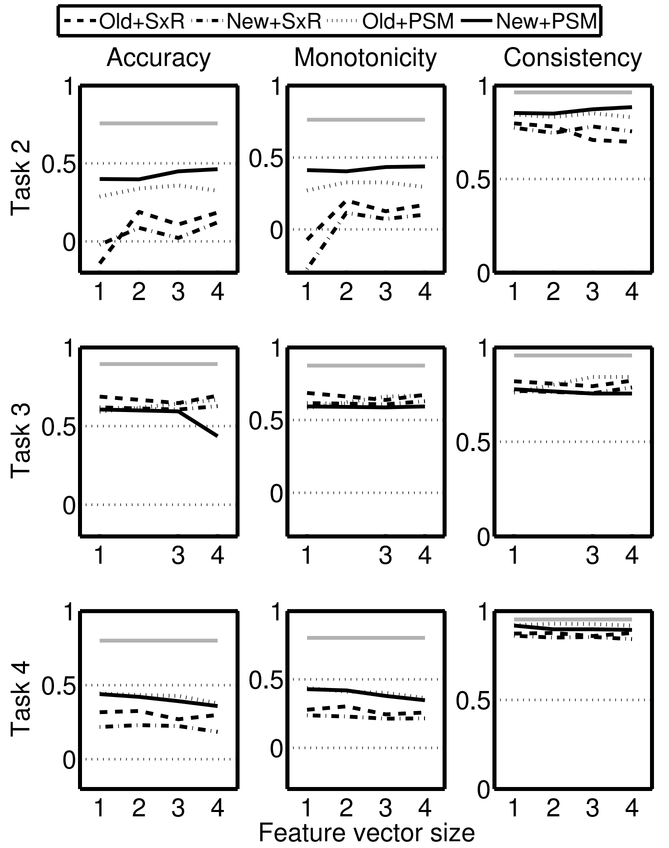


Fig. 7. Prediction performance of specific scores as a function of the size of the feature vector, corresponding to different feature vector configurations shown in Table III. The four curves correspond to the use of the state-of-the-art [21] (Old) *vs.* the proposed (New) distortion decomposition algorithm and of energy ratio-based (SxR) *vs.* PEMO-Q-based (PSM) salience features. The solid curves correspond to the proposed TPS, IPS and APS measures. The gray line indicates the upper performance bound corresponding to performance of the MOS.

The benefit of the proposed measures is important for task 2. The TPS using all four distortion salience features increases accuracy and monotonicity by 0.5 or more compared to nonlinear mapping of the ISR in [21]. Again, this is due both to the proposed distortion decomposition algorithm and to the use of PEMO-Q. The parameters of the neural network reported in Table VI show that one sigmoid mainly depends on q_j^{target} while the second one mainly depends on q_j^{overall}. By contrast, the best results for tasks 3 and 4 are obtained when using a single feature corresponding to the interference component or to the artifacts component respectively. The APS provides a smaller performance improvement on the order of 0.1 compared to nonlinear mapping of the SAR in [21] while the IPS provides small performance decrease compared to nonlinear mapping of the SIR in [21]. Note that an additional comparison showed that when using a single feature, the use of non-linear mapping or of the raw feature as a prediction measure do not change accuracy and monotonicity significantly. Hence, the main benefit of non-linear mapping comes from its ability to combine the features rather than to provide a non-linearity across each feature dimension.

Hence, the major benefit of the proposed measures concerns the assessment of target distortion rather than that of interference suppression and artificial noise. This can be explained by the fact that the salience of target distortion is badly assessed via an energy ratio due to the strong perceptual correlation between the target signal and the target distortion component. By contrast, interference and artifacts components are relatively independent from the target, so that fewer auditory masking effects arise.

VI. The PEASS toolkit

We released the subjective test guidelines and the Matlab listening test software of Section II, the 80 test sounds and the subjective scores of Section III and Matlab software implementing the OPS, TPS, IPS and APS measures proposed in Section IV as a toolkit called PEASS, standing for Perceptual Evaluation methods for Audio Source Separation[8]. All material is freely available under either the GNU Public License or Creative Commons licenses, except PEMO-Q which is free for academic use only. Among all system configurations tested in Section V, we select the one leading to the best accuracy for each of the four measures, retaining the coefficients reported in Table VI. This toolkit can be used both for the evaluation of existing and future audio source separation algorithm and for the training of future performance measures. This toolkit is also part of the evaluation measures used within the 2010 Signal Separation Evaluation Campaign (SiSEC) [32].

VII. Conclusion

We proposed a dedicated multi-criteria protocol for the subjective evaluation of audio source separation and a family of objective measures aiming to predict the resulting subjective scores. Four quality criteria were considered, namely global quality, preservation of the target source, suppression of other sources and absence of additional artificial noise. We collected a dabatase of 6400 subjective scores for a wide variety of mixtures and separation algorithms and showed that the proposed OPS, TPS, IPS and APS measures increase correlation with subjective scores up to 0.5 compared to nonlinear mapping of the individual state-of-the-art SDR, ISR, SIR and SAR source separation measures. These results show the benefit of a subband-based decomposition of the distortion signal into multiple components and of auditory-based methods for the assessment of the salience of each component, as well as the need of combining multiple salience features for the assessment of global quality and target distortion. While an FIR spatial and time distortion model was used in gammatone subbands, more results in the field of auditory scene analysis would be needed to design a truly auditory-based decomposition.

We hope that the proposed subjective test protocol could become the basis for a future improved standardized subjective test protocol. Also, 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

[8] http://bass-db.gforge.inria.fr/peass/

3D rendering, enabling the evaluation of advanced rendering attributes which cannot be accurately computed from the mixture today. The target signal to be estimated would then be the remix or the rendering of the true sources and the proposed decomposition procedure could 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.

REFERENCES

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

[2] S. Makino, T. Lee, and H. Sawada, Eds., *Blind speech separation*. Springer, 2007.

[3] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

[4] 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, 2010.

[5] E. Vincent, C. Févotte, R. Gribonval, X. Rodet, É. Le Carpentier *et al.*, "A tentative typology of audio source separation tasks," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA)*, 2003, pp. 715–720.

[6] D. P. W. Ellis, "Evaluating speech separation systems," in *Speech Separation by Humans and Machines*. Kluwer, 2004, ch. 20, pp. 295–304.

[7] L. D. Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578 – 2583, 2008.

[8] M. I. Mandel, S. Bressler, B. Shinn-Cunningham, and D. P. W. Ellis, "Evaluating source separation algorithms with reverberant speech," *IEEE Trans. Audio, Speech and Lang. Proces.*, vol. 18, no. 7, pp. 1872 –1883, sept. 2010.

[9] 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.

[10] Ö. 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.

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

[12] 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 *Proc. 2005 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2005, pp. 81–84.

[13] 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.

[14] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evaluation of audio source separation algorithms," in *Proc. UK ICA Research Network Workshop*, 2006.

[15] B. Fox, A. Sabin, B. Pardo, and A. Zopf, "Modeling perceptual similarity of audio signals for blind source separation evaluation," in *Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 454–461.

[16] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of subjective and objective evaluation methods for audio source separation," *Proceedings of Meetings on Acoustics*, vol. 4, no. 1, pp. 050 001–050 001–10, 2008.

[17] M. G. Jafari, E. Vincent, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "An adaptive stereo basis method for convolutive blind audio source separation," *Neurocomputing*, vol. 71, no. 10-12, pp. 2087–2097, 2008.

[18] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," in *Proc. of IWAENC*, Seattle, WA, USA, Sep. 2008.

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

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

[21] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, algorithms and results," in *Proc. 7th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2007, pp. 552–559.

[22] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *J. Audio Eng. Soc*, vol. 50, no. 5, pp. 331–342, 2002.

[23] ITU, "ITU-R Recommendation BS.1387-1: Method for objective measurements of perceived audio quality," 2001.

[24] T. Rohdenburg, V. Hohmann, and B. Kollmeier, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *Proc. 2005 Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2005, pp. 169–172.

[25] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech and Lang. Proces.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[26] T. Etame Etame, R. Le Bouquin Jeannès, C. Quinquis, L. Gros, and G. Faucon, "Towards a new reference system for subjective evaluation of coding techniques," in *Proc. 17th European Signal Processing Conf. (EUSIPCO)*, 2009, pp. 914–918.

[27] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical noise generation analysis for noise reduction methods based on spectral subtraction and mmse stsa estimation," in *Proc. of ICASSP*, apr. 2009, pp. 4433 –4436.

[28] Z. Goh, K.-C. Tan, and T. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," in *Proc. of ICASSP*, vol. 6, no. 3, May 1998, pp. 287 –292.

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

[30] E. Vincent, S. Araki, and P. Bofill, "The 2008 Signal Separation Evaluation Campaign: A community-based approach to large-scale evaluation," in *Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2009, pp. 734–741.

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

[32] S. Araki, A. Ozerov, V. Gowreesunker, H. Sawada, F. Theis *et al.*, "The 2010 Signal Separation Evaluation Campaign (SiSEC2010)– Part II–: Audio source separation challenges," in *Proc. 9th Int. Conf. on Independent Component Analysis and Signal Separation (ICA)*, 2010.

[33] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica*, vol. 88, no. 3, pp. 433–442, 2002.

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

[35] 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.

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

[37] ITU, "ITU-T Recommendation P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.

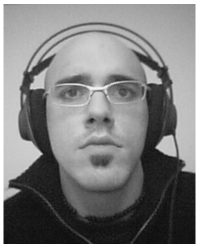
[38] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," *Proc. 2001 IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 749 – 752, 2001.

[39] P. Loizou, *Speech enhancement: theory and practice*. CRC press, Boca Raton, FL, USA, 2007.

Valentin Emiya graduated from Telecom Bretagne, Brest, France, in 2003 and received the M.Sc. degree in Acoustics, Signal Processing and Computer Science Applied to Music (ATIAM) at Ircam, France, in 2004. He received his Ph.D. degree in Signal Processing in 2008 at Telecom ParisTech, Paris, France. Since November 2008, he is a post-doctoral researcher with the METISS group at INRIA, Centre Inria Rennes - Bretagne Atlantique, Rennes, France.

His research interests focus on audio signal processing and include sound modeling and indexing, source separation, quality assessment and applications to music and speech.

Emmanuel Vincent (SM'10) received the mathematics degree of the ÃĽcole Normale SupÃľrieure, Paris, France, in 2001 and the Ph.D. degree in acoustics, signal processing and computer science applied to music from the University of Paris-VI Pierre et Marie Curie, Paris, in 2004. From 2004 to 2006, he has been a Research Assistant with the Centre for Digital Music at Queen Mary, University of London, London, U.K.. He is now a permanent researcher with the French National Institute for Research in Computer Science and Control (INRIA). His research focuses on probabilistic modeling of speech and music audio, applied to blind source separation and information retrieval.



Niklas Harlander received the Dipl.-Ing. (FH) degree in hearing technology and audiology from Oldenburg University of Applied Science, Oldenburg, Germany, in 2005 and the M.Sc. degree in hearing technology and audiology from the University of Oldenburg, Oldenburg, Germany, in 2007. He is currently working as a research associate at the Medical Physics department of the University of Oldenburg. His research interests include classification, noise reduction and perceptual quality estimation of audio signals.

Volker Hohmann received the Physics degree (Dipl.-Phys.) and the doctorat degree in physics (Dr. rer. nat.) from the University of Göttingen, Göttingen, Germany, in 1989 and 1993, respectively. Since 1993, he has been a Faculty Member of the Physics Institute, University of Oldenburg, Oldenburg, Germany, and member of the Medical Physics Group (Prof. B. Kollmeier). He has been active in teaching activities in physics for undergraduate and graduate courses and has research expertise in acoustics and digital signal processing with applications to signal processing in speech processing devices, e.g., hearing aids. He is a Consultant with the Hörzentrum Oldenburg GmbH. He was a Guest Researcher at Boston University, Boston, MA, (Prof. Dr. Colburn) in 2000 and at the Technical University of Catalonia, Barcelona, Spain in 2008. Dr. Hohmann received the Lothar-Cremer price of the German acoustical society (DEGA) in 2008.