Improved perceptual metrics for the evaluation of audio source separation

Emmanuel Vincent

Improved perceptual metrics for the evaluation of audio source separation

Emmanuel Vincent

INRIA, Centre de Rennes - Bretagne Atlantique
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuel.vincent@inria.fr

Abstract. We aim to predict the perceived quality of estimated source signals in the context of audio source separation. Recently, we proposed a set of metrics called PEASS that consist of three computation steps: decomposition of the estimation error into three components, measurement of the salience of each component via the PEMO-Q auditory-motivated measure, and combination of these saliences via a nonlinear mapping trained on subjective opinion scores. The parameters of the decomposition were shown to have little influence on the prediction performance. In this paper, we evaluate the impact of the parameters of PEMO-Q and the nonlinear mapping on the prediction performance. By selecting the optimal parameters, we improve the average correlation with mean opinion scores (MOS) from 0.738 to 0.909 in a cross-validation setting. The resulting improved metrics are used in the context of the 2011 Signal Separation Evaluation Campaign (SiSEC).

Keywords: audio source separation, objective evaluation, PEASS

1 Introduction

Audio source separation is the task of extracting the signal of each sound source from a given mixture. In a number of applications such as speech enhancement for hearing aids or denoising of old music recordings, the separation performance amounts to the subjective judgment of listeners.

A popular set of performance metrics can be obtained by decomposing the estimation error into three components, namely *target distortion*, *interference* and *artifacts*, and measuring the salience of these components via energy ratios termed signal to distortion ratio (SDR), source image to spatial distortion ratio (ISR), signal to interference ratio (SIR) and signal to artifacts ratio (SAR) [11]. Despite the wide use of the associated BSS Eval toolbox[1], *e.g.* within the annual Signal Separation Evaluation Campaign (SiSEC) [11], these metrics are known to poorly correlate with subjective performance for certain mixtures involving *e.g.* low-frequency sounds or time-varying distortion. Two different routes have

[1] http://bass-db.gforge.inria.fr/bss_eval/

been taken to increase correlation: assessing the overall distortion via auditory-motivated measures such as PEAQ [9] or PEMO-Q [8], or combining energy ratios via linear or nonlinear mappings trained on subjective opinion scores [4].

In [3], we combined these two routes via a three-step procedure consisting of

1. decomposing the estimation error into target distortion, interference and artifacts components,
2. assessing the salience of each component via PEMO-Q,
3. combining these saliences via trained nonlinear mappings.

We distributed the resulting metrics termed overall perceptual score (OPS), target-related perceptual score (TPS), interference-related perceptual score (IPS) and artifacts-related perceptual score (APS), as the version 1.0 of a toolkit called PEASS[2]. Each of the above three steps involves one or more design parameters. In [3], we showed that the parameters of the first step have little influence on the prediction performance. In this paper, we evaluate the impact of the parameters of the two latter steps and select the optimal parameters maximizing the correlation with mean opinion scores (MOS). The resulting improved metrics are distributed as the version 2.0 of PEASS and used among others for the evaluation of the algorithms submitted to SiSEC 2011.

The structure of the rest of the paper is as follows. In Section 2, we summarize the computation of the PEASS metrics and highlight the parameters involved in each step. In Section 3, we describe the evaluation protocol and show the effect of each parameter on the prediction performance. We conclude in Section 4.

2 The PEASS metrics

For a given set of separated sources, we aim to predict the perceived quality of the estimated multichannel *spatial image* $\widehat{\mathbf{s}}_j(t)$ of each source j, *i.e.* its contribution to all mixture channels, relatively to the true spatial image $\mathbf{s}_j(t)$ [11]. The PEASS metrics [3] involve three computation steps outlined in the introduction. In the following, we summarize each step with a focus on the two latter steps, including the internal computations of PEMO-Q which were not detailed in [3].

2.1 Distortion decomposition

In the first step, the estimation error $\widehat{\mathbf{s}}_j(t) - \mathbf{s}_j(t)$ is split into three components: target distortion $\mathbf{e}_j^{\text{target}}(t)$, interference $\mathbf{e}_j^{\text{interf}}(t)$ and artifacts $\mathbf{e}_j^{\text{artif}}(t)$ such that

$$\widehat{\mathbf{s}}_j(t) - \mathbf{s}_{ij}(t) = \mathbf{e}_j^{\text{target}}(t) + \mathbf{e}_j^{\text{interf}}(t) + \mathbf{e}_j^{\text{artif}}(t). \quad (1)$$

This is achieved by passing the signals through a bank of gammatone filters [6], partitioning the output into overlapping time frames, performing decomposition (1) in each subband and each time frame by least-squares projection onto the subspaces spanned by delayed versions of the true source spatial image signals, and reconstructing time-domain signals by filterbank inversion. Compared with BSS Eval, this step aims to improve the handling of time-varying distortion.

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

2.2 PEMO-Q component saliences

In the second step, the perceptual salience of these components is assessed as

$$q_j^{\mathrm{o}} = \text{PEMO-Q}(\widehat{\mathbf{s}}_j, \mathbf{s}_j) \tag{2}$$

$$q_j^{\mathrm{t}} = \text{PEMO-Q}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\mathrm{target}}) \tag{3}$$

$$q_j^{\mathrm{i}} = \text{PEMO-Q}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\mathrm{interf}}) \tag{4}$$

$$q_j^{\mathrm{a}} = \text{PEMO-Q}(\widehat{\mathbf{s}}_j, \widehat{\mathbf{s}}_j - \mathbf{e}_j^{\mathrm{artif}}) \tag{5}$$

where PEMO-Q$(\widehat{\mathbf{x}}, \mathbf{x}) \in [-1, 1]$ is the *perceptual similarity* measured by PEMO-Q between a test signal $\widehat{\mathbf{x}}$ and a reference signal $\mathbf{x}$. Compared with BSS Eval, this step accounts for auditory masking and dynamic compression phenomena.

PEMO-Q first computes *internal auditory representations* $\widehat{X}_i$ and X_i of each channel i of $\widehat{\mathbf{x}}$ and $\mathbf{x}$ via the computational auditory model in [2,1]. This model comes in two versions and consists of:

- R1 subband decomposition via a bank of gammatone filters linearly spaced on the equivalent rectangular bandwidth (ERB) scale between $f_{\min}$ and $f_{\max}$,
- R2 for each subband, halfwave rectification, first-order autoregressive (AR) lowpass filtering with 1 kHz cutoff, and summation with a threshold a_{thresh},
- R3 amplitude compression by five consecutive nonlinear feedback loops emphasizing rapid changes up to a maximum amplitude ratio of $r_{\max}$ for each loop,
- R4 either first-order AR lowpass filtering with 8 Hz cutoff (*lowpass version* [2]) or decomposition via a bank of eight first-order AR bandpass filters with center frequencies ranging from 0 to 129 Hz (*modulation version* [1]).

This mimics the effect of haircells in the inner ear and modulation processing in the auditory cortex. The outputs $\widehat{X}_i$ and X_i are either two-dimensional time-frequency representations for the lowpass version or three-dimensional time-frequency-rate representations for the modulation version.

The perceptual similarity between $\widehat{X}_i$ and X_i is then measured by [7,8]

- S1 partial assimilation of the two representations in each time-frequency-rate bin (t, f, m) as $\widehat{X}_{itfm} \leftarrow \alpha X_{itfm} + (1-\alpha)\widehat{X}_{itfm}$ if $|\widehat{X}_{itfm}| < |X_{itfm}|$,
- S2 computation of the time-varying linear cross-correlation between $\widehat{X}_i$ and X_i over time frames of length l_{corr}[3],
- S3 computation of the time-varying root mean square (RMS) amplitude of X_i over time frames of length l_{amp},
- S4 computation of the p-th percentile of the cross-correlation series weighted by the RMS amplitude.

This attempts to model the perception of global similarity based on the local similarities between the signals. Finally, the overall scalar similarity PEMO-Q$(\widehat{\mathbf{x}}, \mathbf{x})$ is selected as the minimum of the channel-wise similarity over all channels i.

[3] A slightly distinct processing is applied in the modulation version. See [8] for details.

2.3 Trained nonlinear mapping

In the third step, the saliences in (2)–(5) are combined by [3]

M1 optional log-mapping from $[-1, 1]$ to $\mathbb{R}$ via $q_j^k \leftarrow \log((1+q_j^k)/(1-q_j^k))$ [7],
M2 selection of one or more saliences forming a *feature vector* $\mathbf{q}_j$,
M3 transformation into a scalar objective score via a feedforward neural network (NN) [5] composed of n_{lay} layers of n_{neur} neurons trained on subjective scores.

Compared with BSS Eval, this accounts for the different perceptual importance of each distortion component by which artifacts may be heard as more disturbing than interference for instance.

Four different perceptual assessment *tasks* were considered in [3]: global quality, preservation of the target source, suppression of other sources, and absence of additional artificial noise. For each task, a different feature vector was selected and a different NN was trained by minimizing the RMS error between the predicted and the actual subjective opinion scores. This resulted in four metrics called OPS, TPS, IPS and APS, respectively.

3 Effect of the design parameters

3.1 Data and evaluation procedure

Each processing block from R1 to M3 involves some design parameters listed above. In order to evaluate their effect on the prediction performance, we consider the set of 6400 subjective scores collected in [3] using the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) protocol [10]. For each of 10 mixtures and each of the four tasks listed in Section 2.3, 20 subjects were asked to score 8 test sounds, including 4 real-world sounds produced by actual source separation algorithms, one *hidden reference* and 3 *anchors*. The scoring scale ranges from 0 to 100, where larger means better. The anchors are artificial sounds with low quality ensuring that the whole scale is used. For information about the variance of subjective scores and outliers, see [3]. In order to avoid overfitting, a 200-fold cross-validation procedure is used. For each fold, the 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. The prediction *accuracy* is assessed via the linear correlation between the predicted scores and the MOS.

3.2 Main results

The version 1.0 of PEASS relies on the following default parameters of PEMO-Q: modulation version, $f_{\min} = 235$ Hz, $f_{\max} = 14500$ Hz, $a_{\text{thresh}} = 10^{-5}$, $r_{\max} = +\infty$, $\alpha = 0.5$, $l_{\text{corr}} = +\infty$ and $l_{\text{amp}} = +\infty$[4]. The mapping consists of a 1.5-layer NN[5] without input log-mapping. For each mixture and subject, all 8 test sounds

[4] p is irrelevant here due to the use of global correlation ($l_{\text{corr}} = +\infty$).
[5] This term refers to a 2-layer NN with linear output layer.

Table 1. Accuracy after successive parameter optimization stages.

Optimization stage	OPS	TPS	IPS	APS	Average
Baseline (version 1.0)	0.799	0.396	0.860	0.896	0.738
Optimal mapping and PEMO-Q version	0.909	0.815	**0.934**	0.870	0.882
Optimal PEMO-Q similarity measure	**0.925**	0.812	0.931	0.924	0.898
Optimal PEMO-Q internal representation	0.922	**0.864**	0.926	**0.925**	**0.909**

were used for training but only the 4 real-world sounds for testing. The best feature vector among 3 or 4 candidates and the best number of neurons were then selected so as to maximize accuracy over the test set [3].

In subsequent experiments, we found this approach to be unsuitable for two reasons. First, the absence of references and anchors in the test set resulted in objective metrics that do not span the whole range from 0 to 100 and thus fail to handle better or poorer sounds than those in that set. Second, the 10 references in the training set drew the NN to better fit scores close to 100 instead of uniformly fitting all scores. In order to avoid these drawbacks, we adopt a consistent approach from now on, whereby all real-world sounds and anchors but only one reference are employed in each training and testing fold. The resulting baseline performance of version 1.0 is displayed in the top row of Table 1.

Due to the large number of design parameters, we optimize these parameters in three successive stages, from higher-level to lower-level ones. For simplicity and computational efficiency, the same parameters are used for all four metrics, except the optimal feature vector and number of neurons which depend on the metric. The resulting performance after each stage is shown in the bottom three lines of Table 1. On average, the accuracy improves from 0.738 to 0.909 when combining all three stages. This huge improvement is mostly due to the optimization of higher-level parameters in the first stage, while the two other stages have less impact. We analyze each stage in more details in the following.

3.3 Detailed impact of the mapping and the version of PEMO-Q

The top half of Table 2 describes the effect of the number of neurons n_{neur} and the feature vectors. By simply selecting the optimal n_{neur} (first row) and features (second row), we greatly improve the performance of the TPS and significantly improve that of the three other metrics, resulting in an average accuracy of 0.868. This is a direct consequence of the consistent training approach discussed above, but also of the fact that all possible feature vectors are tested here. Indeed, none of the optimal feature vectors belongs to the list of candidate vectors previously tested in [3].

Table 3 describes the effect of the other parameters of the nonlinear mapping and the version of PEMO-Q. The use of a 2-layer NN with input log-mapping along with the lowpass version of PEMO-Q appears optimal for all metrics except the APS and yields an optimal average accuracy of 0.882. The corresponding feature vectors are shown in the bottom line of Table 2.

Table 2. Accuracy as a function of the feature vectors and of the baseline or the optimal mapping and version of PEMO-Q, assuming optimal number of neurons and default PEMO-Q parameters.

Mapping and version	Feature vector	OPS	TPS	IPS	APS	Average
baseline	baseline	0.799	0.710	0.860	0.905	0.819
	optimal	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{a}}]$ 0.871	$[q_j^{\mathrm{t}}\ q_j^{\mathrm{i}}]$ 0.747	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{i}}\ q_j^{\mathrm{a}}]$ 0.935	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{t}}\ q_j^{\mathrm{a}}]$ 0.920	0.868
optimal	baseline	0.901	0.801	0.865	0.834	0.850
	optimal	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{a}}]$ 0.909	$[q_j^{\mathrm{t}}\ q_j^{\mathrm{i}}\ q_j^{\mathrm{a}}]$ 0.815	$[q_j^{\mathrm{o}}\ q_j^{\mathrm{i}}\ q_j^{\mathrm{a}}]$ 0.934	$[q_j^{\mathrm{t}}\ q_j^{\mathrm{a}}]$ 0.870	**0.882**

Table 3. Accuracy as a function of the version of PEMO-Q, the optional log-mapping and the number of NN layers n_{lay}, assuming optimal feature vectors and numbers of neurons in each case and default PEMO-Q parameters.

Version	Log-mapping	n_{lay}	OPS	TPS	IPS	APS	Average
filterbank	no	1.5	0.871	0.747	0.935	0.920	0.868
		2	0.877	0.759	0.912	**0.924**	0.868
	yes	1.5	0.884	0.784	0.928	0.916	0.878
		2	0.884	0.761	0.926	0.909	0.870
lowpass	no	1.5	0.886	0.794	0.940	0.869	0.872
		2	0.877	0.788	0.919	0.878	0.866
	yes	1.5	0.903	0.775	**0.939**	0.839	0.864
		2	**0.909**	**0.815**	0.934	0.870	**0.882**

3.4 Detailed impact of the PEMO-Q similarity measure

After fixing the optimal mapping and version of PEMO-Q, we consider the parameters of the PEMO-Q similarity measure in a second stage. The effect of each parameter is illustrated in Figure 1. Among the tested values, the average accuracy appears to increase with p and decrease with α and l_{corr}. This effect is particularly significant for the APS, which may be due to the nonstationary nature of artifacts calling for local rather than global correlation between the reference and the test representation. The optimal values are $\alpha = 0.25$, $l_{\mathrm{corr}} = 100$ ms, $l_{\mathrm{amp}} = 1$ s and $p = 0.5$, yielding an average accuracy of 0.898.

3.5 Detailed impact of the PEMO-Q internal representation

After fixing the optimal parameters of the similarity measure, we consider the parameters of the internal representation in a last stage. The effect of each parameter is illustrated in Figure 2. Among the tested values, the average accuracy appears to increase with f_{min} and r_{max} and decrease with f_{max} and a_{thresh}. This effect is significant for all metrics except the IPS. The optimal parameters are the default f_{min}, f_{max} and r_{max} along with $a_{\mathrm{thresh}} = 10^{-6}$, yielding an average accuracy of 0.909.

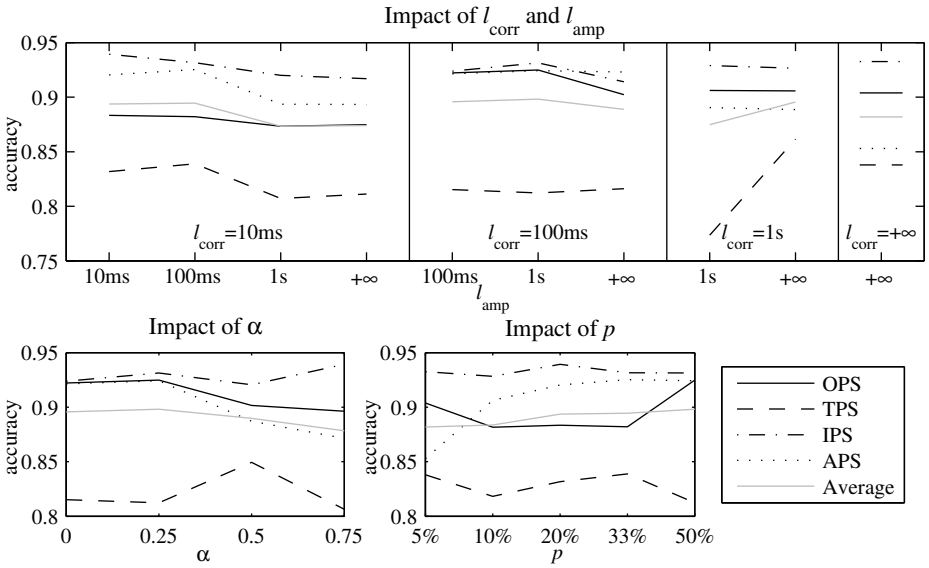

Fig. 1. Accuracy as a function of one of the three parameters (l_{corr}, l_{amp}), α and p given the optimal values of the two other parameters, assuming optimal mapping and version of PEMO-Q and default PEMO-Q internal representation. Note that infinite durations are equivalent to 10 s here, since the duration of the test signals is 5 s.

4 Conclusion and perspectives

We examined the impact of various design parameters over the accuracy of the PEASS metrics. By adopting a consistent training approach together with unconstrained feature selection, we improved the accuracy from 0.738 to 0.868 in a cross-validation setting. By optimizing the parameters of PEMO-Q and the nonlinear mapping, we further increased it to 0.909. These results show that the mapping from the error component saliences to the metrics is crucial, while fine tuning of auditory parameters has smaller impact. The resulting improved metrics have been released as version 2.0 of PEASS and used within SiSEC 2011.

Acknowledgment

This work was supported by the EUREKA Eurostars i3Dmusic project funded by Oseo. We would like to thank Rainer Huber, Volker Hohmann and Valentin Emiya for discussions about this work.

References

1. Dau, T., Kollmeier, B., Kohlrausch, A.: Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc.

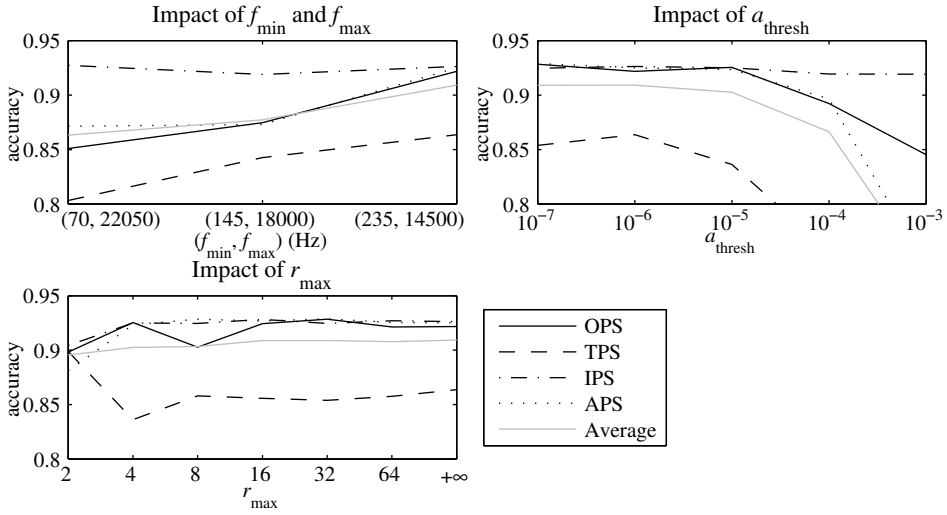

Fig. 2. Accuracy as a function of one of the three parameters ($f_{\min}$, $f_{\max}$), a_{thresh} and $r_{\max}$ given the optimal values of the two other parameters, assuming optimal mapping and version of PEMO-Q and PEMO-Q similarity metric.

Am. 102(5), 2892–2905 (Nov 1997)

2. Dau, T., Püschel, D., Kohlrausch, A.: A quantitative model of the "effective" signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Am. 99(6), 3615–3622 (Jun 1996)
3. Emiya, V., Vincent, E., Harlander, N., Hohmann, V.: Subjective and objective quality assessment of audio source separation. IEEE Trans. Audio Speech Lang. Process. 19(7), 2046–2057 (Sep 2011)
4. Fox, B., Pardo, B.: Towards a model of perceived quality of blind audio source separation. In: Proc. Int. Conf. on Multimedia Expo (ICME). pp. 1898–1901 (2007)
5. Haykin, S.: Neural Networks. Prentice Hall (1999)
6. Hohmann, V.: Frequency analysis and synthesis using a gammatone filterbank. Acta Acustica 88(3), 433–442 (2002)
7. Huber, R.: Objective assessment of audio quality using an auditory processing model. Ph.D. thesis, University of Oldenburg (Dec 2003)
8. Huber, R., Kollmeier, B.: PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception. IEEE Trans. Audio Speech Lang. Process. 14(6), 1902–1911 (Nov 2006)
9. ITU: ITU-R Recommendation BS.1387-1: Method for objective measurements of perceived audio quality (2001)
10. ITU: ITU-R Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems (2003)
11. Vincent, E., Araki, S., Theis, F.J., Nolte, G., Bofill, P., Sawada, H., Ozerov, A., Gowreesunker, B.V., Lutter, D., Duong, N.Q.K.: The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. Signal Processing (to appear)