INFORMED SEPARATION OF SPATIAL IMAGES OF STEREO MUSIC RECORDINGS USING LOW-ORDER STATISTICS

Stanislaw Gorlow[*]

Univ. Bordeaux
LaBRI, UMR 5800
33400 Talence, France

Sylvain Marchand

Univ. Brest
Lab-STICC — CNRS, UMR 6285
29238 Brest, France

ABSTRACT

In this work we address a reverse audio engineering problem, i.e. the separation of stereo tracks of professionally produced music recordings. More precisely, we apply a spatial filtering approach with a quadratic constraint using an explicit source-image-mixture model. The model parameters are “learned” from a given set of original stereo tracks, reduced in size and used afterwards to demix the desired tracks in best possible quality from a preexisting mixture. Our approach implicates a side-information rate of 10 kbps per source or channel and has a low computational complexity. The results obtained for the SiSEC 2013 dataset are intended to be used as reference for comparison with other approaches.

Index Terms— Informed source separation, low-order statistics, professionally produced music recordings, spatial filtering, stereo images

1. INTRODUCTION

Most if not all of today’s professionally produced music has undergone two basic processes: *mixing* and *mastering*. Many established music distribution formats, moreover, are strictly stereo. While in the mastering stage the final mix is prepared and transfered to a data storage device, mixing represents the process that ends up in a summation of individually recorded and edited audio from distinct mono or stereo sources into a composite stereo mixture. The apparent placement of sources between the speakers in a stereo sound field is also known as “imaging” [1] in professional audio engineering. The notion of spatial “images” in a source separation context can e.g. be found in [2]. The separation of stereo images of individual or grouped sources is the central point of the present paper.

For the reason that the total number of source channels is usually greater than the number of mixture channels, mixing is mathematically *underdetermined*. So, demixing constitutes an *ill-posed* source separation problem that cannot be solved without additional assumptions or prior information. We use the knowledge of the mixing process and low-order statistics of the sources as additional information for our algorithm in order to find the optimal solution. The content of the paper is therefore an extension to our previous work on the informed separation of mono sources [3]. We introduce a more general source-image signal model based on common studio practice and also generalize the mixture model to a sum of images of mono and stereo sources. The demixing problem is likewise addressed in an *informed* source separation context [4]. With the proposed approach one can decompose the final mix into distinct tracks or into the background and foreground objects and in the same manner one can separate the vocal from the instrumental track for karaoke.

The organization of the paper is as follows. The problem at hand is given in Section 2. Section 3 illustrates the source-image signal model, the estimation of model parameters, and how the latter can be reduced in size. The extended mixture model is discussed in Section 4. There it is also shown how a source of interest and its image are separated using a linearly constrained spatial filter. The proposed approach is evaluated on five multitracks of changing sound complexity in Section 5. Section 6 concludes the paper with an outlook.

2. PROBLEM STATEMENT

The problem at hand is stated as follows. Given access to the original stereo images of distinct sources, recover a subset of the images in best possible quality from a mixture composed of the original images using a source-image-mixture model. The model parameters shall be estimated from the accessible image signals and used during recovery. The amount of data associated with the model parameters should furthermore be kept to a minimum.

3. PARAMETRIC ANALYSIS

3.1. Source-image signal model

We model the signals in the complex subband domain. Each subband signal is said to be a zero-mean circular symmetric complex Gaussian stochastic process that evolves over time

[*] This research was partially funded by the “Agence Nationale de la Recherche” within the scope of the DReaM project (ANR-09-CORD-006).

n. The set of subband signal components at a given instant n of a single source is deemed to be mutually independent, and so is the set of sources. The sources are thus uncorrelated. A source may be mono or stereo. The two channels of a stereo source are considered pairwise independent, i.e. uncorrelated too, as if a stereo source was comprised of *two* sources: one mono source in the left channel and another mono source in the right channel. A stereo source can thus be thought of as a *centered* spatial image of an acoustic source that is recorded with two independent microphones.

A mono source is assigned a location in the stereo sound field via *pan* control, whereas a stereo source or its centered image is positioned via *balance* control:

$$\begin{aligned} \mathbf{u}_i(n) &= a_{il}\mathbf{e}_l s_{il}(n) + a_{ir}\mathbf{e}_r s_{ir}(n) \\ &= \mathbf{a}_i \circ \mathbf{s}_i(n), \end{aligned} \tag{1}$$

where $\circ$ denotes the Hadamard or entrywise product between the time-invariant steering vector $\mathbf{a}_i = \left[\, a_{il} \; a_{ir} \,\right]^{\mathsf{T}}$ and the ith stereo source $\mathbf{s}_i = \left[\, s_{il} \; s_{ir} \,\right]^{\mathsf{T}}$. In the case of a mono source, $s_{il} = s_{ir} = s_i$. Accordingly, $\mathbf{u}_i$ represents the stereo image of the ith source. In (1), $\{\mathbf{e}_l, \mathbf{e}_r\}$ is the standard basis of $\mathbb{R}^2$, $\mathbf{a}_i \in \mathbb{R}^2$ and $\mathbf{s}_i(n) \in \mathbb{C}^2$. The subband index k is omitted for simplicity. The ith steering vector $\mathbf{a}_i$ is defined as

$$\mathbf{a}_i \triangleq \frac{\mathbf{a}_i'}{\|\mathbf{a}_i'\|} = \begin{bmatrix} \sin\theta_i \\ \cos\theta_i \end{bmatrix} \tag{2}$$

in the case of a mono source, or else as

$$\mathbf{a}_i \triangleq \frac{\mathbf{a}_i'}{a_{i,\mathrm{ref}}'}, \tag{3}$$

where

$$a_{i,\mathrm{ref}}' = \begin{cases} a_{il}' & \text{if } a_{il}' \geqslant a_{ir}' \\ a_{ir}' & \text{otherwise} \end{cases}, \tag{4}$$

in the case of a stereo source. In the above equations, superscript $'$ indicates unnormalized items.

3.2. Model parameter estimation

Consider the stereo image of a distinct source as given. From the stereo signal, we can estimate the model parameters that are used as prior information for source separation, which is detailed in Section 4. These parameters describe the source's location and its instantaneous variance distribution. We apply the following protocol.

First, we compute the zero-lag cross-covariance between the left and the right channel and normalize it by the product of average power in each channel using the root mean square (RMS) as measure:

$$\operatorname{corr}\left[u_{il}(n), u_{ir}(n)\right] = \frac{\operatorname{cov}\left[u_{il}(n), u_{ir}(n)\right]}{\mathrm{RMS}_{il}(n)\mathrm{RMS}_{ir}(n)}. \tag{5}$$

In our case, corr is identical with Pearson's correlation. Thus when $\operatorname{var}_n\left\{\operatorname{corr}\left[u_{il}(n), u_{ir}(n)\right]\right\} \to 0$, the associated source is considered as mono and the pan angle is estimated as

$$\hat{\theta}_i = \operatorname{arccot}\frac{\sum_n \mathrm{RMS}_{ir}(n)}{\sum_n \mathrm{RMS}_{il}(n)}, \tag{6}$$

where arccot is the arccotangent and var is the variance. In the reverse case, when the source is stereo, its power balance is estimated as

$$\hat{a}_{i,\neg\mathrm{ref}} = \frac{\sum_n \mathrm{RMS}_{i,\neg\mathrm{ref}}(n)}{\sum_n \mathrm{RMS}_{i,\mathrm{ref}}(n)} \qquad \text{with } a_{i,\mathrm{ref}} = 1, \tag{7}$$

where $\mathrm{ref} \in \{l, r\}$ is the channel with the greater RMS value and $\neg\mathrm{ref}$ is the complementary channel.

The instantaneous variance distribution of the ith source signal over subbands is given by

$$\phi_{ijk}(n) = \mathrm{E}\left[\left|s_{ijk}(n)\right|^2\right], \tag{8}$$

where E is the expectation and $\left\{\phi_{ijk}(n)\right\}_k$ is the short-time power spectral density (STPSD) of the jth channel at instant n. For a mono source, $\phi_{ijk} = \phi_{ik}$. The STPSD is estimated using the short-time Fourier transform (STFT) according to

$$\hat{\phi}_{ijk}(n) = \left|S_{ij}(k,n)\right|^2 \qquad \text{with } j \in \{l, r\}, \tag{9}$$

or $\hat{\phi}_{ik}(n) = \left|S_i(k,n)\right|^2$, where $\left\{S_{il}(k,n), S_{ir}(k,n)\right\}_k$ is the spectrum of a stereo source and $\left\{S_i(k,n)\right\}_k$ the spectrum of a mono source, respectively. Individual spectra are obtained from the corresponding images by

$$S_i(k,n) = \begin{bmatrix} \sin\hat{\theta}_i & \cos\hat{\theta}_i \end{bmatrix} \mathbf{u}_{ik}(n) \tag{10}$$

in the case of a mono source, or else by

$$S_{i,\mathrm{ref}}(k,n) = u_{i,\mathrm{ref},k}(n) \tag{11a}$$

and

$$S_{i,\neg\mathrm{ref}}(k,n) = \begin{cases} \dfrac{u_{i,\neg\mathrm{ref},k}(n)}{\hat{a}_{i,\neg\mathrm{ref}}} & \text{if } \hat{a}_{i,\neg\mathrm{ref}} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{11b}$$

in the case of a stereo source.

3.3. Parameter quantization and coding

The pan angle θ of a mono source is rounded to the nearest integer value using a mid-tread uniform quantizer defined as

$$Q(x) = \Delta \cdot \left\lfloor \frac{x}{\Delta} + \frac{1}{2} \right\rfloor, \tag{12}$$

where Δ is the step size and $\lfloor\cdot\rfloor$ represents the floor function. The balance parameter $a_{i,\neg\mathrm{ref}}$ for a stereo source is encoded using an A-law or μ-law compressor together with a uniform quantizer as in (12). For a given input $x \in [0, 1]$, the A-law compressor output is

$$C_A(x) = \begin{cases} \dfrac{A \cdot x}{1 + \log A} & \text{if } x < \frac{1}{A} \\ \dfrac{1 + \log(A \cdot x)}{1 + \log A} & \text{otherwise} \end{cases}, \tag{13}$$

where A is the compression parameter and $\log$ is the natural logarithm. The output of the μ-law compressor is

$$C_\mu(x) = \frac{\log(1 + \mu \cdot x)}{\log(1 + \mu)}, \tag{14}$$

where μ is the associated compression parameter. Using A-law or μ-law compression, the signal-to-noise ratio (SNR) is kept constant over the entire range of $a_{i,\neg\mathrm{ref}}$ [5]. The STPSD of a mono source or a channel of a stereo source is quantized on an ERB-like frequency scale according to

$$\bar{\phi}_{iz}(n) = \frac{1}{\mathrm{ub}(z) - \mathrm{lb}(z) + 1} \sum_{k=\mathrm{lb}(z)}^{\mathrm{ub}(z)} \hat{\phi}_{ik}(n), \tag{15}$$

where $\mathrm{lb}(z) = \inf\{k \mid z_k = z\}$, $\mathrm{ub}(z) = \sup\{k \mid z_k = z\}$, z is the quantization index and

$$z_k = \lfloor 21.4 \log_{10}(4.37 f_s/Nk + 1) \rfloor. \tag{16}$$

In (16), f_s is the sampling rate and N is the STFT size. The average power values are then converted from linear scale to logarithmic scale and quantized using (12). These values are encoded using differential pulse-code modulation (DPCM) in combination with Huffman coding. The difference between adjacent power values is calculated in the direction of time or frequency or between channel pairs depending on what gives the lowest entropy.

4. SEPARATION OF STEREO IMAGES

4.1. Mixture model and spatial covariance matrix

The mixture is considered to be obtained by superposition of distinct stereo images that were created according to (1). To account for professionally produced music recordings, $\mathbf{s}_i(n)$ is regarded as having undergone prior processing in the form of linear and nonlinear audio effects [6]. The mixture signal is thus formulated as

$$\begin{aligned} \mathbf{x}_k(n) &= \sum_{i \in \mathsf{I}} \mathbf{a}_i \circ \mathbf{s}_{ik}(n) \\ &= \sum_{p \in \mathsf{P}} \mathbf{a}_p \cdot s_{pk}(n) + \sum_{q \in \mathsf{Q}} \mathbf{a}_q \circ \mathbf{s}_{qk}(n), \end{aligned} \tag{17}$$

where set $\mathsf{P} = \{i \in \mathsf{I} \mid \forall n[s_{ilk}(n) = s_{irk}(n)]\}$ represents the mono sources, while $\mathsf{Q} = \{i \in \mathsf{I} \mid \exists n[s_{ilk}(n) \neq s_{irk}(n)]\} = \mathsf{I} \setminus \mathsf{P}$ represents the stereo sources, respectively.

The local mixture spatial covariance matrix is given by

$$\begin{aligned} \mathbf{R}_{\mathbf{xx},k}(n) &= \mathrm{E}\left[\mathbf{x}_k(n)\mathbf{x}_k^{\mathsf{H}}(n)\right] \\ &= \sum_{p \in \mathsf{P}} \mathbf{a}_p \mathbf{a}_p^{\mathsf{T}} \cdot \phi_{pk}(n) \\ &\quad + \sum_{q \in \mathsf{Q}} \mathbf{a}_q \mathbf{a}_q^{\mathsf{T}} \circ \mathbf{\Phi}_{qk}(n), \end{aligned} \tag{18}$$

where $\{\phi_{pk}(n)\}_k$ is the pth mono source's STPSD and

$$\mathbf{\Phi}_{qk}(n) = \begin{bmatrix} \phi_{q,ll,k}(n) & \phi_{q,lr,k}(n) \\ \phi^*_{q,lr,k}(n) & \phi_{q,rr,k}(n) \end{bmatrix}, \tag{19}$$

where $*$ denotes complex conjugation. In (19), $\{\phi_{q,ll,k}(n)\}_k$ and $\{\phi_{q,rr,k}(n)\}_k$ are the qth stereo source's left- and right-channel STPSDs, while $\{\phi_{q,lr,k}(n)\}_k$ is the short-time cross spectral density (STCSD). Due to the assumed independence of the left and the right channel of a stereo source, $\phi_{q,lr,k}(n)$ is zero for all q, k and n, so that

$$\mathbf{\Phi}_{qk}(n) = \mathrm{diag}\,[\phi_{qlk}(n), \phi_{qrk}(n)]. \tag{20}$$

Using (18) and (20), the mixture spatial covariance matrix is reconstructed from the mixing coefficients and the STPSDs. In the following it is shown how the summation of the stereo images can be "undone" by means of spatial filtering.

4.2. Image separation of a mono source

Let us assume that there are more than two active sources in a time-frequency (TF) point (k, n). In this case, a mono source component is separated from the mixture signal with the aid of the power-conserving minimum-variance (PCMV) spatial filter [3]

$$\hat{\mathbf{w}}_{pk}(n) = \mathbf{R}^{-1}_{\mathbf{xx},k}(n)\mathbf{a}_p \sqrt{\frac{\phi_{pk}(n)}{\mathbf{a}_p^{\mathsf{T}} \mathbf{R}^{-1}_{\mathbf{xx},k}(n)\mathbf{a}_p}} \tag{21}$$

according to

$$\hat{s}_{pk}(n) = \hat{\mathbf{w}}_{pk}^{\mathsf{T}}(n)\mathbf{x}_k(n). \tag{22}$$

The corresponding image component estimate is

$$\hat{\mathbf{u}}_{pk}(n) = \mathbf{a}_p \cdot \hat{s}_{pk}(n). \tag{23}$$

If the number of active sources is at most two, the demixing becomes trivial given that the mixing system is known.

4.3. Image separation of a stereo source

A stereo source component is separated from the mixture in a similar manner, where the left-channel and the right-channel components are estimated simultaneously according to

$$\hat{\mathbf{s}}_{qk}(n) = \hat{\mathbf{W}}_{qk}^{\mathsf{T}}(n)\mathbf{x}_k(n) \tag{24}$$

with the PCMV spatial filter matrix being

$$\hat{\mathbf{W}}_{qk}(n) = \mathbf{R}^{-1}_{\mathbf{xx},k}(n)\boldsymbol{\Phi}^{1/2}_{qk}(n) \cdot \mathrm{diag}\left\{\left[\mathbf{R}^{-1}_{\mathbf{xx},k}(n)\right]_{ll}, \left[\mathbf{R}^{-1}_{\mathbf{xx},k}(n)\right]_{rr}\right\}^{-1/2}. \tag{25}$$

On the analogy of (23), the corresponding image component estimate is given by

$$\hat{\mathbf{u}}_{qk}(n) = \mathbf{a}_q \circ \hat{\mathbf{s}}_{qk}(n). \tag{26}$$

From (20) and (25), it can be seen that when multiple stereo sources are present in the mixture, their component estimates exhibit the same phase between different sources. Only their spectral envelopes are shaped differently. Furthermore, when the mixture is a combination of stereo sources only, $\hat{\mathbf{W}}_{qk}(n)$ in (25) is diagonal. As a result, $s_{qjk}(n)$ is separated from the respective mixture channel using the mono PCMV filter:

$$\hat{s}_{qjk}(n) = \sqrt{\frac{\phi_{qjk}(n)}{\sum_{i\in \mathrm{I}} a^2_{ij}\phi_{ijk}(n)}}\, x_{jk}(n). \tag{27}$$

5. PERFORMANCE EVALUATION

In this section, we evaluate our approach by applying it to a subset of professionally produced music recordings from the SiSEC 2013 [7] dataset. The task is to decompose an artistic mixture into a subset of constituent images that represent the sources of interest alias foreground objects and the image of the background—where applicable. The term “background” refers to the sum of background objects. The original images are given as a reference.

5.1. Performance metrics

We use the evaluation criteria suggested by the SiSEC 2013 committee. These include the performance metrics from the PEASS toolkit [8, 9] and the decoder runtime in seconds per CPU clock rate in GHz. For every mixture, we also give the side-information rate. Furthermore, we include PEMO-Q [10, 11] in our evaluation.

5.2. Experimental design

We use the following testing framework. With respect to the STFT, we employ a 2048-point fast Fourier transform (FFT) with a Kaiser-Bessel derived window of the same length and a 50-% overlap between succeeding frames. The pan angle $\hat{\theta}$ is quantized and coded with 7 bits, while the balance $\hat{a}_{\neg\mathrm{ref}}$ is quantized with 16 bits using the A-law compander with an A of 87.6. The STPSD is quantized with 6 bits per power value using a 76-band nonuniform frequency scale. The probability mass function of the difference between contiguous STPSD values is modeled with a Laplace (μ, b) distribution with $\hat{\mu} = -0.2$ and $\hat{b} = 2$. The simulations are run in MATLAB on an Intel Core i5-520M 2.4-GHz CPU.

5.3. Experimental results

The results of the experiment are summarized in Table 1. As can be observed, the image-to-spatial distortion ratio (ISR) is between 6.66 and 17.4 dB for a stereo source, and is greater or equal to 18.5 dB for a mono source. Similarly, the highest source-to-artifacts ratio (SAR) is obtained for a mono source, which is 27.7 dB. The target-related perceptual score (TPS) shows a weak correlation not only with both the ISR and the SAR, but also with PEMO-Q’s perceptual similarity measure PSM_t, which then again does not take spatial hearing effects into account. The lowest TPS is at 52 %. The measured side-information rate is around 10 kbps per mono source or stereo channel. The execution time of the algorithm is low and also faster than real time.

6. CONCLUSION AND OUTLOOK

In this paper we presented an extension to our previous work on the informed separation of audio sources. By generalizing the mixture model to a sum of stereo images, we have shown how a particular source of interest or its image can be filtered out from a stereo mixture using prior information. From our source-image model we inferred that the pursued approach is most effective when the foreground objects are in mono and only the background object is in stereo. This was validated by the fact that the best separation results with regard to the ISR were obtained for mono sources. The observed inconsistency between different performance metrics makes listening tests indispensable, however.

The sound clips that were used in the experiment can be downloaded from http://www.labri.fr/˜gorlow/eusipco13/. A comparison with other approaches will be made available on the SiSEC 2013 website.

7. REFERENCES

[1] D. Gibson, *The Art of Mixing: A Visual Guide to Recording, Engineering, and Production*. 236 Georgia Street, Vallejo, CA 94590, USA: MixBooks, LLC, 1997, ch. 2.

[2] N. Q. K. Duong, E. Vincent, and R. Gribonval, “Underdetermined reverberant audio source separation using a full-rank spatial covariance model,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, Sep. 2010.

[3] S. Gorlow and S. Marchand, “Informed audio source separation using linearly constrained spatial filters,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 1, pp. 3–13, Jan. 2013.

[4] K. H. Knuth, “Informed source separation: A Bayesian tutorial,” in *Proc. EUSIPCO*, 2005, pp. 1–8.

Track	Type	$\hat{\theta}$ or $\hat{a}_{\neg\mathrm{ref}}$	SDR	ISR	SIR	SAR	OPS	TPS	IPS	APS	PSM$_t$	ODG
Vocal	stereo	0.89	9.76	**16.8**	11.5	21.7	0.38	**0.61**	0.68	0.79	**0.76**	−2.96
Drums	stereo	1.00	8.72	**12.4**	13.3	19.5	0.25	**0.86**	0.66	0.05	**0.34**	−3.30
Guitar	stereo	0.96	9.26	**16.3**	10.1	23.4	0.34	**0.52**	0.47	0.67	**0.76**	−2.97
"The Ones We Love" by Another Dreamer — 59.6 kbps — 10.6 s GHz												
Vocal	stereo	0.99	8.35	**17.1**	9.31	20.9	0.19	**0.54**	0.62	0.86	**0.74**	−3.00
Bass	mono	45.0	8.60	$\boxed{\mathbf{24.2}}$	8.82	$\boxed{27.7}$	0.38	**0.62**	0.52	0.34	**0.54**	−3.21
Piano	stereo	0.83	3.11	**6.92**	4.14	17.4	0.44	**0.63**	0.51	0.60	**0.80**	−2.88
Background	stereo	0.94	4.74	**8.33**	8.17	18.1	0.47	**0.60**	0.58	0.59	**0.69**	−3.07
"Roads" by Bearlin — 69.8 kbps — 7.4 s GHz												
Vocal	stereo	0.90	9.15	**15.5**	10.8	19.5	0.76	**0.62**	0.86	0.68	**0.81**	−2.82
Drums	stereo	0.99	5.15	**6.66**	7.07	15.2	0.27	**0.79**	0.64	0.10	**0.40**	−3.28
Bass	mono	45.0	5.59	$\boxed{\mathbf{18.5}}$	5.24	$\boxed{21.6}$	0.30	**0.80**	0.47	0.07	**−0.10**	−3.38
Claps	stereo	0.99	8.92	**13.8**	11.9	20.6	0.05	**0.96**	0.67	0.00	**−0.03**	−3.37
Background	stereo	0.97	4.76	**10.6**	5.80	14.9	0.46	**0.62**	0.51	0.60	**0.72**	−3.03
"Remember the Name" by Fort Minor — 82.2 kbps — 13.0 s GHz												
Vocal	mono	47.9	14.5	$\boxed{\mathbf{23.0}}$	15.9	$\boxed{27.6}$	0.53	**0.56**	0.88	0.87	**0.85**	−2.66
Guitar	stereo	0.97	14.8	**17.4**	20.5	27.1	0.56	**0.98**	0.77	0.81	**0.88**	−2.53
"Que Pena/Tanto Faz" by Tamy — 31.8 kbps — 5.8 s GHz												
Vocal	stereo	1.00	6.77	**14.5**	7.48	20.2	0.63	**0.72**	0.77	0.56	**0.76**	−2.96
Drums	stereo	0.97	8.39	**14.6**	10.2	19.9	0.49	**0.82**	0.66	0.34	**0.53**	−3.22
Bass	stereo	0.93	5.22	**11.9**	5.87	16.3	0.32	**0.53**	0.52	0.30	**0.24**	−3.32
Background	stereo	0.93	4.61	**11.8**	4.79	17.7	0.40	**0.61**	0.59	0.70	**0.77**	−2.94
Ultimate NZ Tour — 80.7 kbps — 8.4 s GHz												

Table 1. The SiSEC 2013 data subset used in the experiment and the obtained results. Framed are the ISR and SAR values for mono sources.

[5] ITU-T, *Pulse code modulation (PCM) of voice frequencies*, 1993, rec. ITU-T G.711.

[6] N. Sturmel *et al.*, "Linear mixing models for active listening of music productions in realistic studio conditions," in *AES Convention 132*, Apr. 2012.

[7] "SiSEC 2013," http://sisec.wiki.irisa.fr/tiki-index.php, Feb. 2013.

[8] "The PEASS Toolkit – Perceptual Evaluation methods for Audio Source Separation," http://bass-db.gforge.inria.fr/peass/, Feb. 2013, version 2.0.

[9] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

[10] HörTech gGmbH, "PEMO-Q," http://www.hoertech.de/web_en/produkte/pemo-q.shtml, version 1.3.

[11] 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 Lang. Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.