EXAMINING THE PERCEPTUAL EFFECT OF ALTERNATIVE OBJECTIVE FUNCTIONS FOR DEEP LEARNING BASED MUSIC SOURCE SEPARATION

Stylianos Ioannis Mimilakis[*], *Estefanía Cano*[*], *Derry FitzGerald*[†],
Konstantinos Drossos[‡], and Gerald Schuller[◇]

[*]Fraunhofer IDMT, Ilmenau, Germany
[†]AudioSourceRE, Cork, Ireland
[‡]Laboratory of Signal Processing, Tampere University of Technology, Tampere, Finland
[◇]Technical University of Ilmenau, Ilmenau, Germany

ABSTRACT

In this study, we examine the effect of various objective functions used to optimize the recently proposed deep learning architecture for singing voice separation MaD - Masker and Denoiser. The parameters of the MaD architecture are optimized using an objective function that contains a reconstruction criterion between predicted and true magnitude spectra of the singing voice, and a regularization term. We examine various reconstruction criteria such as the generalized Kullback-Leibler, mean squared error, and noise to mask ratio. We also explore recently proposed, for optimizing MaD, regularization terms such as sparsity and TwinNetwork regularization. Results from both objective assessment and listening tests suggest that the TwinNetwork regularization results in improved singing voice separation quality.

Index Terms— music source separation, deep learning, noise-to-mask ratio, perceptual evaluation

1. INTRODUCTION

Audio and music source separation aims at estimating individual audio signals from an observed mixture. An important task in music source separation is the estimation of singing voice and background music [1], with a range of applications spanning from music information retrieval [2] to music repurposing [3, 4]. To that aim, supervised approaches, and specifically approaches based on deep learning, have shown to provide state-of-the-art results [5, 6].

The most adopted solution for deep learning based music source separation relies on the supervised training of deep neural networks, and requires three components: i) the mixture signal which is given as an input to the deep neural network, ii) the target source signal, and iii) the objective function that will compare the estimated and the true target source signals and, if applicable, introduce a penalty with respect to the deep neural networks parameters and/or latent variables. The first two components are usually included in publicly available datasets. For an overview of datasets in music source separation interested readers are kindly referred to [1]. The choice or construction of an objective function comes experimentally by assessing objectively the source separation performance of the deep learning model, using the signal-to-distortion-ratio (SDR) and signal-to-interference-ratio (SIR) metrics [7].

For instance, the work presented in [8] investigates the effect of five objective functions that are commonly used in matrix factorization based music source separation [9] and for the task of multi-channel audio source separation. The results suggest that the generalized Kullback-Leibler (KL) divergence and the mean-squared-error (MSE) distance perform equally well with respect to the SDR and SIR metrics. The work presented in [10] proposes the modification of the binary cross-entropy cost function, commonly used to optimize deep neural networks for classification tasks, so that perceptually relevant time-frequency masks can be approximated by the deep neural networks for the task of speech separation. Similarly, in [11] it is proposed to use a weighting scheme for the cost function under the presence of various types of noise, for the task of speech enhancement. The weighting scheme employs an auditory-perceptual model in order to highlight specific time-frequency regions that are of high importance during training. An intelligibility objective was introduced in [12] for the task of low-latency speech separation, which has shown to outperform the MSE objective. Similarly, the study presented in [13] proposes to use a combination of the intelligibility objective employed in [12], and the SDR-SIR objective metrics presented in [7]. The results from the listening tests conducted in [13] suggest that higher suppression of interfering sources can be achieved by employing the SIR metric into the optimization, while the combination of the SDR and the intelligibility objective increases the quality of source separation.

In contrast to the above mentioned signal processing oriented objective functions, other approaches have been proposed. For instance, the work in [14] examines the objective functions that are used to train a deep clustering network in

speech separation. Specifically, it is shown in [14] that the graph Laplacian distance and the whitened k-means objectives introduce improvements to the deep clustering model, in terms of SDR. The work presented in [15] proposes to use the KL plus a sparsity penalty as the objective function for the Masker and Denoiser (MaD) deep learning architecture for singing voice separation. An improvement of that approach is presented in [16] where the TwinNetwork regularization [17] is used to regularize the recurrent neural networks of the MaD architecture yielding state-of-the-art (SOTA) results in monaural singing voice separation.

In this work we focus on the MaD architecture that for monaural singing voice separation[15, 16], and examine the perceptual effect that the previously mentioned objective functions have upon the separation quality of singing voice. More specifically, we optimize the MaD architecture using the KL objective with the TwinNetwork regularization [16, 17] and the L_1 sparsity regularization [15]. For comparison we also optimize MaD without regularization, using the MSE and a perceptually motivated objective based on the noise to mask ratio (NMR) introduced in [18] and used in the context of low-rank approximation for music signals in [19]. Each cost objective is evaluated using the SDR objective metric that has been initially proposed in [7] and modified for the separation campaign presented in [5]. Furthermore, we subjectively assess the perceptual effect that the objective functions have by means of listening tests using experienced listeners through the usage of MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) standard using the framework presented in [20] and anchor synthesis procedure from [21].

The remainder of this paper is structured as follows. Section 2 presents the MaD architecture. Section 3 introduces the objective functions used to train MaD. The experimental setup is described in Section 4, followed by the results in Section 5. Section 6 concludes this work.

2. MASKER AND DENOISER ARCHITECTURE

The MaD architecture operates in the time-frequency domain and uses the magnitude spectrogram of a mixture signal $|\mathbf{Y}| \in \mathbb{R}_{\geq 0}^{M \times N}$ to predict the magnitude spectrogram of the singing voice, $|\mathbf{Y^s}|$. The MaD architecture consists of two components, the Masker and the Denoiser. The goal of the Masker is to predict a time-frequency mask that when applied to $|\mathbf{Y}|$, it will produce a first estimate of the target source (i.e. the singing voice), denoted as $|\hat{\mathbf{Y}}^s|$. The Denoiser is responsible for enhancing the result of the time-frequency masking operation performed by the Masker, resulting in a better estimate of the singing voice magnitude spectra [16, 15]. An illustration of MaD is given in Figure 1.

More specifically, the Masker uses a bi-directional recurrent neural network (RNN) encoder (RNN_{enc}), that encodes $|\mathbf{Y}|$ by iterating over the time dimension. The output of

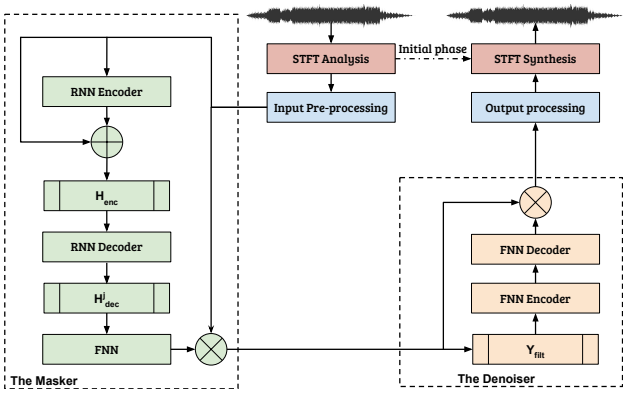

Fig. 1: Illustration of the MaD architecture.

RNN_{enc} is updated by employing residual connections between the output of RNN_{enc} and $|\mathbf{Y}|$. The result is the latent variables $\mathbf{H}_{\text{enc}}$ that are then used as an input to a decoding function that is implemented using a forward RNN decoder, RNN_{dec}. The output of RNN_{dec}, $\mathbf{H}_{\text{dec}}$, is then given to a sparsifying transformation. This transformation is implemented using a feed-forward neural network $\text{FNN}_{\text{M}}$and a rectified linear unit (ReLU), resulting in computing a filter with non-negative values for each time-frequency sub-band, $\mathbf{M}^s$. $\mathbf{M}^s$ is used to filter the mixture $|\mathbf{Y}|$. Formally, the first estimate of the singing voice spectrogram $|\hat{\mathbf{Y}}^s|$ is obtained using:

$$|\hat{\mathbf{Y}}^s| = \mathbf{M}^s \odot |\mathbf{Y}|, \text{ where} \tag{1}$$

$\odot$ is the Hadamard product and the superscript s denotes the source specific indexing for the singing voice. For clarity we will denote the overall function of the Masker for estimating $|\hat{\mathbf{Y}}^s|$ as $\mathcal{M}(\cdot)$, i.e.:

$$|\hat{\mathbf{Y}}^s| = \mathcal{M}(|\mathbf{Y}|) \tag{2}$$

$|\hat{\mathbf{Y}}^s|$ contains interference from the rest of existing music sources in the original mixture, $|\mathbf{Y}|$ [15, 22]. MaD encompasses another trainable module after the Masker, aiming at suppressing those interferences, and (therefore) called the Denoiser. The Denoiser is a denoising autoencoder (DAE), consisting of two feed-forward layers, the encoder FNN_{enc} and the decoder FNN_{dec}. FNN_{enc} and FNN_{dec} are using ReLU as an activation function and the input to the FNN_{enc} is $|\hat{\mathbf{Y}}^s|$. The output of the FNN_{dec} is used to multiply element-wise the $|\hat{\mathbf{Y}}^s|$ by employing the skip-filtering connections, yielding the final estimate of the singing voice magnitude spectrogram by the MaD architecture, $|\hat{\mathbf{Y}}'^s|$. The function implemented by the Denoiser is denoted as $\mathcal{D}(\cdot)$, i.e.:

$$|\hat{\mathbf{Y}}'^s| = \mathcal{D}(|\hat{\mathbf{Y}}^s|). \tag{3}$$

3. OBJECTIVE FUNCTIONS

The objective of the two functions (i.e. for the Masker and Denoiser) respectively is defined as [22]:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_{\mathcal{M}}(\mathcal{M}(|\mathbf{Y}|), |\mathbf{Y}^s|) + \mathcal{L}_{\mathcal{D}}(\mathcal{D}(|\hat{\mathbf{Y}}^s|), |\mathbf{Y}^s|) + \lambda\Omega(\mathcal{M}). \quad (4)$$

With a small abuse of notation for brevity, Ω is a regularization function that operates on the Masker and is scaled by λ which is a factor that controls the strength of the regularization. For the MSE between true $\mathbf{X}$ and predicted $\hat{\mathbf{X}}$ matrices, the loss terms for the Masker and the Denoiser respectively in Eq(4) become:

$$\mathcal{L}_{\text{MSE}}(\mathbf{X}, \hat{\mathbf{X}}) = \frac{1}{MN}||\mathbf{X} - \hat{\mathbf{X}}||_2^2, \quad (5)$$

where $||\cdot||_2^2$ is the squared Frobenious norm. Similarly for the KL divergence between two matrices we have:

$$\mathcal{L}_{\text{KL}}(\mathbf{X}, \hat{\mathbf{X}}) = ||\mathbf{X} \odot \log(\frac{\mathbf{X} + \epsilon}{\hat{\mathbf{X}} + \epsilon}) - \mathbf{X} + \hat{\mathbf{X}}||, \quad (6)$$

where log is the element-wise logarithmic function ,$||\cdot||$ is the $L1$ matrix norm, and ϵ is a small value to ensure numerical stability. Finally, the NMR is defined for matrices that contain magnitude spectral information as:

$$\mathcal{L}_{\text{NMR}}(\mathbf{X}, \hat{\mathbf{X}}) = 10\log_{10}(\frac{1}{MN}||(\mathbf{X} - \hat{\mathbf{X}})^{\odot 2} \odot \mathcal{T}(\mathbf{X})|| + \epsilon), \quad (7)$$

where $\log_{10}$ is the element-wise log base 10 operation, $.^{\odot 2}$ denotes the element-wise exponentiation to the power of 2, and $\mathcal{T}(\mathbf{X})$ is a pre-defined, non-trainable function that computes the *energy* of the inverse masking threshold of the target variable $\mathbf{X}$. The computation of the inverse masking threshold is based on the RASTA model [23], for computing the Bark-scale frequency sub-bands necessary for computing the masking threshold, and on the non-linear superposition approach presented in [24] with the exponential non-linearity set to 0.9. We used the RASTA model because an approximation of the pseudo-inverse can be expressed for the Bark-scaling operation. The pseudo-inverse is used to approximate the masking-threshold at the original dimensionality of the magnitude spectra, allowing the optimization to be performed without further operators (e.g. dimensionality reduction) that will bias the results with respect to the other described cost functions. The non-linear superposition approach is used to compute the masking threshold instead of other approaches, because it provides a fast approximation of the tonal masking without requiring the computation of the tonality of magnitude spectra [18, 24].

Focusing on the regularization term, we examine a very common operator for inducing sparsity over the parameters of the model and commonly used in music source separation for avoiding over-fitting [8]. More specifically and as proposed in [15, 22], we penalize the weights of $\mathbf{W}_{\text{FNN}_\text{M}}$ as:

$$\Omega_{L1}(\mathcal{M}) := ||\mathbf{W}_{\text{FNN}_\text{M}}||. \quad (8)$$

As the Masker relies mostly on recurrent neural networks, we examine the TwinNetwork regularization that have been proposed in the context of music source separation in [16] and is the most recent and reasonably robust method for regularizing recurrent neural networks [17]. TwinNetwork regularization uses the hidden states of a backward RNN to regularize the hidden states of a forward RNN [17], while both of the RNNs are trained to minimize the same cost, enforcing the forward RNN to take into account the future evolution of the signal. Let $\mathbf{H}_{\text{enc}}$ be the output of the RNN_{enc} that is served both to the RNN_{dec}, yielding $\mathbf{H}_{\text{dec}}$, and to the TwinNetwork that outputs $\mathbf{H}_{\text{twin}}$, the regularization is defined as

$$\Omega_{\text{twin}}(\mathcal{M}) := \mathcal{L}_{\text{twin}} = \sum_t ||\psi(\mathbf{h}_{\text{dec}_t}) - \mathbf{h}_{\text{twin}_t}||_2, \quad (9)$$

where ψ is a trainable affine transform that allows small perturbations to norm of errors, and $\mathbf{h}_{*_t}$ is the hidden state vector of the corresponding RNN at the time-state t. All of the above operations are used jointly to train the MaD architecture.

4. EXPERIMENTAL SETUP

To train the MaD architecture and cost functions, we use the 100 two-channel multi-tracks from the MUSDB18 [25] dataset that have been used in the SiSEC 2018 campaign [5]. The multi-track recordings are sampled at 44100 Hz. For each multi-track we use the mixture and target source (singing voice) in the dataset, and average the two available channels. The magnitude spectra are computed by using the STFT analysis for each mixture and the corresponding singing voice signal. The analysis uses a hamming windowing 46 ms long, a factor of 2 for zero-padding, a hop-size of 8.7 ms. For the NMR computation 24 Bark-bands are used. All the model parameters are initialized and optimized according to the proposed methodology presented in [15, 16]. To ensure a fair comparison between the objective functions, the same random seed is used to initialize the parameters of the MaD in each training scheme with the corresponding objective functions. The λ parameter for controlling the sparse regularization (Eq. 8) is equal to $1e-4$ and for the TwinNet regularization (Eq. 9) equal to 0.5. The values were chosen experimentally according to the balance between the range of values for the reconstruction and regularization terms, following [15, 16]. We assess source separation performance with objective metrics SDR and SIR expressed in dB as proposed in the 2018 SiSEC campaign [5] for the evaluation sub-set of the corresponding dataset [25] comprising of 50 additional multi-tracks.

A MUSHRA listening test as defined in [26] was conducted to validate whether perceptual quality improvements

could be observed when using any of the objective functions for separation. A total of 7 participants (after post-screening) with previous experience in audio signal processing participated in the listening test. A selection of six 10-second segments from the evaluation sub-set of MUSDB18 data-set was used as test content in the listening test. The test content was chosen such that it includes (3) male and (3) female singers, ensuring timbre and genre diversity. The anchor signals were generated to resemble distortions produced by source separation algorithms, following the procedure presented in [27]. The participants of the listening test were asked to rate *general separation quality* as formally described in [21]. All participants were asked to rate the test content in a continuous quality scale from 0 to 100, divided into 5 equal intervals with the following quality descriptors: [0-20] bad, [20-40] poor, [40-60] fair, [60-80] good, and [80-100] excellent.

5. RESULTS & DISCUSSION

The results from the objective assessment using the SDR metric are demonstrated in Fig. 2 with box-plots showing the median values of each objective function.

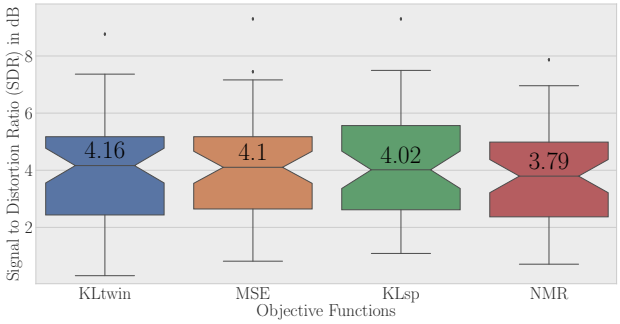

Fig. 2: Boxplot showing the analysis of variance of the SDR metric in dB for the TwinNetwork regularization (KLtwin), mean-squared-error objective (MSE), sparsity regularization (KLsp), and noise-to-mask ratio (NMR). Horizontal lines denote median values also displayed in bold-faced numbers. The whiskers denote the minimum and maximum SDR values.

As can be observed in Fig.2 the TwinNetwork regularization combined with KL (KLtwin), provides a marginal performance increase of 0.06 dB and 0.14 dB with respect to the MSE and the sparse-aware (KLsp) objective functions, respectively. Additionally, the SDR ratings place the NMR as the lowest-performing objective function, with KLsp showing the largest interquartile range IQR, implying a larger degree of variability in the results.

To facilitate a comparison of the results between SDR and mean opinion scores (MOS) obtained with the listening tests, barplots showing mean values with 95% confidence intervals are also given in Fig. 3 and Fig. 4 for the MOS and SDR, respectively. As can be seen in Fig. 3, the four objective functions obtained MOS values between 40 and 60, suggesting fair separation quality according to the MUSHRA specifications. In contrast to the marginal improvement seen for KLTwin in the SDR results, MOS from the subjective listening tests show that this marginal difference in SDR values has an effect upon auditory quality perception. Specifically, the KLTwin received a MOS of 54.49, which is in average, approximately 6 points higher than MSE and KLsp. In addition, KLTwin also outperformed the perceptually motivated objective NMR by 4.3 points of MOS, while depicting small variability in both MOS and SDR, according to Fig. 3 and Fig. 4, respectively. It should be noted that although the NMR received the lowest objective median SDR of 3.79 dB, it was rated as the second best objective function in the listening tests (see Fig. 3). Even when the number of participants in the listening test is not sufficiently high to argue significant differences between objective and subjective evaluation scores, this difference follows similar observations presented in [27], where listening tests at larger scale depict that subjective performance not always coincides with the objective metrics such as SDR.

Focusing on the computational costs during training, the MSE objective involves only the computation a quadratic loss, making it less computationally expensive compared to Kullback-Leibler that involves the computation of the logarithmic function, and element-wise matrix multiplications and divisions. On the other hand, TwinNetwork regularization requires additional training parameters which can significantly increase the training time required for optimizing the MaD architecture. However, the noise-to-mask ratio requires the computation of the inverse masking threshold of the target source magnitude spectra which not only significantly increases the training time but also increases the necessary data passed to the optimization.

From a simple optimization perspective, the additional data obtained from the inverse masking threshold partake in

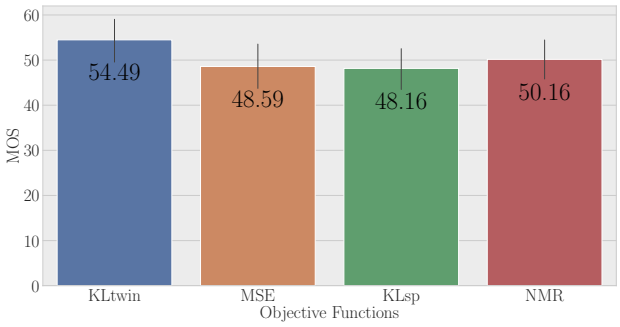

Fig. 3: Mean opinion scores MOS obtained with the listening tests. Mean values with 95% confidence intervals are shown for the four objective functions.

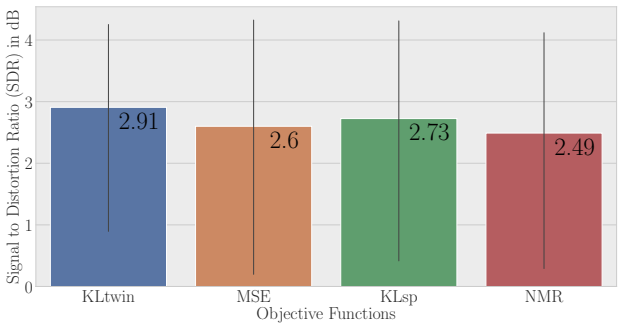

Fig. 4: Barplots showing mean SDR values with 95% confidence intervals for the four objective functions.

training via the element-wise multiplication of the squared errors, which are then passed to the 10-based logarithmic function (Eq.7). By applying the chain rule for computing the partial derivatives with respect to the model parameters, it can be observed that the curvature of the NMR function suffers from plateaus close to *local minima*. A set of other *minima* (not strictly *global* as the NMR function can have many roots), that yield a low cost in the NRM sense, are at non-smooth areas of that curvature making the NMR objective less useful towards optimizing the corresponding parameters of MaD. Based on the perceptual differences in the results, the TwinNetwork regularization appears to be able to provide a better minima, resulting in a perceptually relevant increase of separation performance.

6. CONCLUSIONS

In this study we examined from both objective and subjective stand points, the performance of objective functions used to train the MaD architecture for singing voice separation. Although this study presented a small-scale listening test, the mean opinion scores obtained exhibit noticeable trends between the objective functions and the performed evaluation strategies. More specifically, the TwinNet regularization presented in [17] and proposed in [16] for optimization of the MaD architecture can lead to improvements to both signal-to-distortion ratio metric and perceptual improvements in separation quality of singing voice. Perceptually, the mean-squared-error and Kullaback-Leibler with the L_1 sparsity term have approximately equal performance.

The perceptually motivated objective function based on the noise-to-mask ratio was outperformed by the TwinNetwork regularization that encompasses the Kullback-Leibler reconstruction term. From the above, it is highlighted that techniques or evaluation schemes that are well established in signal processing do not necessarily provide improvements to deep learning approaches trained by stochastic gradient descent. This suggests that devising new objective functions should take into account the dynamics of stochastic gradient descent, the corresponding parameterized functions' gradients, and the perceptual impact of the metric on perceptual quality. Finally, the results presented in this study show trends for the studied deep learning architecture, but might differ for other architectures or approaches, depending on the artifacts and the interference that these approaches produce. In the spirit of reproducible research the source code and the listening tests corpora is available through: `https://github.com/Js-Mim/mss_pytorch/tree/nmr_eval` and `https://github.com/dr-costas/mad-twinnet`.

7. REFERENCES

[1] Z. Rafii, A. Liutkus, F. R. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307–1335, Aug 2018.

[2] B. Pardo, "Finding structure in audio for music information retrieval," *IEEE Signal Processing Magazine*, vol. 23, no. 3, pp. 126–132, May 2006.

[3] G. Roma, E.M. Grais, A.J.R. Simpson, and M. D. Plumbley, "Music remixing and upmixing using source separation," in *Proceedings of the 2nd Workshop on Intelligent Music Production*, September 2016.

[4] S.I. Mimilakis, E. Cano, J. Abeßer, and G. Schuller, "New sonorities for jazz recordings: Separation and mixing using deep neural networks," in *Proceedings of the 2nd Workshop on Intelligent Music Production*. September 2016, AES.

[5] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 Signal Separation Evaluation Campaign," in *Proceedings of the Latent Variable Analysis and Signal Separation: 14th International Conference on Latent Variable Analysis and Signal Separation*, Surrey, United Kingdom, July 2018.

[6] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band DenseNets for audio source separation," in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October 2017.

[7] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[8] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, September 2016.

[9] P. Magron and T. Virtanen, "Towards complex nonnegative matrix factorization with the beta-divergence," in *Proceedings of the 2018 IEEE International Workshop on Acoustic Signal Enhancement*, Sept. 2018.

[10] D. Websdale and B. Milner, "A comparison of perceptually motivated loss functions for binary mask estimation in speech separation," in *Proceedings of Interspeech 2017*, 2017, pp. 2003–2007.

[11] Q. Liu, W. Wang, P. J. B. Jackson, and Y. Tang, "A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions," in *2017 25th European Signal Processing Conference (EUSIPCO)*, Aug 2017, pp. 1270–1274.

[12] G. Naithani, J. Nikunen, L. Bramsløw, and T. Virtanen, "Deep neural network based speech separation optimizing an objective estimator of intelligibility for low latency applications," in *Proceedings of the 2018 IEEE International Workshop on Acoustic Signal Enhancement*, Sep. 2018.

[13] S. Venkataramani, R. Higa, and P. Smaragdis, "Performance based cost functions for end-to-end speech separation," in *in arXiv pre-prints:1806.00511 [eess.AS]*, 2018.

[14] Z. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 686–690.

[15] S. I. Mimilakis, K. Drossos, J. F. Santos, G. Schuller, T. Virtanen, and Y. Bengio, "Monaural singing voice separation with skip-filtering connections and recurrent inference of time-frequency mask," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018.

[16] K. Drossos, S. I. Mimilakis, D. Serdyuk, G. Schuller, T. Virtanen, and Y. Bengio, "MaD TwinNet: Masker-denoiser architecture with twin networks for monaural sound source separation," in *Proc. IEEE International Joint Conference on Neural Networks (IJCNN)*, July 2018.

[17] D. Serdyuk, N.-R. Ke, A. Sordoni, A. Trischler, C. Pal, and Y. Bengio, "Twin Networks: Matching the future for sequence generation," in *Proc. of International Conference on Learning Representations (ICLR)*, April 2018.

[18] K. Brandenburg and T. Sporer, "NMR and Masking Flag: evaluation of quality using perceptual criteria," in *Audio Engineering Society Conference: 11th International Conference: Test & Measurement*, May 1992.

[19] J. Nikunen and T. Virtanen, "Noise-to-mask ratio minimization by weighted non-negative matrix factorization," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 25–28.

[20] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA – A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.

[21] E. Cano, J. Liebetrau, D. Fitzgerald, and K. Brandenburg, "The dimensions of perceptual quality of sound source separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 601–605.

[22] S. I. Mimilakis, K. Drossos, T. Virtanen, and G. Schuller, "A recurrent encoder-decoder approach with skip-filtering connections for monaural singing voice separation," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, September 2017.

[23] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[24] F. Baumgarte, C. Ferekidis, and H. Fuchs, "A nonlinear psychoacoustic model applied to the iso mpeg layer 3 coder," in *Prceedings of the 99th Audio Engineering Society Convention*, 1995.

[25] Z. Rafii, A. Liutkus, F.R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec 2017.

[26] ITU, "Recommendation ITU-R BS.1534-3: Method for the subjective assessment of intermediate quality level of audio systems," 2014.

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