Exploring a Perceptually-weighted DNN-based Fusion Model for Speech Separation

Alessandro Ragano, Andrew Hines

Insight Centre for Data Analytics, University College Dublin, Ireland
alessandro.ragano@ucdconnect.ie, andrew.hines@ucd.ie

Abstract. Deep Neural Network (DNN)-based fusion approaches for single-channel speech separation have recently been introduced but the non uniform perceptual weighting of the human auditory system has not been exploited during the DNN training phase. In addition, the perceived quality of the speech signal has not been assessed using a DNN-based fusion model. We propose a new perceptually-weighted DNN-based fusion model which employs a perceptual cost function and assess the perceived quality of several DNN-based fusion models. Objective and subjective evaluations for speech quality are compared. The results show that the perceptually-weighted DNN-based fusion model displays a significant improvement in terms of Source To Interferences Ratio (SIR) compared to a combined mask. However subjective quality assessment listening tests suggests that the proposed DNN-based fusion model does not result in improved perceived speech quality.

Keywords: Deep Neural Networks, Perceptual Audio Quality, Speech Separation, Fusion.

1 Introduction

Speech separation consists of extracting the speech signal from a mixture signal that contains one or more audio sources. Ideally, the estimated speech signal should be unaffected as much as possible, i.e., without the presence of other sources and without any distortion. The speech separation problem is more complicated when only one channel is provided, i.e., the single-channel scenario. In this paper we propose a Deep Neural Network (DNN)-based fusion method which considers the perceptual importance of each frequency band of the speech signal during the training.

Within the research community, several models have been proposed in order to solve the single-channel speech separation problem. Independent Component Analysis (ICA) has been proposed in [12, 10], where the inherent time structure of audio sources is encoded in the ICA basis functions. Other approaches usually work in the time/frequency domain, where scaling matrices called time/frequency masks are estimated and applied to the mixture in order to extract sources. Rowes proposed the factorial Hidden Markov Models (HMM) in [20], while approaches based on the Non-Negative Matrix Factorization (NMF) are described in [23, 19, 21, 20] where the mixture magnitude spectrogram is factorized in two matrices. Recently, approaches based on Deep Learning [7, 9] have been shown to exhibit better performances than the previous methods establishing the current state-of-the-art performance in single-channel speech enhancement. Even though the existing approaches achieve good performances, the speech separation problem remains far from being solved as the estimated speech signal is usually affected by distortions and background interference. One typical limitation that occurs in speech separation methods concerns robustness, i.e., the achievement of high performance under limited and specific conditions of the mixture. When conditions deviate, separation performance decreases. For example, some time/frequency masks have been shown to be successful for stereophonic scenarios [1, 19, 3] by exploiting some characteristics of the mix (e.g., the speech signal is typically located in the center channel) but they could fail in the mono channel scenario. In order to overcome this limitations, fusion methods have been recently introduced. They involve combining various estimated time/frequency masks for covering several signal aspects that typically occur in real applications. They have been shown to be successful in classification tasks [16] and have been applied to audio source separation problems. A fusion framework for underdetermined audio source separation that employs fusion rules inspired by classification is described in [13]. Compared to [13], significant improvements have been found in [14] where the authors proposed three alternative fusion methods based on standard nonlinear optimization, Bayesian model averaging and DNN. The DNN approach was favoured and has been widely explored in [6, 4, 5]. Although fusion methods have been shown to be successful, they did not account for the perceived quality of the extracted sources. Existing fusion methods assessed the amount of distortion and interference using the Blind Audio Source Separation (BASS) performance measurements [22] that although they perform well, they do not account for perceptual aspects [15, 2]. Regarding the perceived quality of the speech signals, perceptually weighted DNN have recently gained interest [18, 25]. As yet, to the authors' knowledge, no perception-based cost function has been explored with a DNN-based fusion model. In this paper we propose a new DNN-based fusion model that employs a perceptually-weighted cost function which is partly derived from [15] and we also explore how fusion of speech separation time/frequency masks using a DNN affects the perceived quality.

The paper is structured as follows: Section 2 gives a formulation of the BASS problem in the time/frequency domain. Section 3 presents the proposed DNN-based fusion architecture and Section 4 describes the DNN training and the perceptually-weighted cost function. Section 5 shows the experimental results: how the speech signal is subjectively perceived, and the results obtained from the BASS performance measurements. Conclusions are offered in Section 6.

2 Formulation of the Audio Source Separation Problem in the Time/Frequency Domain

In this paper we limit our scope to BASS problems as we only make use of the mixture signal and some a priori statistics of the source signals. In addition, instead of using a convolutive system, we model the channel signal as a linear combination of the source signals as we neglect the presence of environment reverberation. Let us consider $X(k, f)$, $S_1(k, f)$ and $S_2(k, f)$ as the Short Time Fourier Transform (STFT) of the mixture $x(n) = \ s_1(n) \ + \ s_2(n)$, the speech signal $s_1(n)$ and the background signal $s_2(n)$ respectively. Due to the sparsity characteristics of the sources in the time/frequency domain, the mixture magnitude spectrogram $|X(k, f)|$ almost preserves the linear mixing conditions and we can approximate the single channel as follows:

$$|X(k, f)| \approx |S_1(k, f)| + |S_2(k, f)|. \tag{1}$$

Methods that work in the STFT domain usually estimate a spectral weighting matrix $\mathbf{M}$ that assigns a value for each time/frequency element of the mixture spectrogram. More specifically we want to produce 2 masks that applied on the mixture spectrogram gives us the estimation of the magnitude source spectrograms $Z_1(k, f)$ and $Z_2(k, f)$:

$$Z_1(k, f) = M(k, f) \odot |X(k, f)| \tag{2}$$

$$Z_2(k, f) = (1 - M(k, f)) \odot |X(k, f)|. \tag{3}$$

This operation, called time/frequency weighting, discriminates the frequency bins between the sources. In the next section we see how to combine four time/frequency masks to overcome the limitations of using a mask individually.

3 DNN Fusion Architecture

The proposed method is based on using a Feedforward Neural Network (FNN) in order to achieve the combination of time/frequency masks that are characterised by different properties. More specifically, we analyse two DNNs trained with different features and the same cost function, i.e., the Mean Squared Error (MSE) and we propose one DNN which employs a new perceptually-weighted cost function.

The block diagram fusion model architecture is shown in Figure 1 where four time/frequency masks are combined and the single channel scenario was studied. The architecture is partially inspired from [6].

Every speech separation algorithm produces two masks, one for the speech signal and one for the background signal. Since we combine two algorithms, we deal with the fusion of four masks. From the mixture signal an STFT with Hann window and 2048 points is computed. It has been shown that source separation in the time/frequency domain can be estimated only with the magnitude

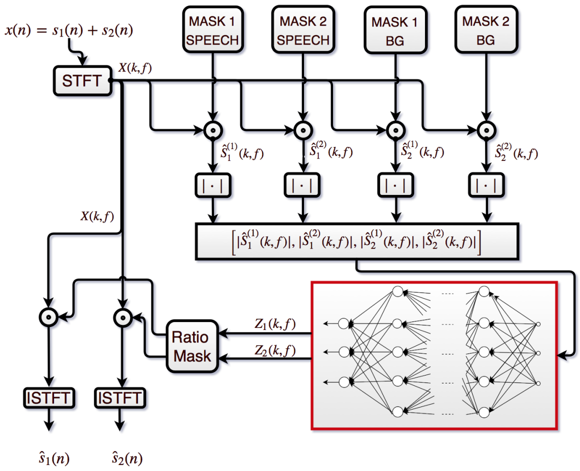


Fig. 1: Block Diagram Fusion Model Architecture: Fusion of four different mask outputs: two speech signal and two background signal are fused using a perceptually-weighted DNN to produce a single ratio mask. Notice that we reconstruct a mask as we chose to apply a mask post-processing musical-noise suppression.

spectrogram by discarding the phase which is useful for the time-domain reconstruction. Next, the Hadamard product between the STFT of the mixture and each mask is used to compute two time/frequency representations for each source. The FNN input is the concatenation of four estimated magnitude spectrograms, where two of them account for the speech signal while the others represent the background signal. Generally, fusion of source separation methods can be conducted by combining different kinds of information such as masks or separate magnitude spectrograms produced by masks themselves. However, in this scenario, we combined the separated magnitude spectrograms as in early experiments we observed better results combining magnitude spectrograms when compared to masks combinations.

The DNN takes the combined magnitude spectrograms as input features and produces new estimated magnitude spectrograms of both speech and background signal. The sources are then combined in order to reconstruct a mask as follows:

$$M_1^{dnn}(k,f) = \frac{Z_1(k,f)^2}{Z_1(k,f)^2 + Z_2(k,f)^2} \tag{4}$$

$$M_2^{dnn}(k,f) = 1 - M_1^{dnn}(k,f). \tag{5}$$

We reconstruct a mask instead of using the output of the FNN as we chose to apply a musical-noise suppression filter which has been developed for masks and not for magnitude spectrograms. Each mask is used to scale the mixture input $X(k, f)$ in order to estimate the STFT of the two sources:

$$\mathbf{Z}_1(k, f) = M_1^{dnn}(k, f) \odot X(k, f) \tag{6}$$

$$\mathbf{Z}_2(k, f) = M_2^{dnn}(k, f) \odot X(k, f). \tag{7}$$

Finally the two sources are reconstructed in the time domain using an Inverse Short Time Fourier Transform (ISTFT) and reapplying the phase of the original mixture. The described architecture can be imagined with more separation algorithms and it is easily extendable for a generic number of speech separators. However extending the architecture necessitates a network modification in terms of number of nodes and network hyperparameters. A key issue of the proposed model concerns the FNN outputs. The FNN can be trained in order to learn how to predict masks or magnitude spectrograms. Predicting masks facilitates easier training as the FNN learns values in a bounded range [0,1], while the prediction of magnitude spectrograms requires unlimited non negative range. Nevertheless we chose to predict magnitude spectrograms as they are less sensitive to the SNR variations of training data with respect to mask prediction. Thus, as shown in [24] we can employ a training dataset with significant SNR variations in the mixtures.

4 DNN training

The employed training dataset contains 250 audio mixes of speech and background signal which repeat themselves with different combinations. We split the dataset in two parts: 80% of time frames were assigned to the training data and the remaining 20% was used for validation. All the time frames have been shuffled using a fixed random seed in order to compare all the models. The mixtures are constructed with different SNR in order to have high variability and accurate reproducibility of real mix scenarios. The SNRs chosen are: 3 dB, 6 dB, 9 dB, 12 dB, 15 dB, 18 dB. The speech signals have been recorded in different languages.

In all the models we perform a z-score standardization such that the magnitudes of the frequency bins have the properties of a standard normal distribution with zero mean and unit variance. This pre-processing guarantees that the update of the weights is not biased by particular directions [17]. The DNN has 4100 node for each hidden layer with 2 hidden layers and 2050 nodes in the output layer. This is due to the fact that we compute a 2048 point STFT and each estimated magnitude spectrogram has 1025 frequency bins. The activation function in the hidden layer is the hyperbolic tangent. The output layer contains the concatenation of the estimated speech and the estimated background and has Rectified Linear Unit (ReLU) as activation function since we want to predict magnitude spectrograms. The choice of the parameters and the network architecture were partly taken from [6]. In order to prevent overfitting we used the early stopping technique which suggests that no overfitting occurred during the training. The early stopping has been preferred to the cross-validation because of the computational cost of the latter method [8].

4.1 Reduced Resolution Magnitude Spectrograms

In order to deal with a manageable feature set, we propose to use a reduced resolution of the magnitude spectrograms which exploits the perceptual mel scale. The mel scale sets the relationship between the perceived frequency of a pure tone respect to its actual measured frequency. Humans discern frequencies differently depending on the frequency range: small changes in the low frequency range are discerned better than the high frequency range. In order to compute the mel spectrogram we applied mel-spaced rectangular filter banks to each time frame obtaining a reduced resolution of each frame. The reduced resolution allows the combination of a greater number of speech separation algorithms and reduces the training time. However, doing this operation with linearly spaced filter banks would have not been beneficial since it does not reflect the human hearing resolution. On the other hand, mel-spaced filter banks reduce the number of frequency bins while maintaining the speech quality. Unlike a typical conversion with 40 filters we used 192 filters keeping the signal quality while maintaining sufficient frequency information to train the network. Both MSE models have been trained with batch size of 256 and learning rate equal to 0.01.

4.2 Modified Distortion To Signal Ratio

We propose a fusion model which employs a Modified Distortion To Signal Ratio (MDSR) as a cost function. The MSE gives the same importance to each frequency bin that is involved in the error. It computes the energy of the error without taking into account that some frequency bands are perceptually more important than others. The MDSR instead, takes into account the perceived audio quality of blindly separated audio signals and it is mainly derived from the Distortion To Signal Ratio (DSR) called DSX in [15].

Given a time frame, the DSX is formulated as:

$$DSX(k)_{seg,s_j} = \frac{\sum_{i=1}^{I}\sum_{l=1}^{L}\left(\frac{E_{error,j}(k,l,i)}{E_{target,j}(k,l,i)}w(k,l,i)\right)}{\sum_{i=1}^{I}\sum_{l=1}^{L}w(k,l,i)}. \tag{8}$$

The ratio between the energy error and the energy target can be seen as a classic DSR except that it is computed in the Bark bands. The Bark scale is a frequency scale on which equal distances correspond with perceptually equal distances, more details regarding the Bark scale are showed in [26]. This is weighted by the perceived loudness $w(k,l,i)$ which boosts up the DSR in the perceptually important bands.

The estimation of the perceived loudness is given by:

$$w(k,l,i) = E_{target,s_1,s_2}(k,l,i)^{0.25}. \tag{9}$$

Unlike the BASS measurements, that exhibit poor correlation with subjective perception, the DSX has been shown to be linked with the perceived quality

[15]. This is why we identified it as a potential fitness function for improving the DNN-based fusion model performances.

In order to adapt the DSX for network training, a number of modifications were required. In [15] the author used orthogonal projections in order to compute the error and the target energy, while we will consider the energy of the squared difference between the target signal and the estimated signal in the Bark scale and the square of the target signal. This approximation simplifies the differentiability requirements of the cost function in the training phase. The number of channels was also adapted. The DSX formulation is composed by summation over Bark band and over channels while we only consider the single channel as we are in the single channel scenario. Another difference concerns the concatenation of speech and background while the original DSX formulation is unique for each source. Therefore in our model the $E_{target(k,l)}$ contains 48 Bark bands where the first 24 address the energy of the target speech signal, while the remaining 24 represent the energy of the target background signal. The other measures such as loudness and energy error are modified at the same way. Therefore, the adapted version of the DSX is:

$$MDSR(k) = \frac{\sum_{l=1}^{2L} \left(\frac{E_{error,s_1 s_2}(k,l)}{E_{target,s_1 s_2}(k,l)} w(k,l) \right)}{\sum_{l=1}^{2L} w(k,l)} \tag{10}$$

where $L = 24$ is the number of Bark bands. This model has been trained with learning rate equal to 1e-5.

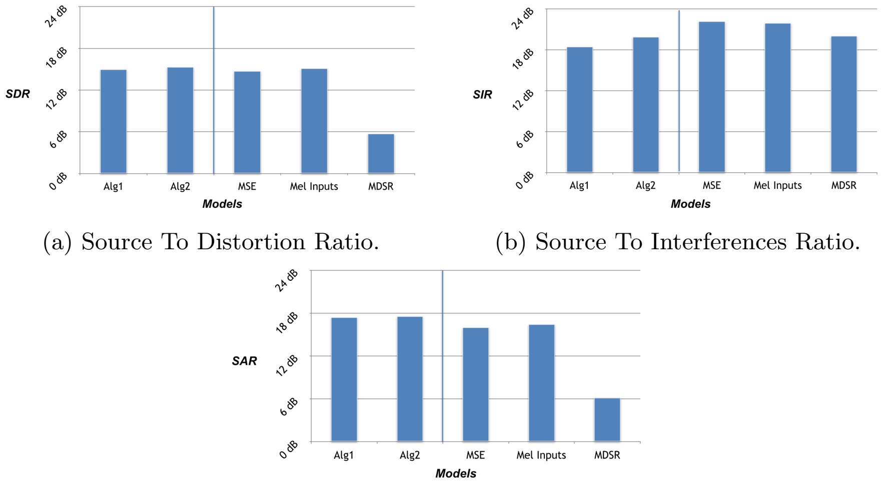
Even though the results in [15] suggested positive expectations we will show in the next section that this function does not produce anticipated improvements in terms of the perceived quality.

5 Experimental Results and Discussions

Model Name	Model Meaning	SOA/Contribution
Alg1	The first algorithm used in the fusion	State of the Art
Alg2	The second algorithm used in the fusion	State of the Art
MSE	FNN fusion with MSE	Contribution
Mel Inputs	FNN fusion with mel magnitude spectrograms	Contribution
MDSR	FNN fusion with MDSR	Contribution

Table 1: Algorithms used for the performance evaluations.

This section provides an evaluation of the experimental results. Both objective and subjective measurements are analysed, in order to see how they correlate.



(a) Source To Distortion Ratio.

(b) Source To Interferences Ratio.

(c) Source To Artifacts Ratio.

Fig. 2: Comparison of performances computed on the estimated speech signal. The graph shows the average values over the whole test dataset.

5.1 Evaluation of the Blind Audio Source Separation Performance Measurements

The objective measurements have been computed with the BASS performance measurements, described in [22]. The BASS performance measurements take into account some aspects like distortion, interference and artifacts. We computed 3 measurements, Source To Distortion Ratio (SDR), Source To Interferences Ratio (SIR) and Source to Artifacts Ratio (SAR), for all the models. The results have been computed with a test dataset composed by 65 audio mixtures whose Signal To Noise Ratio (SNR) range goes from 3 dB to 18 dB. Our model predicts both the speech signal and background signal, but since we are more interested in the speech signal we decided to do not consider performances regarding the background signal. We use graph that show performances in dB and we evaluate them on all the models, which are listed in Table 1. Each value in dB represents the average over the all test dataset of the measurement computed on the extracted speech signal.

Source To Distortion Ratio Each combination displayed similar performance to each individual algorithm meaning that there is no significant improvement in terms of SDR. The MDSR shows 5 dB of SDR which is not considered to be a promising result.

Source To Interferences Ratio The SIR measurement has been computed for assessing the amount of interference of the background signal that occurs in the estimated speech signal. Each DNN-based combiner achieves an improvement of roughly 4 dB with respect to each mask individually. In addition, all of the DNN-based combiner achieve an improvement of 10 dB with respect to the test mixtures which have an average SIR of 11.52 dB.

Source To Artifacts Ratio The last BASS performance measurement represents how much of artifacts is present with respect to the target signal. We observed that every combiner introduces more artifacts than each algorithm individually.

5.2 Evaluation of the perceived quality

The subjective performance measurements have been assessed with a listening test based on the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) test [11]. We asked to 6 expert listeners to evaluate the overall sound quality of the speech signal. As indicated in the ITU BS.1534-1 recommendations, the listeners had to assign a score from 0 (bad) to 100 (excellent) for different versions of the speech signal produced by the following models:

- Algorithm 1
- Algorithm 2
- Fusion with FNN - MSE
- Fusion with FNN - mel spectrograms
- Fusion with FNN - MDSR
- The target speech signal (Hidden Reference)
- A highly distorted version of the speech signal (Anchor).

In order to have accurate audio fidelity we used the professional STAX headphones. Unlike the original MUSHRA, where the anchor is a low-pass filtered version of the reference, we employed a strongly degraded version of the speech signal as we are not assessing a lossy audio compression algorithm. We asked listeners to evaluate how the overall sound quality was different with respect to the target speech signal. This means that they had to assign low score when the estimated speech signal showed artifacts and also when the background signal was considered relevantly present. We chose 5 audio mixtures taken from a different dataset with varied language and speaker gender. Each audio mixture has been taken as 6 dB of input SNR since it reflects more real scenarios and our training dataset range goes from 3 dB to 18 dB. All the signals have been normalized to have the same integrated loudness. In order to evaluate how the DNN was going to act in presence of varied background signals we decided to include background that were different from the ones used in the training dataset, i.e., drum tracks, guitar tracks and impulsive noise. The results, showed in Figure 3, represent the mean and the 95% confidence intervals computed using Student's t-distribution for each test item and algorithm, and the mean over all items. The approximation by looking the mean of the models suggests that on average, fusion models have been rated worse than one of the combined time/frequency mask. The observed result is not line with the objective BASS performance measurements

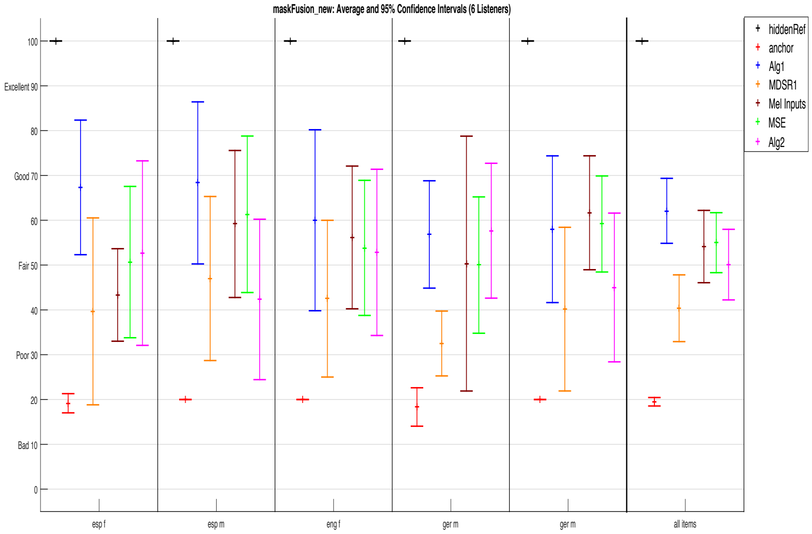


Fig. 3: Listening test based on MUSHRA. For each model it shows the average and 95% confidence intervals of the scores related to each audio item and the mean over all items.

where the DNN-based fusion models outperform each mask individually. MDSR did not perform as anticipated with feedback and comments of the listeners confirming a presence of the noise gate effect in various signals, which is considered annoying since it is usually preferred an higher constant noise to a lower but variable noise. After informal listening we believe that those signals are the ones produced by the perceptually-weighted DNN. Therefore we conclude that the MDSR does not show promising results in terms of the perceived quality and more work is still required. It should be noted that MDSR was shown to correlate source separation with perceived quality in [15]. This work has shown that this correlation does not translates into improved quality when MDSR is applied in a cost function. In addition, it can be observed that using the reduced resolution is convenient as it reduces the training time and it did not significantly impact performance quality.

6 Conclusions

We have proposed a new perceptually-weighted DNN-based fusion model that takes into account the perceived quality of blindly separated audio signals. We assessed both objective and subjective measurements in order to see if the perceptual model outperforms each mask individually and others DNN-based fusion models.

The experimental results from the BASS performance measurements show that the perceptually-weighted DNN-based fusion model outperforms only one of the two algorithms in terms of SIR. The DNN-based fusion model with MSE outperforms each mask individually in terms of SIR and presents similar performances in terms of SDR. We also conducted a listening test for assessing the perceived speech quality. Our experimental results show that using any DNN-based fusion models does not improve the perceived quality of the speech signals compared to each mask individually.

7 Acknowledgements

This work was performed at Fraunhofer IIS, Erlangen, Germany with the cooperation of Politecnico di Milano, Milan, Italy. This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

The authors would like to sincerely thank Dr. Jouni Paulus and Prof. Jürgen Herre of Fraunhofer IIS for their support and their excellent advice during this research project.

References

1. Cobos, M., López, J.J.: Stereo audio source separation based on time-frequency masking and multilevel thresholding. Digital Signal Processing: A Review Journal **18**(6), 960–976 (2008)
2. Emiya, V., Vincent, E., Harlander, N., Hohmann, V.: Subjective and Objective Quality Assessment of Audio Source Separation. IEEE Transactions on Audio, Speech and Language Processing **19**(7), 2046 –2057 (2011)
3. Geiger, J.T., Grosche, P., Parodi, Y.L.: Dialogue enhancement of stereo sound. European Signal Processing Conference (EUSIPCO). pp. 869–873 (2015)
4. Grais, E.M., Roma, G., Simpson, A.J.R., Plumbley, M.D.: Single Channel Audio Source Separation using Deep Neural Network Ensembles. AES: Journal of the Audio Engineering Society (9494), 236–246 (2016)
5. Grais, E.M., Roma, G., Simpson, A.J.R., Plumbley, M.D.: Two-Stage Single-Channel Audio Source Separation Using Deep Neural Networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing **25**(9), 1469–1479 (2017)
6. Grais, E.M., Roma, G., Simpson, A.J., Plumbley, M.D.: Combining Mask Estimates for Single Channel Audio Source Separation Using Deep Neural Networks. Interspeech 2016, Proceedings of. pp. 3339–3343 (2016)
7. Grais, E.M., Sen, M.U., Erdogan, H.: Deep neural networks for single channel source separation. Acoustics, Speech and Signal Processing (ICASSP). pp. 3734–3738 (2014)
8. Hines, A., Kendrick, P., Barri, A., Narwaria, M., Redi, J.A.: Robustness and prediction accuracy of Machine Learning for objective visual quality assessment. In: European Signal Processing Conference. pp. 2130–2134 (2014)
9. Huang, P.S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P.: Deep Learning for Monaural Speech Separation. Acoustics, Speech and Signal Processing (ICASSP). pp. 1562–1566 (2014)

10. Hyvärinen, A., Oja, E.: Independent Component Analysis: Algorithms and Applications. Neural Networks **13**(45), 411–430 (2000)
11. International Telecommunication Unions: BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems pp. 1–18 (2003)
12. Jang, G.J., Lee, T.W.: A Maximum Likelihood Approach to Single-channel Source Separation. Journal of Machine Learning Research **1**(7-8), 1365–1392 (2003)
13. Jaureguiberry, X., Richard, G., Leveau, P., Hennequin, R., Vincent, E.: Introducing a simple fusion framework for audio source separation. IEEE International Workshop on Machine Learning for Signal Processing, MLSP (2013)
14. Jaureguiberry, X., Vincent, E., Richard, G.: Fusion methods for speech enhancement and audio source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **24**(7), 1266–1279 (2016)
15. Kastner, T.: Evaluating Physical Measures for Predicting the Perceived Quality of Blindly Separated Audio Source Signals. AES: Journal of the Audio Engineering Society pp. 1–13 (2009)
16. Kittler, J., Hater, M., Duin, R.P.: Combining classifiers. Proceedings - International Conference on Pattern Recognition **2**(3), 897–901 (1998)
17. LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.R.: Efficient backprop. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **7700 LECTU**, 9–48 (2012)
18. Liu, Q., Wang, W., Jackson, P.J.B.: A Perceptually-Weighted Deep Neural Network for Monaural Speech Enhancement in Various Background Noise Conditions. European Signal Processing Conference (EUSIPCO). pp. 1310–1314 (2017)
19. Ozerov, A., Févotte, C.: Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Transactions on Audio, Speech and Language Processing **18**(3), 550–563 (2009)
20. Roweis, S.S.T.: One microphone source separation. Advances in neural information processing systems pp. 793–799 (2001)
21. Smaragdis, P., Brown, J.C.: Non-Negative Matrix Factorization for Polyphonic Music Transcription. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics pp. 177–180 (2003)
22. Vincent, E., Gribonval, R., Févotte, C.: Performance Measurement in Blind Audio Source Separation. IEEE Transactions on Audio, Speech and Language Processing **14**(4), 1462–1469 (2006)
23. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Transactions on Audio, Speech and Language Processing **15**(3), 1066–1074 (2007)
24. Zhang, X.L., Wang, D.: A Deep Ensemble Learning Method for Monaural Speech Separation. IEEE Transactions on Audio, Speech and Language Processing **21**(5), 1475–1487 (2016)
25. Zhen, K., Sivaraman, A., Sung, J., Kim, M.: On Psychoacoustically Weighted Cost Functions Towards Resource-Efficient Deep Neural Networks for Speech Denoising. arXiv preprint arXiv:1801.09774 (2018)
26. Zwicker, E.: Subdivision of the audible frequency range into critical bands (frequenzgruppen). The Journal of the Acoustical Society of America **33**(2), 248–248 (1961)