Deep Filtering: Signal Extraction Using Complex Time-Frequency Filters

Preprint · April 2019

CITATIONS
0

READS
196

2 authors:

Wolfgang Mack
International Audio Laboratories Erlangen
2 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Emanuël A. P. Habets
Friedrich-Alexander-University of Erlangen-Nürnberg
254 PUBLICATIONS **3,399** CITATIONS
SEE PROFILE

Deep Filtering: Signal Extraction Using Complex Time-Frequency Filters

Wolfgang Mack, and Emanuël A. P. Habets, *Senior Member, IEEE*
{wolfgang.mack, emanuel.habets}@audiolabs-erlangen.de

***Abstract*—Signal extraction from a single-channel mixture with additional undesired signals is most commonly performed using time-frequency (TF) masks. Typically, the mask is estimated with a deep neural network (DNN) and element-wise applied to the complex mixture short-time Fourier transform (STFT) representation to perfom extraction. Ideal mask magnitudes are zero for solely undesired signals in a TF bin and infinity for total destructive interference. Usually, masks have an upper bound to provide well-defined DNN outputs at the cost of limited extraction capabilities. We propose to estimate with a DNN a complex TF filter for each mixture TF bin which maps an STFT area in the respective mixture to the desired TF bin to address destructive interference in mixture TF bins. The DNN is optimized by minimizing the error between the extracted and the ground-truth desired signal allowing to train without having to specify ground-truth TF filters but learn filters by error reduction. We compare our approach with complex and real valued TF masks by separating speech from a variety of different sound and noise classes from the Google AudioSet corpus. We also process the mixture STFT with notch filters and zero whole time-frames to demonstrate the reconstruction capabilities of our approach. The proposed method outperformed the baselines especially when notch filters and time-frame zeroing were applied.**

***Index Terms*—Signal Extraction, Signal Enhancement, Time-Frequency Masking**

I. Introduction

Real world signals are often corrupted by undesired noise sources or interferers like white self-noise of microphones, background sounds like babble noise, or traffic, but also impulsive sounds like clapping. Preprocessing, like notch filtering, or specific room acoustics which cause spatial comb filters can also contribute to a decreased quality of the recorded signal. Extracting and/or reconstructing the desired signal from such a mixture is highly desired when high-quality signals are needed. Possible applications are for example enhancing recorded speech signals, separating different sources from each other or packet-loss concealment. Signal extraction methods can broadly be categorized in single- and multi-channel approaches. In this paper, we focus on single-channel approaches and address the problem of extracting a desired signal from a mixture of desired and undesired signals.

Common approaches perform this extraction in the short-time Fourier transform (STFT) domain, where either the desired spectral magnitude (e.g. [1]) or a time-frequeny (TF) mask is estimated which then is element-wise applied to the complex mixture STFT to perform extraction. Estimating TF masks is usually prefered over directly estimating spectral magnitudes due to performance reasons [2]. Typically, TF masks are estimated from a mixture representation by a deep neural network (DNN) (e.g. [2]–[9]) where the output layer often directly yields the STFT mask. Two common approaches exist to train such DNNs. First, a ground-truth mask is defined and the DNN learns the mixture to mask mapping by minimizing an error function between the ground-truth and estimated masks (e.g. [3], [5]). In the second approach, the DNN learns the mapping by directly minimizing an error function between estimated and desired signal (e.g. [8], [10], [11]). Erdogan et al. [12] showed that direct optimization is equal to mask optimization weighted with the squared mixture magnitude. Consequently, the impact of high energy TF bins on the loss is increased and the impact of low energy decreased. Furthermore, no ground-truth mask has to be defined as it is implicitly given in the ground-truth desired signal.

For different extraction tasks, different types of TF masks have been proposed. Given a mixture in STFT domain where the signal in each TF bin either belongs solely to the desired or the undesired signal, extraction can be performed using binary masks [13] which have been used e.g. in [5], [7]. Given a mixture in STFT domain where several sources are active in the same TF bin, ratio masks (RMs) [14] or complex ratio masks (cRMs) [15] can be applied. Both assign a gain to each mixture TF bin to estimate the desired spectrum. The real-valued gains of RMs perform TF bin wise magnitude correction from mixture to desired spectrum. The estimated phase is in this case equal to the mixture phase. cRMs apply a complex instead of a real gain and additionally perform phase correction. Speaker separation, dereverberation, and denoising have been achieved using RM (e.g. [6], [8], [10], [11], [16]) and cRM (e.g. [3], [4]). Ideally, the magnitude of RMs and cRMs is zero if only undesired signals are active in a TF bin and infinity if the desired and undesired signals overlap destructively in a certain TF bin. Outputs approaching infinity cannot be estimated with a DNN. For obtaining well-defined DNN outputs, it is possible to estimate a compressed mask (e.g. [4]) with a DNN and perform extraction after decompression to obtain mask values with high magnitudes. Weak noise on the DNN output, however, can lead to a huge change in the estimated masks resulting in big errors. Furthermore, when the desired and undesired signals in a TF bin add up to zero, also a compressed mask cannot reconstruct the respective magnitude from zero by multiplication. Often, the case of destructive interference is ignored (e.g. [6], [11], [17]) and mask values bounded to one are estimated because higher values also come with the risk of noise amplification. Besides masks, also complex-valued TF filters (e.g. [18]) have

Wolfgang Mack and Emanuël Habets are with the International Audio Laboratories Erlangen (a joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer IIS), Germany.

been applied for the purpose of signal extraction. Current TF filter approaches usually incorporate a statistics estimation step (e.g. [18]–[21]) which can be crucial given a large variety of unknown interference signals with fast changing statistics as present in real-world scenarios.

In this paper, we propose to use a DNN to estimate a complex-valued TF filter for each TF-bin in the STFT domain to address extraction also for highly non-stationary signals with unknown statistics. The filter is element-wise applied to a defined area in the respective mixture STFT. The result is summed up to obtain an estimate of the desired signal in the respective TF bin. The individual complex filter values are bounded in magnitude to provide well-defined DNN outputs. Each estimated TF bin is a complex weighted sum of a TF bin area in the complex mixture. This allows to address the case of destructive interference in a single TF bin without the noise-sensitivity of mask compression. It also allows to reconstruct a TF bin which is zero by taking into account neighboring TF bins with non-zero magnitudes. The combination of DNNs and TF filters mitigates both the shortcomings of TF masks and of existing TF filter approaches.

The paper is structured as follows. In Section II, we present the signal extraction process with TF masks and subsequently, in Section III, we describe our proposed method. Section IV contains the data sets we used and Section V the results of the experiments to verify our theroretical considerations.

II. STFT Mask based Extraction

In this section, we review the extraction process with TF masks and provide implementation details of the masks we used as baselines in the performance evaluation.

A. Objective

We define the complex single-channel spectrum of the mixture as $X(n,k)$, of the desired signal as $X_{\mathrm{d}}(n,k)$, and of the undesired signal as $X_{\mathrm{u}}(n,k)$ in STFT domain where n is the time-frame and k is the frequency index. We consider the mixture $X(n,k)$ to be a superposition

$$X(n,k) = X_{\mathrm{u}}(n,k) + X_{\mathrm{d}}(n,k). \tag{1}$$

Our objective is to obtain an estimate of $X_{\mathrm{d}}(n,k)$ by applying a mask to $X(n,k)$, i.e.,

$$\widehat{X}_{\mathrm{d}}(n,k) = \widehat{M}(n,k) \cdot X(n,k), \tag{2}$$

where $\widehat{X}_{\mathrm{d}}(n,k)$ is the estimated desired signal and $\widehat{M}(n,k)$ the estimated TF mask. For a binary mask, $\widehat{M}(n,k)$ is $\in \{0,1\}$, for a RM $\widehat{M}(n,k)$ is $\in [0,b]$ with the upper-bound $b \in \mathbb{R}^+$, and for a cRM $|\widehat{M}(n,k)|$ is $\in [0,b]$, and $\widehat{M}(n,k)$ is $\in \mathbb{C}$. The upper-bound b is typically one or close to one. Binary masks classify TF bins, RMs perform magnitude correction and cRMs additionally perform phase correction from $X(n,k)$ to $\widehat{X}_{\mathrm{d}}(n,k)$. Addressing the extraction problem is in this case equal to addressing the mask estimation problem.

Usually TF masks are estimated with a DNN which is either optimized to estimate a predefined ground-truth TF mask for all $N \cdot K$ TF bins, where N is the total number of time-frames and K the number of frequency bins per time-frame

$$J_{\mathrm{M}} = \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{n=1}^{N} |M(n,k) - \widehat{M}(n,k)|^2, \tag{3}$$

with the ground-truth mask $M(n,k)$, or to reduce the reconstruction error between $X_{\mathrm{d}}(n,k)$ and $\widehat{X}_{\mathrm{d}}(n,k)$

$$J_{\mathrm{R}} = \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{n=1}^{N} |(X_{\mathrm{d}}(n,k) - \widehat{X}_{\mathrm{d}}(n,k))^2|, \tag{4}$$

or the magnitude reconstruction

$$J_{\mathrm{MR}} = \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{n=1}^{N} (|X_{\mathrm{d}}(n,k)| - |\widehat{X}_{\mathrm{d}}(n,k)|)^2. \tag{5}$$

Optimizing the reconstruction error is equivalent to a weighted optimization of the masks reducing the impact of TF bins with low energy and increasing the impact of high energy TF bins on the loss [12]. For destructive interference in (1), the well known triangle inequality given by

$$|X_{\mathrm{d}}(n,k) + X_{\mathrm{u}}(n,k)| < |X_{\mathrm{d}}(n,k)| < |X_{\mathrm{d}}(n,k)| + |X_{\mathrm{u}}(n,k)|, \tag{6}$$

holds, requiring $1 < |M(n,k)| \leqslant \infty$. Hence, the global optimum cannot be reached above the mask upper-bound b.

B. Implementation

For mask estimation, we use a DNN with a batchnorm layer followed by three bidirectional long short-term memory (BLSTM) layers [22] with 1200 neurons per layer and a feed-forward output layer with tanh activation yielding the output O with dimension $(N, K, 2)$ representing an imaginary and real output per TF bin $\in [-1, 1]$.

For mask estimation, we designed the model to have the same number of trainable parameters and the same maximum of $|\widehat{M}|$ for the RM and cRM approaches. We used a real-valued DNN with the stacked imaginary and real part of X as input and two outputs, defined as O_{r} and O_{i}, per TF bin. These can be interpreted as imaginary and real mask components. For RM estimation, we computed $\widehat{M}(n,k) = \sqrt{O_{\mathrm{r}}(n,k)^2 + O_{\mathrm{i}}(n,k)^2}$ resulting in $\widehat{M}(n,k) \in [0, \sqrt{2}]$. For the cRM $Re\{\widehat{M}(n,k)\} = O_{\mathrm{r}}(n,k)$ and $Im\{\widehat{M}(n,k)\} = O_{\mathrm{i}}(n,k)$. This setting yields a phase dependent maximal cRM magnitude between 1 and $\sqrt{2}$, where 1 is achieved for a pure real or imaginary mask value and $\sqrt{2}$ for $|O_{\mathrm{r}}(n,k)| = |O_{\mathrm{i}}(n,k)| = 1$ resulting in an amplification disadvantage of the cRM compared to the RM. We trained two DNNs to estimate a RM optimized with (5) and a cRM optimized with (4). We computed the complex multiplication of $X(n,k)$ and $\widehat{M}(n,k)$ in (2) for the cRM by

$$Re\{\widehat{X}_{\mathrm{d}}\} = Re\{\widehat{M}\} \cdot Re\{X\} - Im\{\widehat{M}\} \cdot Im\{X\}, \tag{7}$$

$$Im\{\widehat{X}_{\mathrm{d}}\} = Im\{\widehat{M}\} \cdot Re\{X\} + Re\{\widehat{M}\} \cdot Im\{X\}. \tag{8}$$

Note that (n,k) is omitted for brevity. We trained 100 epochs, used the Adam [23] optimizer, a dropout [24] of 0.4 in the BLSTMs, a batch size of 64, an initial learning rate of 1e-4 multiplied by 0.9 after each episode the validation loss did not decrease.

III. Proposed STFT Filter based Extraction

In this section we show how to estimate X_d using an STFT domain filter instead of TF masks. We refer to this filter as a deep filter (DF).

A. Objective

We propose to obtain $\widehat{X}_\mathrm{d}$ from X by applying a complex filter

$$\widehat{X}_\mathrm{d}(n,k) = \sum_{i=-I}^{I} \sum_{l=-L}^{L} H^*_{n,k}(l+L, i+I) \cdot X(n-l, k-i), \quad (9)$$

where $2 \cdot L + 1$ is the filter dimension in time-frame direction and $2 \cdot I + 1$ in frequency direction and $H^*_{n,k}$ is the complex conjugated 2D filter of TF bin (n,k). Note that, without loss of generality, we used in (9) a square filter only for reasons of presentation simplicity. The filter values are like mask values bound in magnitude to provide well-defined DNN outputs

$$|H^*_{n,k}(l+L, i+I)| \leqslant b \;\forall l, i \in \mathbb{N} : l, i \in [-L, L], [-I, I]. \quad (10)$$

The DNN is optimized according to (4) which allows training without having to define ground-truth filters (GTFs) and to directly optimize the reconstruction mean squared error (MSE). The decision for GTFs is crucial because there are usually infinitely many combinations of different filter values that lead to the same extraction result. If a GTF was selected randomly for a TF bin from the set of infinitely many GTFs, training would fail because there would not be consistency between the selected filters. We can interpret this situation as a partially observable process for the GTF designer and fully observable for the DNN. From the input data properties, the DNN can decide exactly which filter to take without ambiguities. The GTF desginer has a infinitely large set of possible GTFs but cannot interpret the input data to decide which GTF to take so that the current DNN update is consistent w.r.t. previous updates. By training with (4), we avoid the problem of GTF selection.

B. Implementation

We used the same DNN as proposed in Section II-B changing the output shape to $(N, K, 2, 2 \cdot L + 1, 2 \cdot I + 1)$, where the last 2 entries are the filter dimensions. The complex multiplication in (9) was performed as shown in (7) and (8). In our experiments, we set $L = 2$ and $I = 1$ resulting in a filter dimension of $(5, 3)$. Similar as for cRMs in Subsection II-B, the maximum of $|H_{n,k}(l,i)|$ is phase dependent $\in [1, \sqrt{2}]$ for the employed output layer activation. As all $|H_{n,k}(l,i)|$ can be at least 1, a DNN can theoretically optimize (4) to its global optimum zero, if

$$c \cdot \sum_{i=-I}^{I} \sum_{l=-L}^{L} |X(n-l, k-i)| \geqslant |X_\mathrm{d}(n,k)|, \quad (11)$$

where $c \in \mathbb{R}^+$ is the maximal magnitude all filter values can reach and with $c = 1$ in our setting. Hence, to address destructive interference, the summation of all mixture magnitudes considered by a filter weighted with c must be at least equal to the desired TF bin magnitude. As filters exceed the spectrum for TF bins at the edge, we zero padded the spectrum with L zeros on time and I on frequency axis.

IV. Data Sets

We used the AudioSet [25] as interferer (without the speech samples) and LIBRI [26] as desired speech data corpora. All data was downsampled to 8 kHz sampling frequency and had a duration of 5 s. For the STFT we set the hop size to 10 ms, the frame length to 32 ms, and used the Hann window. Consequently, in our tests $K = 129$ and $N = 501$.

We degraded the desired speech samples by adding white noise, interference from AudioSet, notch-filtering and random time-frame zeroing (T-kill). Each degradation was applied to a sample with a probability of 50 percent. For the AudioSet interference, we randomly selected five seconds of AudioSet and desired speech from LIBRI to compute one training sample. Speech and interference were mixed with a segmental signal-to-noise-ratio (SNR) $\in [0, 6]$ dB, speech and white noise with SNR $\in [20, 30]$ dB. For notch-filtering, we randomly selected a center frequency with a quality factor $\in [10, 40]$. When T-kill was applied, every time-frame was zeroed with a probability of 10 percent. We generated 100000 training, 5000 validation and 50000 test samples using the respective sets of LIBRI and with the aforementioned degradations. To avoid overfitting, training, validation and test samples were created from distinct speech and interference samples from AudioSet and LIBRI. We divided the test samples in three subsets, namely Test 1, Test 2, and Test 3. In Test 1, speech was solely degraded by interference from AudioSet. In Test 2, speech was only degraded by both, notch-filtering and T-kill. In Test 3, speech was degraded by interference, notch-filtering, and T-kill simultaneously. All subsets include samples with and without white noise.

V. Performance Evaluation

For performance evaluation, we used the signal-to-distortion-ratio (SDR), the signal-to-artifacts-ratio (SAR), the signal-to-interference-ratio (SIR) [27], the reconstruction MSE (see (4)), the short-time objective intelligibility (STOI) [28], [29], and the test data set.

First, we tested how clean speech is degraded when processed. The MSEs after RM, cRM, and DF application were -33.5, -30.7, and -30.2 dB, respectively. The errors are very small and we assume them to be caused by noise on the DNN outputs. RMs produce the smallest MSE as noise on the DNN outputs solely affects the magnitude, then cRMs as phase and magnitude is affected and finally, DFs introduce the highest MSE. In an informal listening test, no difference was perceived. Table I shows the average results of Test 1 - 3. In Test 1, DFs, cRMs and RMs showed to generalize well to unseen interference. Processing with cRMs instead of RMs did not result in performance improvements although cRMs perform phase in addition to magnitude correction. This can result from the amplification disadvantage of cRMs compared to RMs caused by the employed DNN architecture described in Subsection II-B. For the metric STOI, DFs and RMs

TABLE I: Average Results SDR, SIR, SAR, MSE (in dB), STOI for RM, cRM, and DF for test samples degraded with AudioSet interference in Test 1, with a notch-filter and a time-frame zeroing (T-kill) in Test 2, and the combination in Test 3; unpr. MSE 1.60, -7.80, 1.12 and STOI 0.81, 0.89, 0.76 for Test 1, 2, 3, respectively

Test 1: Interference					
	MSE	STOI	SDR	SAR	SIR
RM	-10.23	**.86**	15.09	15.81	25.55
cRM	-10.20	.85	15.06	15.78	26.30
Proposed DF	**-10.83**	**.86**	**15.67**	**16.44**	**26.59**
Test 2: T-kill and Notch					
	MSE	STOI	SDR	SAR	SIR
RM	-7.80	.89	12.25	12.39	29.50
cRM	-7.80	.89	12.25	12.45	27.40
Proposed DF	**-18.63**	**.94**	**26.37**	**27.40**	**34.16**
Test 3: Interference, T-kill, and Notch					
	MSE	STOI	SDR	SAR	SIR
RM	-6.00	.82	9.81	10.04	24.73
cRM	-5.94	.81	9.77	10.15	25.20
Proposed DF	**-9.94**	**.85**	**14.77**	**15.21**	**26.21**

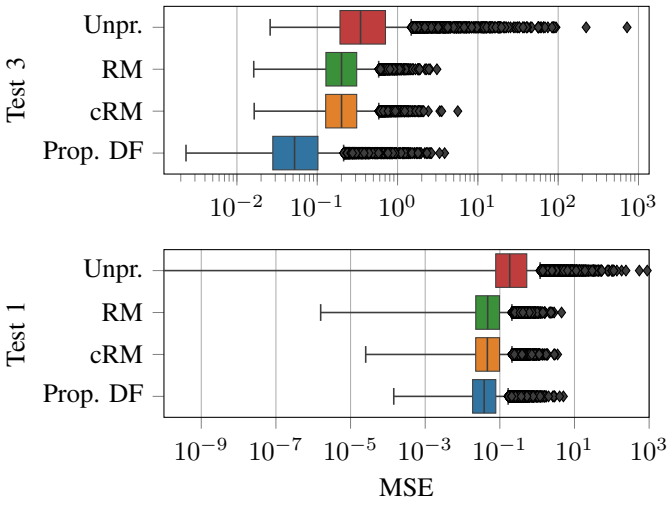

Fig. 1: MSE results of Test 1 and Test 3

performed on par whereas for the other metrics DFs performed better and achieved a further improvement of 0.61 dB in SDR. Boxplots of the MSE results are depicted in Figure 1. We assume this to be caused by the advanced reconstruction capabilities of DFs with respect to destructive interference. In Test 2, DFs clearly outperformed cRMs and RMs as expected because the test conditions provided a compareable scenario to destructive interference. Figure 2 depicts log-magnitude spectra of clean speech, degraded speech by zeroing every fifth time-frame and frequency axis and after enhancement with DF. Traces of the grid are still visible in low but not in high energy spectral regions as focused on by the loss in (4). In Test 3, DFs performed best as they are able to address all degradations whereas RMs and cRMs cannot. The baselines cRMs and RMs performed on par.

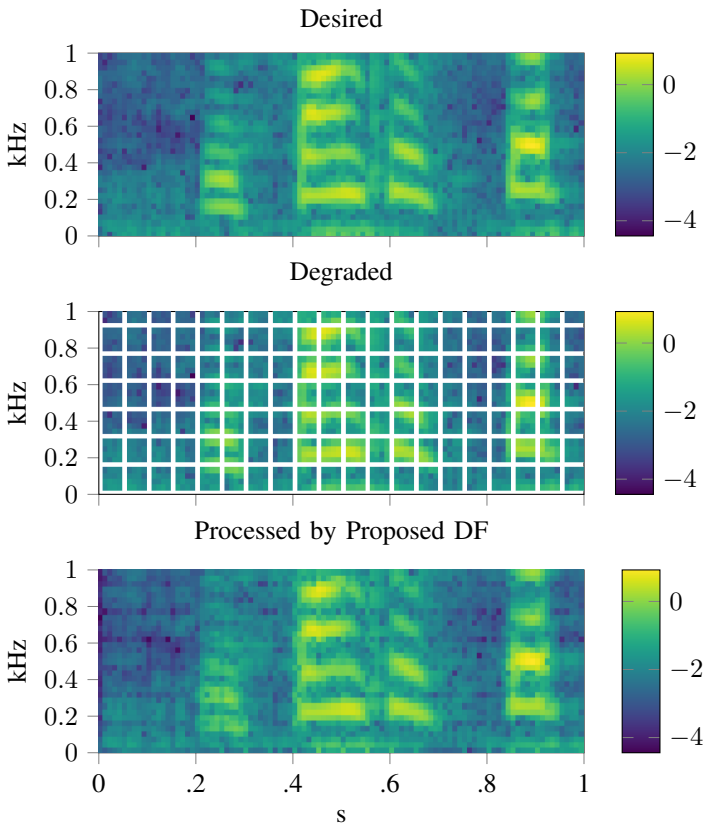

Fig. 2: Excerpt of log-magnitude STFT spectra of desired speech, degraded by zeroing every fifth time-frame and frequency, and after processing with DF. The degradation in this figure was performed for illustration purposes only unlike the random time-frame zeroing in the data sets.

VI. Conclusion

We extended the concept of time-frequency masks for signal extraction to complex filters to increase the interference reduction and decrease the signal distortion, and to adress destructive interference of desired and undesired signals. We proposed to estimate the filters with a deep neural network which is trained by minimizing the MSE between the desired and estimated signal and avoids defining ground-truth filters for training which would be crucial due to the necessity to consistently define filters for network training given infinity many possibilities. The filter and the mask methods were able to perform speech extraction given unknown interference signals from AudioSet which shows their generaliziability and introduced only a very small error when processing clean speech. Our approach outperformed a complex ratio mask in all and a ratio mask baseline in all but one metric where the performance was on par. Beside interference reduction, we tested whether data loss simulated by time-frame zeroing or filtering with notch filters can be addressed and showed that solely our proposed method was able to reconstruct the desired signal. Hence, with deep filters, signal extraction and/or reconstruction seems to be feasible under very adverse conditions given packet-loss, or unknown interference.

REFERENCES

[1] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[2] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[3] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 3, pp. 483–492, Aug. 2016.

[4] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 5590–5594.

[5] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 31–35.

[6] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 246–250.

[7] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech Conf.*, Sep. 2016, pp. 545–549.

[8] D. Yu, M. Kolbk, Z. H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 241–245.

[9] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 61–65.

[10] M. Kolbaek, D. Yu, Z.-H. Tan, J. Jensen, M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[11] W. Mack, S. Chakrabarty, F.-R. Stöter, S. Braun, B. Edler, and E. A. P. Habets, "Single-channel dereverberation using direct MMSE optimization and bidirectional LSTM networks," in *Proc. Interspeech Conf.*, Sep. 2018, pp. 1314–1318.

[12] H. Erdogan and T. Yoshioka, "Investigations on data augmentation and loss functions for deep learning based speech-background separation," in *Proc. Interspeech Conf.*, Sep. 2018, pp. 3499–3503.

[13] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Kluwer Academic, 2005, pp. 181–197.

[14] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*, G. R. Naik and W. Wang, Eds. Springer, 2014, pp. 349–368.

[15] F. Mayer, D. S. Williamson, P. Mowlaee, and D. Wang, "Impact of phase estimation on single-channel speech separation based on time-frequency masking," *J. Acoust. Soc. Am.*, vol. 141, no. 6, pp. 4668–4679, 2017.

[16] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. of the 12th Int. Conf. on Lat.Var. An. and Sig. Sep.*, ser. LVA/ICA. New York, USA: Springer-Verlag, 2015, pp. 91–99.

[17] X. Li, J. Li, and Y. Yan, "Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions," in *Proc. Interspeech Conf.*, Aug. 2017, pp. 1203–1207.

[18] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, ser. SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag, 2011.

[19] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 273–276.

[20] D. Fischer, S. Doclo, E. A. P. Habets, and T. Gerkmann, "Combined single-microphone Wiener and MVDR filtering based on speech interframe correlations and speech presence probability," in *Speech Communication; 12. ITG Symposium*, Oct. 2016, pp. 1–5.

[21] D. Fischer and S. Doclo, "Robust constrained MFMVDR filtering for single-microphone speech enhancement," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, 2018, pp. 41–45.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[23] J. B. D. Kingma, "Adam: A method for stochastic optimization," in *Proc. IEEE Intl. Conf. on Learn. Repr. (ICLR)*, May 2015, pp. 1–15.

[24] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[25] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 776–780.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 5206–5210.

[27] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "MIR_EVAL: A transparent implementation of common MIR metrics," in *Intl. Soc. of Music Inf. Retrieval*, Oct. 2014, pp. 367–372.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[29] M. Pariente, "pystoi," https://github.com/mpariente/pystoi, 2018.