A Hands-on Comparison of DNNs for Dialog Separation Using Transfer Learning from Music Source Separation

Martin Strauss[1], Jouni Paulus[1,2], Matteo Torcoli[2], Bernd Edler[1]

[1]International Audio Laboratories Erlangen[*], Germany
[2]Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

{martin.strauss, bernd.edler}@audiolabs-erlangen.de,
{jouni.paulus, matteo.torcoli}@iis.fraunhofer.de

Abstract

This paper describes a hands-on comparison on using state-of-the-art music source separation deep neural networks (DNNs) before and after task-specific fine-tuning for separating speech content from non-speech content in broadcast audio (i.e., dialog separation). The music separation models are selected as they share the number of channels (2) and sampling rate (44.1 kHz or higher) with the considered broadcast content, and vocals separation in music is considered as a parallel for dialog separation in the target application domain. These similarities are assumed to enable transfer learning between the tasks. Three models pre-trained on music (Open-Unmix, Spleeter, and Conv-TasNet) are considered in the experiments, and fine-tuned with real broadcast data. The performance of the models is evaluated before and after fine-tuning with computational evaluation metrics (SI-SIRi, SI-SDRi, 2f-model), as well as with a listening test simulating an application where the non-speech signal is partially attenuated, e.g., for better speech intelligibility. The evaluations include two reference systems specifically developed for dialog separation. The results indicate that pre-trained music source separation models can be used for dialog separation to some degree, and that they benefit from the fine-tuning, reaching a performance close to task-specific solutions.

Index Terms: dialog separation, deep neural network, transfer learning, fine-tuning

1. Introduction

A common complaint about the audio mix in TV-shows or movies is the low intelligibility of the spoken dialogs as a consequence of loud background music or sound effects [1]. The right balance between dialog and background depends on many personal factors, such as hearing acuity, listening environment, and individual taste [2]. The ideal solution is when a person can adjust the volume of the dialog content individually from the background sounds according to their personal preferences and needs. Such end-user interactivity is enabled by object-based audio systems, such as MPEG-H Audio [3], where the dialog and background can be transported as distinct objects and mixed at the receiver. The required audio tracks can be saved in new productions, but there exists a significant amount of archived content for which only the final audio mixture is available. For obtaining the object of these mixtures, source separation methods are employed [4, 5, 6]. Although audio source separation is a major research topic since many years [7], techniques applicable for broadcast content are not easily available. Among others, the main reasons for this can be identified as: 1) Broadcast

content is usually multi-channel, e.g., stereo, which makes the separation process different from the single-channel methods in the literature. 2) Broadcast content needs to be of high perceptual quality and requires high sampling rates (usually 48 kHz). 3) Real broadcast content with clean reference signals that can be used for training is generally not available to researchers.

Approaches that specifically aim at this application are still rare. Examples of solutions with non-DNN (deep neural network) signal processing include [8] extracting a phantom center channel containing the dialog from a stereo mix, [9] where source separation based on spatio-level filtering is combined with a classifier induced gating, and an earlier work from the authors [4] using a late-fusion of multiple parallel methods. Although the proposed systems are of relatively low computational complexity and show favourable subjective performance, the development of their overall performance has somewhat stagnated.

Approaches using DNNs are successful in many tasks, and can be considered as the de-facto state-of-the-art also in source separation of speech signals [10]. Even though significant progress has been made in both, single-channel [11, 12] and multi-channel speech enhancement [13, 14, 15, 16, 17], the proposed systems are still largely impractical for an application in the broadcast environment mainly being trained on low sampling rates or supporting only single input channel. The method of [18] uses an LSTM for predicting complex-valued 2D filters for speech separation with the benefit over instantaneous filters that these can handle also missing input data to some degree. Many of the speech separation models developed at 8 or 16 kHz sampling rates suffer from a model size explosion when applied on the sampling rate required by the application, because of cubical or worse dependency on the frequency resolution. The authors in [13] addresses this problem by sharing the long-short-term-memory (LSTM) parameters across the frequencies.

Making an efficient use of the limited training data is important and in [19] the authors describe a Generative Adversarial Network based system for this, but demonstrate it only for music separation. Still, this might be applicable also for dialog separation in broadcast content.

At the same time, progress in music source separation has been made by developing models trained on stereo data with high sampling rates, and some of these models are able to separate the vocals from the rest of the musical piece [20]. Assuming that speech and singing are close enough as signals and transferring knowledge from those pre-trained music networks towards dialog separation, could help to overcome multiple of the aforementioned issues.

This paper investigates the application of music source separation models and applying transfer learning for solving the

---

dialog separation task. We fine-tune three state-of-the-art neural networks (NN): Open-Unmix [21], Spleeter [22], and Conv-TasNet [23], with real-world broadcast material and compare their performance against two systems developed specifically for the task. The performance is evaluated using both computational measures (Scale-Invariant BSSEval, 2f-model), as well as a listening test simulating the intended application of enhancing the relative level of the dialog content in the mixture.

2. Problem Formulation

Let us consider a stereo audio mixture $\mathbf{y} \in \mathbb{R}^{N \times 2}$ of length N, composed of a summation of the two components $\mathbf{x} \in \mathbb{R}^{N \times 2}$ and $\mathbf{b} \in \mathbb{R}^{N \times 2}$, i.e.,

$$\mathbf{y} = \mathbf{x} + \mathbf{b}. \tag{1}$$

Signal $\mathbf{x}$ represents the dialog component consisting of one or more unknown talkers. The background signal $\mathbf{b}$ potentially includes a wide range of sounds, e.g., music and sound effects, but also ambient nature or crowd noise. The aim of dialog separation is to extract the single components $\hat{\mathbf{x}}$ and $\hat{\mathbf{b}}$, with them being close to the underlying $\mathbf{x}$ and $\mathbf{b}$, so that they can be remixed with an adapted relative level of the dialog, e.g., by a background attenuation factor μ:

$$\hat{\mathbf{y}} = \hat{\mathbf{x}} + \mu\hat{\mathbf{b}}. \tag{2}$$

As a result, the audibility and potentially also intelligibility of the dialog in $\hat{\mathbf{y}}$ is increased, ideally without decreasing the overall perceptual quality of the audio signal.

3. Separation Methods

In the following we consider three DNNs applicable for separating vocals from music with minor modification to match our data and fine-tuning them with dialog-specific data (Sec. 3.1 – 3.3). The performance of these is compared with two solutions specifically intended for dialog separation (Sec. 3.4).

3.1. Open-Unmix (UMX)

UMX [21] was created as the reference system in the SiSEC 2018 evaluation campaign [24]. The model operates in Short-time Fourier transform (STFT) domain and is a DNN-based on a three-layer bidirectional LSTM with fully-connected encoding and decoding layers to estimate the separation masks. One network is trained for each target source (vocals, bass, drums, other) using the mean-squared-error loss between the target and the separated magnitude spectrograms. During inference, a multi-channel Wiener filter post-processing step is applied. For this work the model trained on the uncompressed version of the MUSDB18 (MUSDB18-HQ) [25] dataset was used. Note, that although a specific speech enhancement version of this model is available, it was not further considered in this work, since it was trained on single-channel audio and at 16 kHz sampling rate. We selected the model with vocals target for our experiments. The model has approximately 8.9 M trainable parameters.

3.2. Spleeter

Spleeter [22] is a U-Net like convolutional NN operating in STFT-domain with 12 layers. Although each target source has its own network, these are optimized jointly by considering the combined loss. The training target is a soft mask for each source based on the corresponding magnitude spectrograms and the L_1-loss. Also here, during inference an additional Wiener filtering is performed. The network was trained on a large internal

dataset of Deezer. According to the authors, the network was only trained on a bandwidth up to 11 kHz for performance reasons. Still, we found that separation up to 22 kHz works reasonably well. For the experiments, we created our own PyTorch implementation of the model using the pre-trained weights of the 2-source-model and fine-tuning it on our data. Additionally, the batch normalization layers were frozen for the fine-tuning, because this improved the stability. The model has approximately 19.6 M trainable parameters.

3.3. Conv-TasNet for music source separation

The Conv-TasNet model for music source separation [23] is a multi-channel version of the Conv-TasNet for speaker-speaker separation [11]. The network operates in time domain by learning soft masks for each source, based on a learnt linear encoder-decoder representation using the L_1-loss of the time-domain signals. The model is fully-convolutional using dilated 1-D convolutional blocks to incorporate temporal information. There was no openly available model trained for two target components. For this reason we employed the model for four sources trained on 150 additional songs. The separation result of the three non-vocal sources was summed together to create a background component before being forwarded to the loss computation. Due to GPU memory constraints, the batch size had to be reduced to 1 for the fine-tuning. We are aware that the paper [23] proposes Demucs as an alternative for source separation with better results. However, we were unable to fit it to our GPU, which is why it was not considered further. Note, that for the remainder of this paper we use "Conv-TasNet" to refer to the music separation model. The model has approximately 11 M trainable parameters.

3.4. Systems specifically developed for dialog separation

3.4.1. Audionamix Instant Dialogue Cleaner (IDC)

IDC [26] is a commercially available plugin for dialog separation, e.g., for production or broadcasting engineers. According to the Audionamix website, it is built on DNNs and operates in real-time. For our study, the estimated dialog component was obtained by setting the plugin to maximum background attenuation (i.e., background gain was set to $-\infty$ dB, speech gain was set to 0 dB, and the *Strength* parameter was set to 0).

3.4.2. DNN Dialog Separation (DNN-DS)

The second comparison method DNN-DS is the signal separation core of a solution under development. The full system has been employed recently in field tests in Germany [27]. Since the pre- and post-processing steps of the full system have a significant effect on the output quality and they have been tuned specifically for this method, they were disabled for allowing a better comparison of plain network performance. The method uses a fully-convolutional structure in STFT-domain with a receptive field of approximately 1 s. The model has approximately 370'000 trainable parameters, and for this study, it was trained using the same data as is used for the fine-tuning of the candidate systems.

3.5. Data

The data used in this work is comprised of real-world broadcast material, hand-cleaned to obtain clean reference components by retaining signal parts with only clean speech (or silence) or non speech-like sounds in the respective temporally aligned stems.

Table 1: *Experimental results using the computational evaluation metrics. The **Input** column describes the values of the original mix. The columns containing **base** and **transfer** represent the corresponding methods without and with fine-tuning using broadcast content. **Conv-TasNet retrained** shows the results of training Conv-TasNet completely from scratch (i.e. random weights initialization instead of pre-trained weights). In the first two rows, the mean of the absolute SI-SDR and SI-SIR values are displayed. The SI-SDRi and SI-SIRi are computed as the average of the per-item improvement. The last row shows the 2f-model score of the analysed methods.*

Method	Input	IRM	Conv-TasNet retrained	UMX base	Spleeter base	Conv-TasNet base	UMX transfer	Spleeter transfer	Conv-TasNet transfer	IDC	DNN-DS
SI-SDR [dB]	6.5 ± 7.0	16.8 ± 5.6	15.5 ± 5.8	10.9 ± 5.7	12.0 ± 5.9	13.8 ± 5.2	13.0 ± 5.7	14.4 ± 6.1	**16.8 ± 5.1**	5.1 ± 5.5	14.5 ± 5.9
SI-SIR [dB]	6.5 ± 7.1	22.4 ± 6.6	23.9 ± 6.1	18.8 ± 7.4	18.4 ± 6.1	19.1 ± 6.4	20.4 ± 6.8	23.5 ± 7.0	**28.4 ± 6.8**	20.6 ± 8.7	23.6 ± 7.5
SI-SDRi [dB]		10.3 ± 4.2	9.0 ± 3.9	4.4 ± 5.1	5.5 ± 3.7	7.3 ± 4.2	6.5 ± 4.6	7.9 ± 4.1	**10.3 ± 4.2**	-1.5 ± 4.6	8.0 ± 4.1
SI-SIRi [dB]		15.9 ± 6.9	17.4 ± 7.5	12.3 ± 6.9	12.0 ± 6.6	12.6 ± 7.7	13.9 ± 7.5	17.1 ± 8.3	**21.9 ± 8.8**	14.1 ± 6.6	17.1 ± 8.3
2f-model	20.6 ± 11.9	**70.1 ± 9.3**	34.3 ± 6.9	31.0 ± 8.7	33.9 ± 11.6	36.8 ± 11.8	35.1 ± 9.7	38.6 ± 13.4	46.4 ± 12.2	23.8 ± 8.2	35.0 ± 12.4

All items were in stereo at 48 kHz sampling rate. The content ranged from nature documentaries to sports and drama. The data contained both female and male talkers in multiple languages, though the majority was in German. As a pre-processing step the data was downsampled to 44.1 kHz to match the network requirements. Then the training items in the dataset were cut into 15 s chunks. The dataset was separated into a training (15 h), a validation (1.45 h) and a test set (0.9 h).

3.6. Training strategy

The input training samples were randomly selected from each clip. Since the subject networks utilize a different amount of input data (UMX 6 s, Conv-TasNet 2.8 s and Spleeter 12 s), we wanted to make sure that each model was trained on a similar amount of data. Therefore, per epoch the items of UMX and Conv-TasNet were sampled two and four times respectively to match the duration of input data for Spleeter.

A learning rate decay by a factor of 0.3 and a patience of 5 epochs was included into training. The maximum number of epochs was limited to 100. To prevent overfitting, an additional early-stopping mechanism was implemented, stopping the training if the validation loss did not decrease for 10 consecutive epochs. All data augmentation was turned off, since the methods used varied in between the different networks.

4. Evaluation Methods

The performance of the models is evaluated before and after fine-tuning with computational evaluation metrics (Sec. 4.1), as well as with subjective evaluation via a listening test (Sec. 4.2).

4.1. Computational evaluation metrics

To quantify the outcome of our experiments we compute the improvement in Scale-Invariant Signal-to-Distortion-Ratio (SI-SDRi) and Signal-to-Inference-Ratio (SI-SIRi) [28] in dB between the separated and reference components of the test set. In addition, the average SI-SDR and SI-SIR values of the input mix as well as the corresponding output values of the separated dialog components are reported. As a further metric we computed the 2f-model score [29]. It is a value in the range 0 – 100 approximating the result of a MUSHRA listening test, and it has been shown to correlate well with perceptual audio quality for source separation [30]. Since oracle ideal ratio mask (IRM) is often considered as the training target for source separation DNNs, we computed the metrics for it as an additional orientation point. The IRM results were obtained using a sine window of 43 ms, 50% overlap and an exponent of 2 defining a Wiener filter-like separation mask.

4.2. Listening test

A listening test was conducted with a subset of 8 items to evaluate the perceptual quality. The items were selected to obtain 4 items with female and 4 with male talkers each, as well as 4 in German and 4 in English language; 4 items featured music as background, while 4 items featured noisy background, e.g., atmospheres from forest, seaside, or football stadium. Each item was 8 s long. The computational evaluation was repeated on the selected items, confirming that the selection was not biased towards a specific model. The separated components were mixed to simulate a dialog enhancement application scenario by:

1. Create a reference condition by attenuating the background component by 12 dB compared to the input mix.

2. Calculate the integrated loudness (as per [31]) of the reference background considering only the non-dialog parts, with gating deactivated.

3. For each separation method tested (i.e., test condition), obtain the background component by subtracting the estimated dialog component from the input mixture.

4. For each test condition, determine a background attenuation to obtain the same background loudness during dialog inactivity as in the reference condition. Maximal attenuation is 120 dB.

5. Normalize each item to -23 LUFS integrated loudness (gating deactivated).

The test was conducted following the MUSHRA methodology [32]. Including a 3.5 kHz low-pass anchor and the hidden reference, the test consisted of 10 conditions. Note, that the IRM was not included as a condition into the listening test, since it is an ideal method, not available in practice. Due to the COVID-19 restrictions, the test had to be conducted at home office, each participant with their own computer and headphones. The test participants were fellow post-graduate students and colleagues of the authors, with various levels of experience in subjective evaluation of audio and at least a basic knowledge of German and English. The authors did not participate in the test. The test results from 14 listeners were collected, and 6 participants were post-screened because they did not always rate the hidden reference as the best condition.

5. Experimental results

5.1. Computational evaluation metrics

The results of the computational evaluation metrics are displayed in Table 1. All the fine-tuned models outperform the respective base versions showing a benefit introduced by the
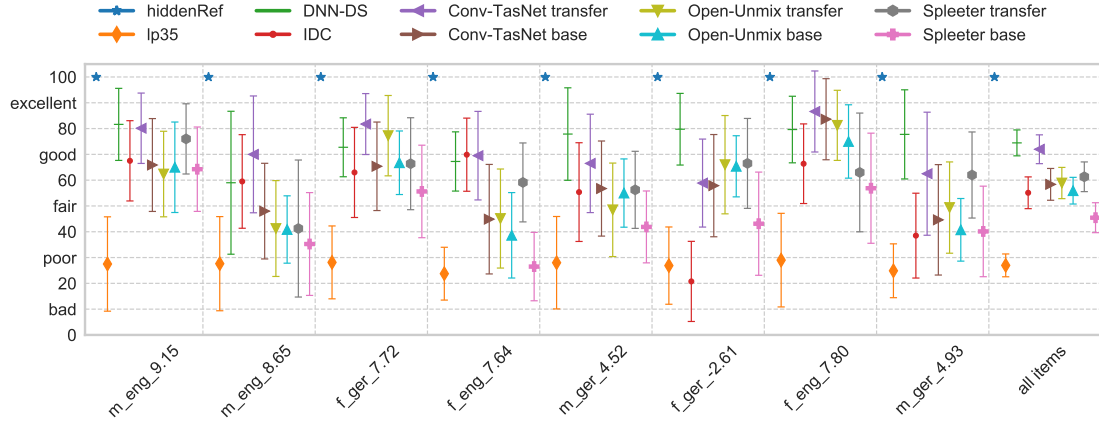
Figure 1: *Listening test results. Mean and 95% confidence interval (using Student's t-distribution, 8 listeners) are depicted per item and test condition. The test items include female (**f**) and male (**m**), as well as German (**ger**) and English (**eng**) samples, and different input SI-SIR (reported at the end of the each sample name). On the right, the average results are displayed. (Best viewed in colors)*

task-specific fine-tuning step and suggesting that transfer learning from music separation towards dialog separation is possible.

UMX has the lowest SI-SDRi and 2f-model scores both before and after fine-tuning. One possible explanation for this is the smaller amount of training data UMX uses in the original training. Considering SI-SIRi, UMX shows a better performance for the base model than the base version of Spleeter but, after fine-tuning the Spleeter results exceed UMX by a large margin. Conv-TasNet consistently shows better results than the other two methods before and after fine-tuning. The margin even increases after fine-tuning suggesting that Conv-TasNet benefits from this more than the other models. IDC shows a better performance than the base networks and the fine-tuned UMX considering SI-SIRi, but after fine-tuning it is being outperformed by Spleeter and Conv-TasNet. The SI-SDRi for IDC is negative indicating that the processing induces significant artifacts. Inspecting some of the output signals confirms this and reveals that occasionally the dialog components are missed entirely. It has to be noted that IDC operates causally in real-time, and therefore has more challenging processing constraints than the other methods. The performance of DNN-DS is significantly better than both versions of UMX, close to the fine-tuned Spleeter, and is outperformed by Conv-TasNet in the computational metrics. In average SI-SIRi, Conv-TasNet exceeds DNN-DS by 4 dB indicating a better dialog/background separation. IRM is outperformed by Spleeter, Conv-TasNet and DNN-DS in SI-SIRi while showing the best 2f-model results by a large margin. The beneficial effect of transfer learning can be further highlighted by considering a Conv-TasNet model that was trained from scratch (random weights initialization) using the same parameters and data. The metrics in Table 1 indicate that training from scratch can outperform the respective base model for the task at hand, but performs worse than the fine-tuned version pretrained on music. Similar behaviour was observed with Spleeter and UMX.

5.2. Listening test

The results from the listening test are displayed in Figure 1. It can be seen that DNN-DS reaches the best performance with a slight advantage over Conv-TasNet. The advantage is particularly noticeable in the test items with lower input SI-SIR

(<5 dB), while the networks are on par or Conv-TasNet takes the lead for higher input SI-SIR (≥5 dB). The overall slight advantage of DNN-DS is contrary to the computational evaluation where Conv-TasNet was ahead in all metrics, reminding that even state-of-the-art computational evaluation measures may give results different from subjective evaluation [30]. One thing to note is that Conv-TasNet uses almost 30 times more trainable parameters than DNN-DS to reach this performance. Furthermore, we assume that Conv-TasNet benefits significantly from the remixing in the application. Inspecting some separated dialog samples reveals quite audible clipping-like artifact. Remixing the dialog component with an attenuated version of the background component appears to hide these artifacts reasonably well. A point for future work is to test the method proposed in [33] replacing the learned encoder and decoder with STFT. This should both reduce the model size considerably and address the distinctive timbral artifact.

IDC performs better in the listening test than the base UMX and the base Spleeter, but stays behind the other methods. This confirms the low scores observed in the computational evaluation metrics. The fine-tuned Spleeter network shows still the second best results of the music models, but with an increased margin from the best performing models. This could be explained by phantom speech or chorusing-like artifacts that Spleeter introduces. This effect is even emphasized in the base Spleeter version, showing the lowest listening test scores of all music networks. The categorization of UMX is largely confirmed by the listening test.

6. Conclusion

This paper described a comparison of applying three state-of-the-art music separation DNN models for dialog separation before and after task-specific fine-tuning. When evaluated in the remixing application, the considered DNNs range between perceptual quality levels described as fair and good. This shows that dialog separation can be solved to a degree using transfer learning from music separation, with competitive performance compared to solutions specifically developed for the task. The results also showed that the methods all benefit to varying degrees from fine-tuning training with task-specific material.

7. References

[1] M. Armstrong, "From Clean Audio to Object Based Broadcasting," *BBC Research & Development White Paper, WHP324*, 2016. [Online]. Available: http://www.bbc.co.uk/rd/publications/whitepaper324

[2] D. Geary, M. Torcoli, J. Paulus, C. Simon, D. Straninger, A. Travaglini, and B. Shirley, "Loudness Differences for Voice-Over-Voice Audio in TV and Streaming," *Journal of the Audio Engineering Society*, vol. 68, no. 11, pp. 810–818, 2020.

[3] C. Simon, M. Torcoli, and J. Paulus, "MPEG-H Audio for Improving Accessibility in Broadcasting and Streaming," *arXiv preprint arXiv:1909.11549*, 2019.

[4] J. Paulus, M. Torcoli, C. Uhle, J. Herre, S. Disch, and H. Fuchs, "Source Separation for Enabling Dialogue Enhancement in Object-based broadcast with MPEG-H," *Journal Audio Eng. Soc.*, vol. 67, no. 7/8, pp. 510–521, 2019.

[5] W. Lim, S. Beack, and T. Lee, "Multi-level stereo attention model for center channel extraction," in *Proc. of 2019 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2019.

[6] C. Uhle, O. Hellmuth, and J. Weigel, "Speech Enhancement of Movie Sound," in *Proc. Audio Eng. Soc. Convention*, 2008.

[7] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement.* Wiley, 2018.

[8] J. T. Geiger, P. Grosche, and Y. L. Parodi, "Dialogue enhancement of stereo sound," in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2015, pp. 869–873.

[9] A. S. Master, L. Lu, H.-M. Lehtonen, H. Mundt, H. Purnhagen, and D. Darcy, "Dialog Enhancement via Spatio-Level Filtering and Classification," in *Proc. Audio Eng. Soc. Convention*, 2020.

[10] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[11] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[12] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," *ACM Trans. Graph.*, vol. 37, no. 4, 2018.

[13] X. Li and R. Horaud, "Multichannel Speech Enhancement Based On Time-Frequency Masking Using Subband Long Short-Term Memory," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, pp. 298–302.

[14] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-Supervised Multichannel Speech Enhancement With a Deep Speech Prior," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2197–2212, 2019.

[15] C. Liu, S. Fu, Y. Li, J. Huang, H. Wang, and Y. Tsao, "Multichannel Speech Enhancement by Raw Waveform-Mapping Using Fully Convolutional Networks," *IEEE/ACM Transactions on Speech and Language Processing*, vol. 28, pp. 1888–1900, 2020.

[16] O. X. Luo, "Deep Learning for Speech Enhancement - A Study on WaveNet, GANs, and General CNN-RNN Architectures," Master's thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2019.

[17] A. Pandey and D. Wang, "On Adversarial Training and Loss Functions for Speech Enhancement," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5414–5418.

[18] W. Mack and E. A. P. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Signal Process. Lett.*, vol. 27, pp. 61–65, 2020.

[19] D. Stoller, S. Ewert, and S. Dixon, "Adversarial Semi-Supervised Audio Source Separation Applied to Singing Voice Extraction," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2391–2395.

[20] E. Cano, D. Fitzgerald, A. Liutkus, M. D. Plumbley, and F. Stöter, "Musical Source Separation: An Introduction," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 31–40, 2019.

[21] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - A Reference Implementation for Music Source Separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[22] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.

[23] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music Source Separation in the Waveform Domain," *arXiv preprint arXiv:1911.13254*, 2019.

[24] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 Signal Separation Evaluation Campaign," in *International Conference on Latent Variable Analysis and Signal Separation*, 2018, pp. 293–30.

[25] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "MUSDB18-HQ - an uncompressed version of MUSDB18," 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373

[26] Instant Dialogue Cleaner (IDC) plugin. [Online]. Available: https://audionamix.com/instant-dialogue-cleaner/

[27] M. Torcoli, C. Simon, J. Paulus, D. Straninger, A. Riedel, V. Koch, S. Wirts, D. Rieger, H. Fuchs, C. Uhle, and S. Meltzer, "Dialog+ in Broadcasting: First Field Tests using Deep-Learning Based Dialogue Enhancement," in *Int. Broadcasting Conv. (IBC) Technical Papers*, 2021, under review.

[28] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – Half-baked or Well Done?" in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[29] T. Kastner and J. Herre, "An Efficient Model for Estimating Subjective Quality of Separated Audio Source Signals," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, pp. 95–99.

[30] M. Torcoli, T. Kastner, and J. Herre, "Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of their Application Domain Dependence," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1530–1541, Mar. 2021.

[31] International Telecommunication Union, "Recommendation ITU–R BS.1770-4 Algorithms to measure audio programme loudness and true-peak audio level," 2015.

[32] ——, "Recommendation ITU–R BS.1534-3 Method for the subjective assessment of intermediate quality level of audio systems," 2015.

[33] F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, and D. Yu, "A comprehensive study of speech separation: spectrogram vs waveform separation," in *Proc. Interspeech Conf.*, 2019, pp. 4574–4578.