INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

End-to-End Monaural Speech Separation with Multi-Scale Dynamic Weighted Gated Dilated Convolutional Pyramid Network

Ziqiang Shi[1], *Huibin Lin*[1], *Liu Liu*[1], *Rujie Liu*[1], *Shoji Hayakawa*[2], *Shouji Harada*[2], *Jiqing Han*[3]

[1]Fujitsu Research and Development Center
[2]Fujitsu Laboratories Ltd.
[3]Harbin Institute of Technology
shiziqiang@cn.fujitsu.com

Abstract

The monaural speech separation technology is far from satisfactory and has been a challenging task due to the interference of multiple sound sources. While deep dilated temporal convolutional networks (TCN) have been proved to be very effective in sequence modeling, this work investigates how to extend TCN to result in a new state-of-the-art approach for monaural speech separation. First a novel gating mechanisms is introduced and added to result in gated TCN. The gated activation can control the flow of information. Further in order to remedy the temporal scale variation problem caused by word length and pronunciation characteristics of different people, a multi-scale dynamic weighted pyramids gated TCNs is proposed, where a "weightor" network is used to determine the weights of different gated TCNs dynamically for each utterance. Since the strengths of different branches with different temporal receipt fields appear complementary, the combination outperforms single branch system. For the objective, we propose to train the network by directly optimizing utterance level signal-to-distortion ratio (SDR) in a permutation invariant training (PIT) style. Our experiments on the the WSJ0-2mix data corpus results in 18.4dB SDR improvement, which shows our proposed networks can leads to performance improvement on the speaker separation task.

Index Terms: speech separation, cocktail party problem, temporal convolutional neural network, gating mechanism, pyramid network

1. Introduction

Multi-talker monaural speech separation has a wide range of applications. For example, in a home environment or conference environment where many people talk, the human auditory system can easily track and follow the speech of the target speaker from the mixed speech of multiple speakers. In this case, it is necessary to separate the clean speech signal of the target speaker from the mixed speech to complete the subsequent recognition work. There are two difficulties with this task. The first problem is that since we don't have any prior information about the user, a practical system must be independent of the speaker. The second difficulty is that there is no way to use the beamforming algorithm for a single microphone signal. Many traditional methods, such as Computational Auditory Scene Analysis (CASA) [1, 2, 3], non-negative matrix factorization (NMF) [4, 5] and probability model [6], do not solve these two difficulties well.

Recently, a large number of deep learning-based techniques have been proposed for this task. These methods can be briefly grouped into three categories. The first type is based on deep clustering (DPCL) [7, 8], which maps the time-frequency (TF) points of the spectrogram to the embedding vectors, and then these embedding vectors are clustered into several classes corresponding to different speakers, and finally these clusters are used as masks to inversely transform the spectrogram to the separated clean voices; the second is the permutation invariant training (PIT) [9, 10], which minimizes the lowest error output in all possible permutations of the N mixed sources allocation to solve the label permutation problem; the third category is end-to-end speech separation in the time domain [11, 12, 13, 14], which is a natural way to overcome the obstacles of the upper bound source-to-distortion ratio improvement (SDRi) in short-time Fourier transform (STFT) mask estimation based methods and real-time processing requirements in actual use.

This paper is based on the end-to-end approach [11, 12, 13, 14], which has achieved better results than DPCL based or PIT based methods. Since most DPCL and PIT based methods use STFT as the representations, which has several limitations. Firstly, it is not clear whether the STFT is optimal transformation of the signal for speech separation [15]. Secondly, most STFT-based methods generally assume that the phase of the separated signal is equal to the mixture phase, which is usually incorrect and imposes an obvious upper limit on the separation performance by using the ideal masks. As a method for overcoming the above problems, several speech separation models have recently been proposed, which run directly on the time domain speech signal [11, 12, 13, 14]. Recurrent neural networks (RNN), including long short-term memory (LSTM) are the preferred methods for simulating time-dependent sequences. However, one of their main drawbacks is the exploding and vanishing gradient problem and the difficulty of parallel training and separation. Additionally, recent literature also suggests that feed-forward convolutional models empirically out-perform recurrent models while being parallelizable and easier to train with more stable gradients [16]. Inspired by these preliminary results, we propose FurcaPorta and FurcaPy[1]. FurcaPorta added a novel gating mechanism to the TCN, where in each dilated convolutional module corresponding to each dilated factor, two gates are introduced, one gate controls the inflow of information, and one gate controls the processing and outflow of information. Furthermore, due to the influence of different word lengths or different speech speeds, multiple branches of a variety of temporal receipt field scales are introduce to characterize speech, and the weights of different scales are automatically determined by a "weightor" network.

[1]"Furca" is Latin for "fork", and we use this word to mean the mixed speech is split into two streams by our network like water. "Porta" is Latin for "gate". "Py" is the abbreviation of Pyramid.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-1292

The remainder of this paper is organized as follows: Section 2 introduces monaural speech separation with TCN. Section 3 describe our proposed FurcaPorta, FurcaPy and the separation algorithm in detail. The experimental setup and results are presented in Section 4. We conclude this paper in Section 5.

2. Speech separation with TCN

In this section, we will review the formal definition of the monaural speech separation task and the original TCN architecture.

The goal of monaural speech separation is to estimate the individual target signals from a linearly mixed single-microphone signal, where the target signals overlap in the TF domain. Let $x_i(t), i = 1, .., S$ denote the S target speech signals and $y(t)$ denotes the mixed speech respectively. If we assume the target signals are linearly mixed, which can be represented as:

$$y(t) = \sum_{i=1}^{S} x_i(t),$$

then monaural speech separation aims at estimating individual target signals from given mixed speech $y(t)$. In this work it is assumed that the number of target signals is known.

In order to deal with this ill-posed problem, Luo et al. [12] introduce TCN [17, 16] to do this task. TCN is proposed as an alternative to RNN in various tasks [17, 16]. Each layer in the TCN contains a 1-D convolution block with an increased dilation factor. The dilation factor is increased exponentially to ensure a suitable large time context window to take advantage of the long-range dependence of the speech signal, as shown in Figure 1. Dilated convolution in WaveNet has been a huge success for audio generation [18]. Dilated convolutions with different dilation factors have different receptive fields. Stacked dilated convolution provides a very large receptive fields for the network, with only a few layers, because the dilation range grows exponentially. This allows the network to capture the temporal dependence of various resolutions with input sequences. TCN introduces a temporal hierarchy: the upper layer can access longer input subsequences and learn representations on a larger time scale. Local information from lower layers is propagated through the hierarchy through residuals and skip connections.

There are two important elements in the original TCN [16] as shown in Figure 1, one is the dilated convolutions, and the other is residual connections. Dilated convolutions follow the work of [18], it is defined as

$$(x *_d k)(p) = \sum_{s+dt=p} x(s)k(t),$$

where x is the 1-D input signal, k is the filter(a.k.a. kernel), and d is the dilation factor. Therefore, dilation is equivalent to introducing a fixed step size between every two adjacent filter taps. The general way to increase the receipt field of the TCN is to increase the dilation factor d. In this work we increase d exponentially with the depth of the network and $d = 2$ as shown in Figure 1, and this TCN has four layers of 1-D Conv modules with dilation factors of $1, 2, 4, 8$ respectively. As shown in Figure 1, each 1-D Conv module is a residual block [19], which contains one layer of dilated convolution (Depth wise conv [20]), two layers of $1\times$ 1 convolutions ($1\times$ 1 Conv), two non-linearity activation layers (parametric rectified linear unit,

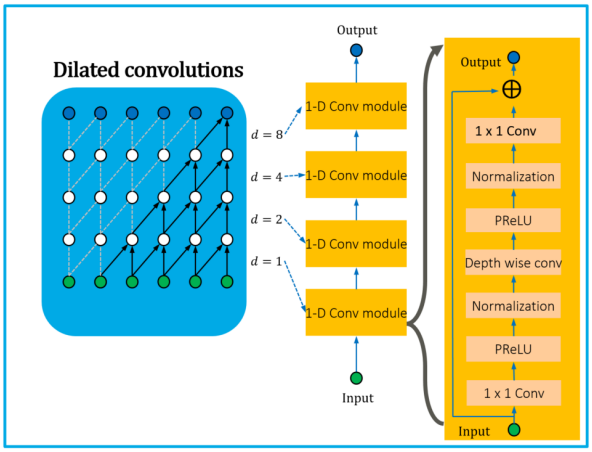

Figure 1: *The structure of TCN.*

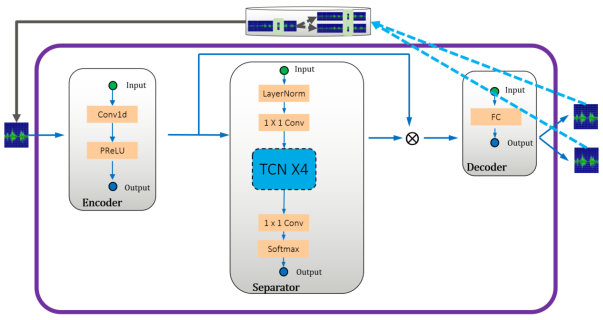

Figure 2: *The pipeline of TCN based speech separation in [12].*

PReLU [21]), and two normalization layers (Normalization). For normalization, we applied global normalization [12] to the convolutional filters.

Luo et al. proposed a TCN based speech separation method [12], which consists of three processing stages, as shown in Figure 2: encoder (Conv1d is followed by a PReLU), separator (consisted in the order by a LayerNorm, a 1×1conv, 4 TCN layers, 1×1conv, and a softmax operation) and decoder (a FC layer). First, the encoder module is used to convert short segments of the mixed waveform into their corresponding representations. Then, the representation is used to estimate the multiplication function (mask) of each source and each encoder output for each time step. The source waveform is then reconstructed by transforming the masked encoder features using a linear decoder module.

3. End-to-end Speech separation with FurcaPorta and FurcaPy

The main work of this paper is to make several improvements to the TCN (Figure 1) and TCN based framework (Figure 2) for speech separation. First, we introduced a novel gating mechanism in this TCN, as shown in Figure 3. Nonlinear gated activation had been used in prior work on sequence modeling [18, 22], it can control the flow of information and may help the network to model more complex interactions. Two gates are added to each 1-D convolutional module in the plain TCN, one gate is corresponding to the first 1×1 convolutional

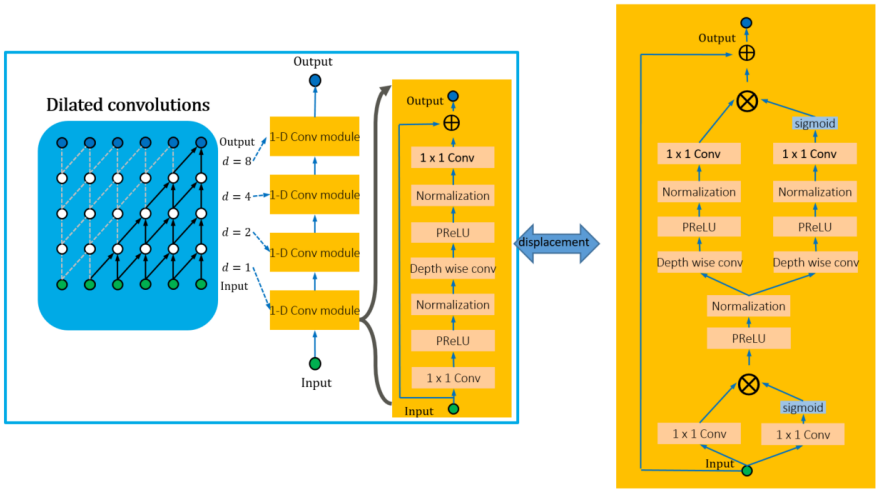

Figure 3: *The structure of gated TCN.*

layer in the 1-D convolutional module and this gate is used to control the inflow information. The other gate is corresponding to all the layers from the depth-wise convolutional layer to the output 1×1 convolutional layer and this gate is used to control the processing and outflow of information. This gated TCN based speech separation pipeline is called FurcaPorta in this work.

3.1. FurcaPy: Multi-scale dynamic weighted gated dilated convolutional pyramids network

Since in real life the utterance always have the feature of temporal scale variation caused by different word lengths and pronunciation characteristics (e.g. speed) of different people, thus different temporal receipt fields may help in speech separation. The temporal receipt field is fixed in previous network structure. In order to remedy the temporal scale variation problem, a multi-scale dynamic weighted **py**ramids gated TCNs based pipeline which is called FurcaPy is proposed as shown in Figure 4 and there are three kinds of different temporal receipt fields in this description. FurcaPy's encoder and decoder are the same as the previous FurcaPorta, they differ only in the separator. In the separator of FurcaPy, each branch in the pyramid consists of a different number of gated TCNs. The length of the temporal receptive field of each branch is several times the length of the temporal receiving field of a single gated TCN. If the receptive field of a single gated TCN is assumed to be L, then the length of the receptive field of all branches in the Figure 4 is $3L$,$4L$, and $5L$ respectively. The total output is obtained by weighted averaging the outputs of the different branches corresponding to different receipt fields. Additionally, a "weightor" module is designed to determine which temporal receipt field is more suitable for current input utterance signal, that means the weights of different gated TCNs are determined dynamically by a "weightor" network for each utterance. The "weightor" is composed of a common multi layer 1-D convolutional neural network as shown in Figure 4 and it consist of Conv1d, PReLU, LayerNormal, 3 layers of 1×1 Conv and max pooling, and Softmax.

3.2. Perceptual metric: Utterance-level SDR objective

Since the loss function of many STFT-based methods is not directly applicable to waveform-based end-to-end speech separation, perceptual metric based loss function is tried in this work. The perception of speech is greatly affected by distortion [23, 24]. Generally in order to evaluate the performance of speech separation, the BSS_Eval metrics signal-to-distortion ratio (SDR), signal-to-Interference ratio (SIR), signal-to-artifact ratio (SAR) [25, 26], and short-time objective intelligibility (STOI) [27] have been often employed. In this work we directly use SDR, which is most commonly used metrics to evaluate the performance of source separation, as the training objective. SDR measures the amount of distortion introduced by the output signal and define it as the ratio between the energy of the clean signal and the energy of the distortion.

SDR captures the overall separation quality of the algorithm. There is a subtle problem here. We first concatenate the outputs of FurcaPy into a complete utterance and then compare with the input full utterance to calculate the SDR in the utterance level instead of calculating the SDR for one frame at a time. These two methods are very different in ways and performance. If we denote the output of the network by s, which should ideally be equal to the target source x, then SDR can be given as [25, 26]

$$\begin{aligned} \tilde{x} &= \frac{\langle x, s\rangle}{\langle x, x\rangle} x, \\ e &= \tilde{x} - s, \\ \text{SDR} &= 10 * \log_{10} \frac{\langle \tilde{x}, \tilde{x}\rangle}{\langle e, e\rangle}. \end{aligned}$$

Then our target is to maximize SDR or minimize the negative SDR as loss function respect to the s.

In order to solve tracing and permutation problem, the PIT training criteria [9, 10] was employed in this work. We calculate the SDRs for all the permutations, pick the maximum one, and use the negative as the loss. In this work, it is called the uSDR loss.

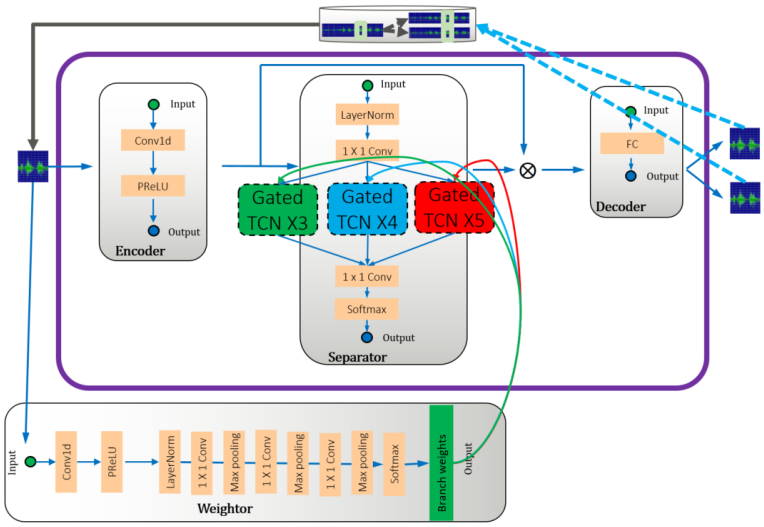

Figure 4: *The structure of FurcaPy.*

4. Experiments

4.1. Dataset and neural network

We evaluated our system on two-speaker speech separation problem using WSJ0-2mix dataset [7, 8], which contains 30 hours of training and 10 hours of validation data. The mixtures are generated by randomly selecting 49 male and 51 female speakers and utterances in Wall Street Journal (WSJ0) training set si_tr_s, and mixing them at various signal-to-noise ratios (SNR) uniformly between 0 dB and 5 dB. 5 hours of evaluation set is generated in the same way, using utterances from 16 unseen speakers from si_dt_05 and si_et_05 in the WSJ0 dataset.

4.2. Results

We evaluate the systems with the SDR improvement (SDRi) [25, 26] metrics used in [8, 28, 29, 30, 9]. The original SDR, that is the average SDR of mixed speech $y(t)$ for the original target speech $x_1(t)$ and $x_2(t)$ is 0.15.

$$M_s = \frac{|X_s(t,f)|}{\sum_{s=1}^{S}|X_s(t,f)|} \tag{1}$$

applied to the STFT $Y(t,f)$ of $y(t)$ to obtain the separated speech, which is evaluated to show the upper bounds of STFT based methods, where $X_s(t,f)$ is the STFT of $x_s(t)$.

In this experiment, as baselines, we reimplemented several classical approaches DPCL [7], TasNet [11] and Conv-TasNet [12]. Table 1 lists the average SDRi obtained by our methods and almost all the results in the past two years, where IRM means the ideal ratio mask. Compared with these baselines an average increase of 2dB SDRi is obtained. FurcaPorta and FurcaPy has achieved the most significant performance improvement compared with baseline systems, and it break through the upper bound of STFT based methods a lot (nearly 6dB).

Table 1: *SDRi (dB) in a comparative study of different separation methods on the WSJ0-2mix dataset. * indicates our reimplementation of the corresponding method.*

Method	SDRi
DPCL [7]	5.9
uPIT-BLSTM [10]	10.0
cuPIT-Grid-RD [29]	10.2
DANet [30]	10.5
ADANet [28]	10.5
DPCL*	10.7
DPCL++ [8]	10.8
CBLDNN-GAT [31]	11.0
TasNet [11]	11.2
TasNet*	11.8
Chimera++ [32]	12.0
IRM	12.7
FurcaNet	13.3
Conv-TasNet [12]	15.0
Conv-TasNet*	15.8
FurcaPorta	17.3
FurcaPy	18.4

5. Conclusion

In this paper we investigated the effectiveness of deep dilated temporal convolutional networks modeling for multi-talker monaural speech separation. Benefits from the strength of end-to-end processing, the novel gating mancinism and dynamic multiple temporal receipt fields modeling of sequence signal, FurcaPorta and FurcaPy leads to 9% and 16% relative improvement, and we achieve the new state-of-the-art on the public WSJ0-2mix data corpus. For further work, although SDR is widely used and can be useful, but it has some weaknesses [33]. In the future, maybe we can use SNR to evaluation our models. It would be interesting to see how consistent the SDR and SNR are.

6. Acknowledgment

We would like to thank Jian Wu at Northwestern Polytechnical University, Yi Luo at Columbia University, and Zhong-Qiu Wang at Ohio State University for valuable discussions on WSJ0-2mix database, DPCL, and end-to-end speech separation.

7. References

[1] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE press, 2006.

[2] Y. Shao and D. Wang, "Model-based sequential organization in cochannel speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 289–298, 2006.

[3] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Transactions on audio, speech, and language processing*, vol. 21, no. 1, pp. 122–131, 2013.

[4] P. Smaragdis *et al.*, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on audio speech and language processing*, vol. 15, no. 1, p. 1, 2007.

[5] J. Le Roux, F. J. Weninger, and J. R. Hershey, "Sparse nmf–half-baked or well done?" *Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023*, 2015.

[6] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Ninth International Conference on Spoken Language Processing*, 2006.

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 31–35.

[8] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[9] M. Kolbæk, 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/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

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

[11] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," *arXiv preprint arXiv:1711.00541*, 2017.

[12] ——, "Tasnet: Surpassing ideal time-frequency masking for speech separation," *arXiv preprint arXiv:1809.07454*, 2018.

[13] S. Venkataramani, J. Casebeer, and P. Smaragdis, "Adaptive front-ends for end-to-end source separation," in *Proc. NIPS*, 2017.

[14] Z. Shi, H. Lin, L. Liu, R. Liu, S. Hayakawa, S. Harada, and J. Han, "Furcanet: An end-to-end deep gated convolutional, long short-term memory, deep neural networks for single channel speech separation," *arXiv preprint arXiv:1902.00651*, 2019.

[15] Z. Shi, H. Lin, L. Liu, R. Liu, and J. Han, "Is cqt more suitable for monaural speech separation than stft? an empirical study," *arXiv preprint arXiv:1902.00631*, 2019.

[16] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[17] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.

[18] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[22] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[23] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of the modified bark spectral distortion as an objective speech quality measure," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 541–544.

[24] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech processing in the auditory system*. Springer, 2004, pp. 231–308.

[25] C. Févotte, R. Gribonval, and E. Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4214–4217.

[28] Y. Luo, Z. Chen, and N. Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 4, pp. 787–796, 2018.

[29] C. Xu, W. Rao, X. Xiao, E. S. Chng, and H. Li, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," pp. 6–10, 2018.

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

[31] C. Li, L. Zhu, S. Xu, P. Gao, and B. Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," 2018.

[32] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[33] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" *arXiv preprint arXiv:1811.02508*, 2018.