FurcaNeXt: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks

Ziqiang Shi*[1], Huibin Lin[1], Liu Liu[1], Rujie Liu[1], Jiqiang Han[2], and Anyan Shi[3]

[1]Fujitsu Research & Development Center, Beijing, China
[2]Harbin Institute of Technology, Harbin, China
[3]Shuangfeng First, Beijing, China

Feb., 2019

Abstract

Deep dilated temporal convolutional networks (TCN) have been proved to be very effective in sequence modeling. In this paper we propose several improvements of TCN for end-to-end approach to monaural speech separation, which consists of 1) multi-scale dynamic weighted gated dilated convolutional pyramids network (FurcaPy), 2) gated TCN with intra-parallel convolutional components (FurcaPa), 3) weight-shared multi-scale gated TCN (FurcaSh), 4) dilated TCN with gated difference-convolutional component (FurcaSu), that all these networks take the mixed utterance of two speakers and maps it to two separated utterances, where each utterance contains only one speaker's voice. 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 public 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.

1 Introduction

Multi-talker monaural speech separation has a vast range of applications. For example, a home environment or a conference environment in which many people talk, the human auditory system can easily track and follow a target speaker's voice from the multi-talker's mixed voice. In this case, a clean speech signal of the target speaker needs to be separated from the mixed speech to complete the subsequent recognition work. Thus it is a problem that must be solved in order to achieve satisfactory performance in speech or speaker recognition tasks. There are two difficulties in this problem, the first is that since we don't have any priori information of the user, a truly practical system must be speaker-independent. 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) [30, 20, 10], Non-negative matrix factorization (NMF) [23, 13], and probabilistic models [29], do not solve these two difficulties well.

More recently, a large number of techniques based on deep learning are proposed for this task. These methods can be briefly grouped into three categories. The first category is based on deep clustering (DPCL) [8, 11], which maps the time-frequency (TF) points of the spectrogram into the embedding vectors, 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) [12, 35], which solves the label permutation problem by minimizing the lowest error output among all possible permutations for N mixing sources assignment; the third category is end-to-end speech separation in time-domain [17, 18, 27], which is a natural way to overcome the obstacles of the upper bound source-to-distortion ratio improvement (SDRi) in short-time

---

*Corresponding author: shiziqiang@cn.fujitsu.com; shiziqiang7@gmail.com

Fourier transform (STFT) mask estimation based methods and real-time processing requirements in actual use.

This paper is based on the end-to-end method [17, 18, 27], which has achieved better results than DPCL based or PIT based approaches. Since most DPCL and PIT based methods use STFT as front-end. Specifically, the mixed speech signal is first transformed from one-dimensional signal in time domain to two-dimensional spectrum signal in TF domain, and then the mixed spectrum is separated to result in spectrums corresponding to different source speeches by a deep clustering or mask estimation method, and finally the cleaned source speech signal can be restored by an inverse STFT on each spectrum. This framework has several limitations. Firstly, it is unclear whether the STFT is the optimal (even assume the parameters it depends on are optimal, such as size and overlap of audio frames, window type and so on) transformation of the signal for speech separation. Secondly, most STFT based methods often assumed that the phase of the separated signal to be equal to the mixture phase, which is generally incorrect and imposes an obvious upper bound on separation performance by using the ideal masks. As an approach to overcome the above problems, several speech separation models were recently proposed that operate directly on time-domain speech signals [17, 18, 27]. Inspired by these first results, we propose FurcaNeXt, which is a general name for a series of fully end-to-end time-domain separation methods, includes 1) multi-scale dynamic weighted gated dilated convolutional pyramids network (FurcaPy): 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; 2) deep gated dilated temporal convolutional networks (TCN) with intra-parallel convolutional components (FurcaPa): replace two convolutional related modules in each dilated convolutional module by two intra-parallel convolutional modules , which can reduce the variance of this model. The intra-parallel convolutional modules replicate weight matrices and take the average from the feature maps produced by those layers. This convenient technique can effectively improve separation performance. 3) weight-shared multi-scale gated TCN (FurcaSh): a simple design is proposed to achieve the functions of FurcaPy but without increasing the number of network parameters. 4) dilated TCN with gated difference-convolutional component (FurcaSu): inspired by the work of Highway network [24], in which two additional non-linear transformations acts as gates that can dynamically pass part of its inputs and suppress the other part, conditioned on the input itself. Authors simplify the Highway network through multiple ways [34]. After further simplification we propose to use two identical transformation function branches to implemented a simplified version of the highway network module.

The remainder of this paper is organized as follows: section 2 introduces monaural speech separation with TCN. Section 3 describe our proposed FurcaNeXt 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 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, in which 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. [18] introduce TCN [14, 2] to do this task. TCN is proposed as an alternative to RNN in various tasks [14, 2]. 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 temporal context window to take advantage of the long range dependence of the speech signal, as shown in Figure 1. Dilated convolution has made a huge success in WaveNet for audio generation [26].
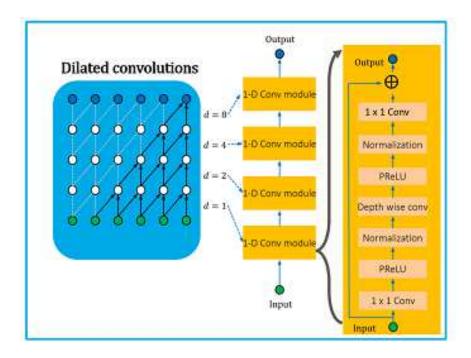
Figure 1: The structure of TCN.

Dilated convolutions with different dilations 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 the input sequences. The TCN introduces a time hierarchy: the upper layer can access longer input subsequences and learn representations on larger time scales. Local information from lower layers spreads through the hierarchy by means of residuals and skip connections.

There are two important elements in the original TCN [2] as shown in Figure 1, one is the dilated convolutions, and the other is residual connections. Dilated convolutions follow the work of [26], 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(aka 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 $1, 2, 4, 4$ respectively. As shown in Figure 1, Each 1-D Conv module is a residual block [7], which contains one layer of dilated convolution (Depth wise conv [9]), two layers of 1×1 convolutions (1×1 Conv), two non-linearity activation layers (Parametric Rectified Linear Unit, PReLU [6]), and two normalization layers (Normalization). For normalization, we applied global normalization [18] to the convolutional filters.

Luo et al. proposed a TCN based speech separation method [18], 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.
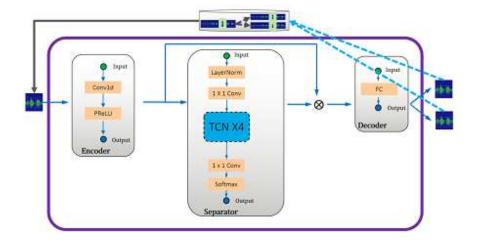
3

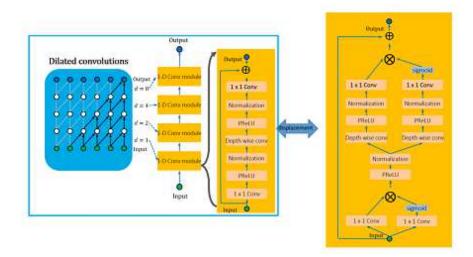Figure 2: The pipeline of TCN based speech separation in [18].

Figure 3: The structure of gated TCN.

3 Speech separation with FurcaNeXt

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 gating operations in this TCN, as shown in Figure 3. Nonlinear gated activation had been used in prior work on sequence modeling [26, 4], 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 is corresponding to the first 1×1 convolutional layer in the 1-D convolutional module, the other is corresponding to all the layers from the depth-wise convolutional layer to the output 1×1 convolutional layer. 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
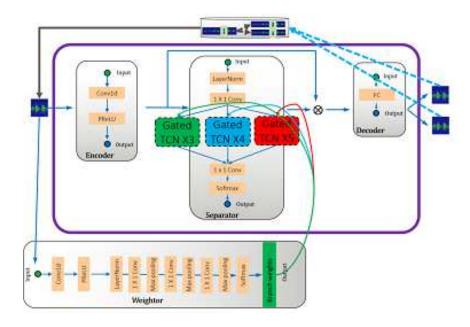
4

Figure 4: The structure of FurcaPy.

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 FurcaSh: Weight-shared multi-scale gated TCN

FurcaPy will increase the number of parameters of the network several times, and the processing speed of the network will decrease a lot. In many cases, there is no way to meet the requirements of real-time processing for such network. In order to deal with these problems, a new structure is proposed that can achieve the two-level multi-scale receptive fields without increasing the number of network parameters. As shown in Figure 5 and Figure 6, two levels of multi-scale temporal receptive fields is introduced, one is in the dilated 1-D conv module level, that is the outputs corresponding to different dilated factors are summed and averaged to result in the final output of this gated TCN; the other is that since there are 4 gated TCNs in the FurcaSh pipeline, the outputs of different gated TCNs are summed and averaged to result in the separator. So there are two different levels of multi-scale temporal receipt field in this structure.

3.3 FurcaPa: Deep gated dilated temporal convolutional networks (TCN) with intra-parallel convolutional components

The performance of a single predictive model can always be improved by ensemble, that is to combines a set of independently trained networks. The most commonly used method is to do the average of the model, which can at least help to reduce the variance of the performance. As shown in Figure 7, in the different layers of each 1-D convolutional module of gated TCN in FurcaPorta, two identical **pa**rallel branches are
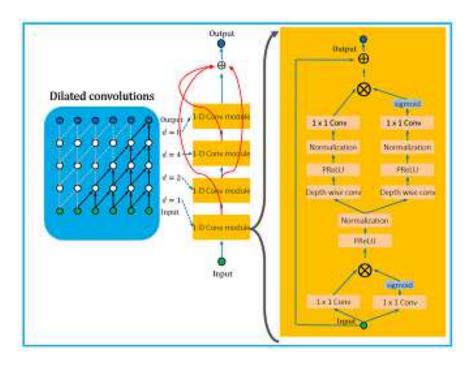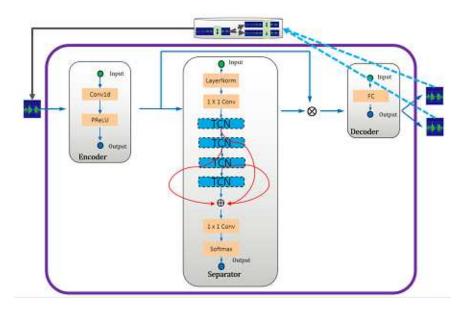
Figure 5: The structure of FurcaSh.

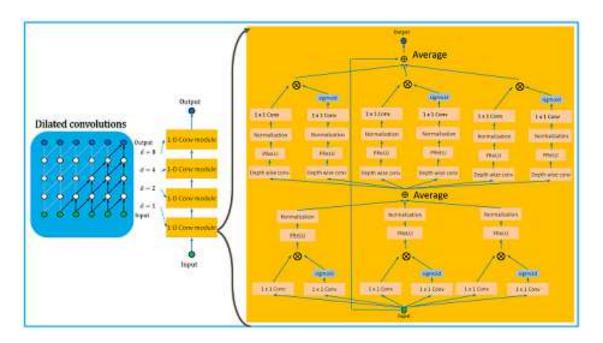

Figure 6: The structure of FurcaSh.

Figure 7: The structure of FurcaPa.

added. This structure is called FurcaPa. The total output of each intra-parallel convolutional components is obtained by averaging the outputs of all the different branches. In each single dilated 1-D convolutional module layer, two intra-parallel convolutional components are introduced, the first one is near the input layer and covers the Conv1d, PReLU, and Normalization layers; the other one is near the output and it covers the rest all layers, including the Depthwise conv, PReLu, Normalization and 1×1 Conv layers. The reason why we do this ensemble in two places is to reduce the sub-variances of each block.

3.4 FurcaSu: Dilated TCN with gated difference-convolutional component

Highway network can be simplified and generalized to have better performance [34]. Follow the work of [34], we further simplify the Highway network module, as shown in Figure 8, in each single dilated 1-D convolutional module layer, two Highway network module or gated difference-convolutional components as we called are introduced, the first one is near the input layer and covers the Conv1d, PReLU, and Normalization layers; the other one is near the output and it covers the rest all layers, including the Depthwise conv, PReLu, Normalization and 1×1 Conv layers. Different from the original use of three different transformation functions, in order to simplify the design and improve performance, here we use three identical transformation branches, one branch as the attention gates, the other two are signal transformations, and their results are subtracted and then gated.

3.5 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 [33, 1]. 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) [5, 28], and short-time objective intelligibility (STOI) [25] 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 FurcaNet 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.
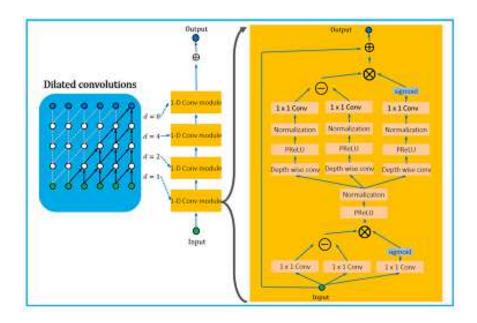
7

Figure 8: The structure of FurcaSu.

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 [5, 28]

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

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 [12, 35] is employed in this work. We calculate the SDRs for all the permutations, pick the maximum one, and take the negative as the loss. It is called the uSDR loss in this work.

4 Experiments

4.1 Dataset and neural network

We evaluated our system on two-speaker speech separation problem using WSJ0-2mix dataset [8, 11], 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.

We evaluate the systems with the SDR improvement (SDRi) [5, 28] metrics used in [11, 16, 32, 3, 12]. 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. Table 1 lists the average SDRi obtained by the different structures in FurcaNeXt and almost all the results in the past two years, where IRM means the ideal ratio mask

$$M_s = \frac{|X_s(t, f)|}{\sum_{s=1}^{S} |X_s(t, f)|} \tag{4.1}$$

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 $x_s(t)$.

8

In this experiment, as baselines, we reimplemented several classical approaches, such as DPCL [8], Tas-Net [17] and Conv-TasNet [18]. Table 1 lists the SDRis 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 nearly 2.6dB SDRi is obtained. 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 [8]	5.9
uPIT-BLSTM [35]	10.0
cuPIT-Grid-RD [32]	10.2
DANet [3]	10.5
ADANet [16]	10.5
DPCL*	10.7
DPCL++ [11]	10.8
CBLDNN-GAT [15]	11.0
TasNet [17]	11.2
TasNet*	11.8
Chimera++ [31]	12.0
FurcaX [21]	12.5
IRM	12.7
FurcaNet [22]	13.3
Conv-TasNet [18]	15.0
Conv-TasNet*	15.8
FurcaPorto	17.3
FurcaSu	17.9
FurcaSh	18.0
FurcaPa	18.2
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. We propose a series structure under the name of FurcaNeXt do to speech separation. Benefits from the strength of end-to-end processing, the novel gating mancinism and dynamic improvements, the best performance of structure in FurcaNeXt achieve 18.4dB SDRi on the the public WSJ0-2mix data corpus, results in 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 [19]. 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

References

[1] Assmann, P., Summerfield, Q.: The perception of speech under adverse conditions. In: Speech processing in the auditory system, pp. 231–308. Springer (2004)

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

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

[4] Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. arXiv preprint arXiv:1612.08083 (2016)

[5] Févotte, C., Gribonval, R., Vincent, E.: Bss_eval toolbox user guide–revision 2.0 (2005)

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

[7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

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

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

[10] Hu, K., Wang, D.: An unsupervised approach to cochannel speech separation. IEEE Transactions on audio, speech, and language processing 21(1), 122–131 (2013)

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

[12] Kolbæk, M., Yu, D., Tan, Z.H., Jensen, J., Kolbaek, M., Yu, D., Tan, Z.H., Jensen, J.: Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 25(10), 1901–1913 (2017)

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

[14] Lea, C., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks: A unified approach to action segmentation. In: European Conference on Computer Vision. pp. 47–54. Springer (2016)

[15] Li, C., Zhu, L., Xu, S., Gao, P., Xu, B.: Cbldnn-based speaker-independent speech separation via generative adversarial training (2018)

[16] Luo, Y., Chen, Z., Mesgarani, N.: Speaker-independent speech separation with deep attractor network. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(4), 787–796 (2018)

[17] Luo, Y., Mesgarani, N.: Tasnet: time-domain audio separation network for real-time, single-channel speech separation. arXiv preprint arXiv:1711.00541 (2017)

[18] Luo, Y., Mesgarani, N.: Tasnet: Surpassing ideal time-frequency masking for speech separation. arXiv preprint arXiv:1809.07454 (2018)

[19] Roux, J.L., Wisdom, S., Erdogan, H., Hershey, J.R.: Sdr-half-baked or well done? arXiv preprint arXiv:1811.02508 (2018)

[20] Shao, Y., Wang, D.: Model-based sequential organization in cochannel speech. IEEE Transactions on Audio, Speech, and Language Processing 14(1), 289–298 (2006)

[21] Shi, Z., Lin, H., Liu, L., Liu, R., Hayakawa, S., Han, J.: Furcax: End-to-end monaural speech separation based on deep gated (de)convolutional neural networks with adversarial example training. In: Proc. ICASSP (2019)

[22] Shi, Z., Lin, H., Liu, L., Liu, R., Hayakawa, S., Harada, S., Han, J.: 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)

[23] Smaragdis, P., et al.: Convolutive speech bases and their application to supervised speech separation. IEEE Transactions on audio speech and language processing 15(1), 1 (2007)

[24] Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. arXiv preprint arXiv:1505.00387 (2015)

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

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

[27] Venkataramani, S., Casebeer, J., Smaragdis, P.: Adaptive front-ends for end-to-end source separation. In: Proc. NIPS (2017)

[28] 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)

[29] Virtanen, T.: Speech recognition using factorial hidden markov models for separation in the feature space. In: Ninth International Conference on Spoken Language Processing (2006)

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

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

[32] Xu, C., Xiao, X., Li, H., XU, C., RAO, W., XIAO, X., CHNG, E.S., LI, H.: Single channel speech separation with constrained utterance level permutation invariant training using grid lstm (2018)

[33] Yang, W., Benbouchta, M., Yantorno, R.: 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, pp. 541–544. IEEE (1998)

[34] Yousef, M., Hussain, K.F., Mohammed, U.S.: Accurate, data-efficient, unconstrained text recognition with convolutional neural networks. arXiv preprint arXiv:1812.11894 (2018)

[35] Yu, D., Kolbæk, M., Tan, Z.H., Jensen, J.: 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. pp. 241–245. IEEE (2017)