FURCAX: END-TO-END MONAURAL SPEECH SEPARATION BASED ON DEEP GATED (DE)CONVOLUTIONAL NEURAL NETWORKS WITH ADVERSARIAL EXAMPLE TRAINING

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

Fujitsu Research and Development Center, Beijing, China[1]
Fujitsu Laboratories Ltd. Kawasaki, Japan[2]
Harbin Institute of Technology, Harbin, China[3]

ABSTRACT

Deep gated convolutional networks have been proved to be very effective in single channel speech separation. However current state-of-the-art framework often considers training the gated convolutional networks in time-frequency (TF) domain. Such an approach will result in limited perceptual score, such as signal-to-distortion ratio (SDR) upper bound of separated utterances and also fail to exploit an end-to-end framework. In this paper we present an integrated simple and effective end-to-end approach called FurcaX[1] to monaural speech separation, which consists of deep gated (de)convolutional neural networks (GCNN) that takes 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 SDR in a permutation invariant training (PIT) style. We execute generative adversarial training (GAT) throughout the training, which makes the separated speech indistinguishable from the real one. Our experiments on the the public WSJ0-2mix data corpus demonstrate that this new scheme can produce more discriminative separated utterances and leading to performance improvement on the speaker separation task.

Index Terms— Speech separation, cocktail party problem, gated convolutional neural network, generative adversarial training, permutation invariant training

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) [1, 2, 3], Non-negative matrix factorization (NMF) [4, 5], and probabilistic models [6], 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) [7, 8], 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) [9, 10], 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 [11, 12], 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 an extension of the end-to-end speech enhancement method [13] to speech separation. 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 method, and finally the cleaned source speech signal can be restored by an

[1]"Furca" is Latin for "fork", and we use this word to mean the speech is split into two streams by our network like water. "X" is the shape of the separator in our framework.

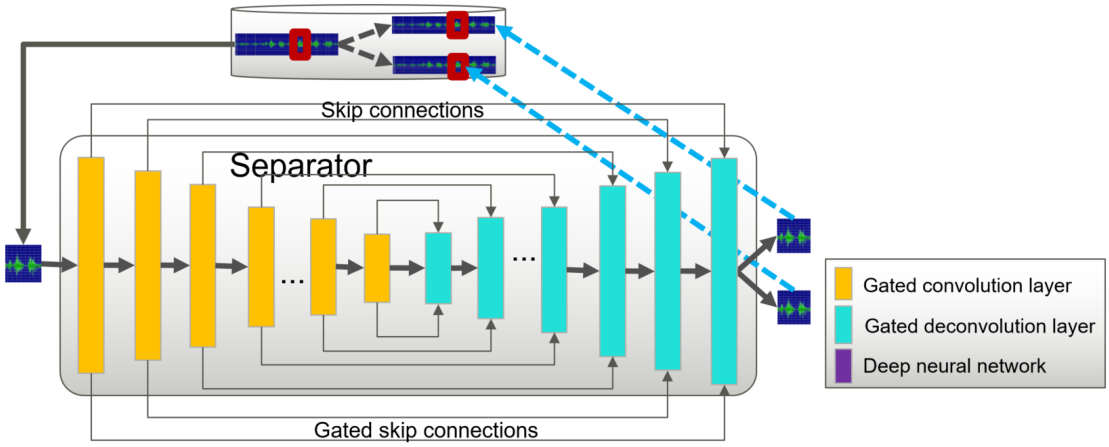

Fig. 1. Architecture of the separator in FurcaX.

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 [13, 11, 12]. But both of the works [11, 12] are based on very short frames, for example 5ms in [11], which results in severe label permutation problem. While Pascual et al. proposed a successfull model call SEGAN to do speech enhancement in a long scale of 1s utterance. Inspired by these first results, we propose FurcaX, a fully end-to-end time-domain separation systems, based on deep gated (de)convolutional network (GCNN) [14] and generative adversarial training [15].

The remainder of this paper is organized as follows: section 2 first introduces monaural speech separation, then describe our proposed FurcaX and the separation algorithm in detail. The experimental setup and results are presented in Section 3. We conclude this paper in Section 4.

2. THE FURCAX MODEL

The proposed end-to-end deep learning approach consists of two main components: one is the FurcaX pipeline, which consists of a deep GCNN separator and a deep GCNN discriminator; and the other is the perceptual loss function.

In this section, we first review the formal definition of the monaural speech separation task and the deep GCNN architecture. We then show the details of the FurcaX architecture we investigated. Finally the perceptual metric as a loss function is introduced.

Fig. 2. Architecture of the discriminator in FurcaX.

2.1. Monaural speech separation

The goal of monaural speech separation is to estimate the individual target signals in 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 in given mixed speech $y(t)$. In this work it is assumed that the number of target signals is known.

In this work, we propose an end-to-end deep learning approach to separate the mixed voice. The input of the FurcaX is a mixed utterance $y(t)$, and the output of the network are the separated utterances, ideally it is best to be exactly the same as $x_i(t), i = 1, .., S$. In order to do

this, the mixed speech is firstly framed. Then each frame of the mixed utterance $y(t)$ is directly as raw wave forward propagated through the FurcaX, and the output activations are the separated frames, each frame is corresponding only one speaker. Finally the separated frames are concatenate together to form the output utterances.

2.2. Network architecture

The proposed FurcaX model is similar to [13], but with fine adjustment. The FurcaX separation system comprises a separator and a discriminator, and the structure is illustrated in Fig. 1 and Fig. 2. A deep GCNN proposed in [14] is adopted here to build the main frame. GCNN is implemented by stacking multiple 1D gated (de)convolutional (GConv or GDeconv) layers on top of each other.

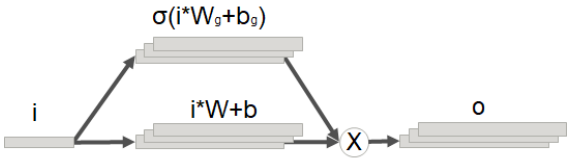

Fig. 3. Architecture of a 1D GConv or GDeconv layer.

Fig. 3 shows the structure of a 1D GConv and GDeconv layer. The main difference between a GConv(GDeconv) layer and a plain convolutional layer is that a gated linear unit (GLU) [14], namely the the gates $\sigma(i * W_g + b_g)$ of Eq. (1) is used as a nonlinear control function instead of tanh activation or regular rectified linear units (ReLUs) [14]:

$$o = (i * W + b) \otimes \sigma(i * W_g + b_g), \quad (1)$$

where i and o are the input and output, W, b, W_g, and b_g are learned parameters, σ is the sigmoid function and $\otimes$ is the element-wise product between vectors or matrices. Similar to LSTMs, GLUs play the role of controlling the information passed on in the hierarchy. This special gating mechanism allows us to effectively capture long-range context dependencies by deepening layers without encountering the problem of vanishing gradient.

The separator consists of 11 stacked GConv layers and 10 stacked GDeconv layers. The input to the separator is a speech frame of size 16384 (about 2s speech in our setting). The dimensions of the outputs from the successive layers of the separator are: 1×16384 (input), 16×8192, 32×4096, 32×2048, 64×1024, 64×512, 128×256, 128×128, 256×64, 256×32, 512×16, 1024×8, 512×16, 256×32, 256×64, 128×128, 128×256, 64×512, 64×1024, 32×2048, 32×4096, 16×8192, 2×16384 (output). Since different from plain convolution layer, GConv layer has two data flows, so unlike the skip connections in [13], our separator has two kinds of skip connections, one is the ordinary skip connections same as in [13], the other is the gated skip connections, as it is shown in Fig. 1.

In order to overcome the limitation of ordinary loss function's strong assumption on how the distribution of the separator is shaped, we use the generative adversarial training. The discriminator can be understood as learning some sort of loss for separators output to look real. The discriminator consists of 11 stacked GConv layers and 1 full connected layer. The input to the discriminator is two speech frames of size 16384 each. The dimensions of the outputs from the successive layers of the separator are: 2×16384 (input), 32×8192, 64×4096, 64×2048, 128×1024, 128×512, 256×256, 256×128, 512×64, 512×32, 1024×16, 2048×8, 1×8, 1(output). The generator and discriminator are trained jointly. During training we need to provide the correct reference $x_i(t), i = 1, .., S$ to the corresponding output layer for supervision.

2.3. Loss function

2.3.1. 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. We directly use the BSS_Eval metric signal-to-distortion ratio (SDR) [16, 17], which is most commonly used metrics to evaluate the performance of source separation, as the training objective.

SDR captures the overall separation quality of the algorithm. There is a subtle problem here. We first concatenate the outputs of FurcaX 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 [16, 17]

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

Then our target is to maximize SDR or minimize the negative SDR as loss function respect to the s. The PIT training criteria [9, 10] 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.

2.3.2. GAT loss

In this work we train the separator S by GAT. Here we use two-speaker separation as an example. Thus the loss

function is modified. We use LSGAN [18] based method. The discriminator D is trained to recognize the target utterance pair x_1, x_2 as real and the separated pair s_1, s_2 as fake. The separator S is trained to fool the discriminator D that let D believes the utterances separated from the mixed voice m to be real. At the same time, uSDR loss is as a regularization to guide the training. In order to balance GAN loss and uSDR loss, λ is taken as hyperparameter in this experiment:

$$\begin{aligned} \min_{\mathrm{D}} \mathcal{L}(\mathrm{D}) &= \mathrm{E}[(\mathrm{D}(x_1, x_2) - 1)^2] + \mathrm{E}[(\mathrm{D}(s_1, s_2))^2], \\ \min_{\mathrm{S}} \mathcal{L}(\mathrm{S}) &= \mathrm{E}[(\mathrm{D}(\mathrm{S}(m)) - 1)^2] + \lambda \mathrm{uSDR}(s_1, s_2; x_1, x_2). \end{aligned}$$

where E is the expectation operator, $\mathcal{L}(\mathrm{D})$ and $\mathcal{L}(\mathrm{S})$ are the losses of D and S respectively.

3. EXPERIMENTS

3.1. 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. To reduce the computational cost, the waveforms were down-sampled to 8 kHz. In this work, we shift the window around raw waveform by about 200ms and produce a set of frames at about 2s (16384 samples) intervals.

3.2. Results

We evaluate the systems with the SDR improvement (SDRi) [16, 17] metrics used in [8, 19, 20, 21, 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. Table 1 lists the average SDRi obtained by FurcaX and almost all the results in the past two years, where IRM means the ideal ratio mask

$$M_s = \frac{|X_s(t,f)|}{\sum_{s=1}^{S} |X_s(t,f)|} \tag{2}$$

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 two classical approaches DPCL [7] and TasNet [11]. Compared with these baselines an average increase of 0.5dB SDRi is obtained. FurcaX has achieved the most significant performance improvement compared with baseline systems, and it almost (0.2dB less) break through the upper bound of STFT based methods.

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 [20]	10.2
DANet [21]	10.5
ADANet [19]	10.5
DPCL*	10.7
DPCL++ [8]	10.8
CBLDNN-GAT [22]	11.0
TasNet [11]	11.2
TasNet*	11.8
Chimera++ [23]	12.0
FurcaX	12.5
IRM	12.7

4. CONCLUSION

In this study, we proposed an end-to-end architecture called FurcaX for monaural speech separation. FurcaX can combine the advantages of gated convolution operation and generative adversarial training, and at the same time it can directly use perceptual metrics such as SDR as regularized optimization objective. Our results on two-speaker mixed speech separation task indicate that FurcaX can achieve a state-of-the-art performance. Future research would include extending the experiment to three-speaker mix task to see whether it is independent of the number of sound sources.

5. 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. We also would like to thank Anyan Shi at infansPISTRIS Technology for the motivation and encouragement of this work.

6. REFERENCES

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

[2] Yang Shao and DeLiang 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] Ke Hu and DeLiang 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] Paris 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, pp. 1, 2007.

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

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

[7] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji 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] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey, "Single-channel multi-speaker separation using deep clustering," *arXiv preprint arXiv:1607.02173*, 2016.

[9] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, Jesper Jensen, Morten Kolbaek, Dong Yu, Zheng-Hua Tan, and Jesper 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] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper 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] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," *arXiv preprint arXiv:1711.00541*, 2017.

[12] Shrikant Venkataramani, Jonah Casebeer, and Paris Smaragdis, "Adaptive front-ends for end-to-end source separation," in *Proc. NIPS*, 2017.

[13] Santiago Pascual, Antonio Bonafonte, and Joan Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[14] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier, "Language modeling with gated convolutional networks," *arXiv preprint arXiv:1612.08083*, 2016.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[16] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.

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

[18] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2813–2821.

[19] Yi Luo, Zhuo Chen, and Nima 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.

[20] Chenglin Xu, Xiong Xiao, Haizhou Li, CHENGLIN XU, WEI RAO, XIONG XIAO, ENG SIONG CHNG, and HAIZHOU LI, "Single channel speech separation with constrained utterance level permutation invariant training using grid lstm," 2018.

[21] Zhuo Chen, Yi Luo, and Nima 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.

[22] Chenxing Li, Lei Zhu, Shuang Xu, Peng Gao, and Bo Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," 2018.

[23] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.