Noise-Tolerant Time-Domain Speech Separation with Noise Bases

Kohei Ozamoto*, Kuniaki Uto*, Koji Iwano† and Koichi Shinoda*
* Tokyo Institute of Technology, Tokyo, Japan
E-mail: {ozamoto, uto}@ks.c.titech.ac.jp, shinoda@c.titech.ac.jp Tel/Fax: +81-3-5734-3481
† Tokyo City University, Yokohama, Japan
E-mail: iwano@tcu.ac.jp Tel/Fax: +81-45-910-2598/+81-45-910-2599

***Abstract*—Time-domain Audio Separation Network (TasNet) is a deep-learning-based speech separation system that performs end-to-end learning in time domain. By directly using speech signal in a time domain, TasNet outperforms conventional systems based on a time-frequency domain in terms of reconstruction quality and latency. In TasNet, the separation accuracy is greatly degraded when speech contains noise. We propose TasNet with noise basis signals (TasNet-NB), a method to improve separation performance under noisy environments by adding noise basis signals to speaker's basis signals. It utilizes curriculum learning to gradually reduce the noise signal-to-noise ratio and the noise reconstruction loss as one of the objective functions in training. We evaluate the method on WHAM! dataset and show that it improves SI-SDRi from 13.7 dB to 14.6 dB.**

I. Introduction

Speech separation aims for separating each speaker's speech from each other when multiple people are speaking at the same time. The separation is a necessary pre-processing for higher-level speech recognition, such as automatic speech recognition (ASR), speaker identification and emotion recognition. In conventional speech separation systems, the separation is generally done in a time-frequency (T-F) domain into which the original mixed signal is decomposed by discrete short-time Fourier transform (DSTFT) using time-frequency masks [1]. Then, each individual speech source is retrieved by inverse discrete short-time Fourier transform (IDSTFT) from the estimated spectrum of the original speech [2]. Those systems utilize the frequency and amplitude of speech while its phase is discarded. As a result, waveforms reconstructed without the phase information inevitably result in a decrease in quality. In these T-F domain systems, speech signals are decomposed into a set of segments with a short time window length, and hence, a relatively large time window length is needed for high frequency resolution, which leads to high latency.

Recently proposed Time-domain Audio Separation Network (TasNet) [3] and its derivatives [4], [5], [6] perform end-to-end learning by directly using speech signal in time domain. It avoids both of the above-mentioned shortcomings of the T-F domain systems, i.e., the degradation of reconstruction quality and high latency. In TasNet, a waveform is converted into a set of features by convolutions. The filters used for the convolutions are called the basis signals, and they are trained to improve the separation performance between speakers. The network parameters including the basis signals and speaker separation masks are trained by using the reconstruction loss of speakers' speeches. While TasNet generally has higher separation performance than T-F domain systems, the performance degrades significantly when the speech to be separated contains environmental noise [7].

Different from speech enhancement in which accurate reconstruction of noise is not required in general, speech separation inherently requires reconstruction of noise as well as speakers' speeches to differentiate target signals (the speakers' speech) from noise [7]. TasNet-based approach in noisy environment can be benefited from explicitly adding noise basis signals and, then, evaluating the reconstruction loss of noise in addition to the loss for spearkers' speeches.

In this paper, we propose TasNet with noise basis signals (TasNet-NB), a method for creating noise basis signals together with speaker basis signals for time-domain systems. The noise reconstruction loss is used as one of the objective functions in training. To further improve the separation accuracy, we also employ curriculum learning, in which the relative amount of noise in the dataset is gradually increased.

II. Related works

A. TasNet

1) Overview: TasNet has a structure with an encoder and a decoder between which a separator is inserted (Fig. 1). The encoder directly converts a mixed input signal segment into a feature vector. The separator decomposes it into a set of feature vectors, each of which represents individual speaker characteristics. Finally, from each speaker's feature vector, the decoder retrieves his/her speech.

Given a time window length L, a set of overlapping segments with a stride of $L/2$, $\{\mathbf{x}_k\}_{k=1}^{K}$, is extracted from incoming mixed speech signals, where $\mathbf{x}_k \in \mathbb{R}^{1\times L}$ and K are the k-th segment and the number of segments, respectively. Note that L is significantly smaller than that used in time-frequency domain systems. The objective is to separate the discrete speech segment $\mathbf{x}_k$ into a set of C speakers' speech $\{\mathbf{s}_{i,k}\}_{i=1}^{C}$, where $\mathbf{s}_{i,k}$ is the i-th speaker's speech element in $\mathbf{x}_k$.

This work was supported by JST CREST JPMJCR1687 and by MEXT KAKENHI 19H04133.

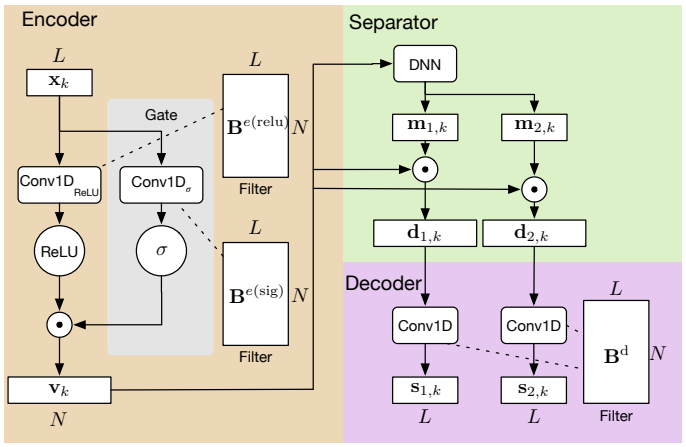

Fig. 1. Network of TasNet (2 speakers).

The encoder maps the segment $\mathbf{x}_k$ to a N-dimensional feature vector $\mathbf{v}_k \in \mathbb{R}^{1\times N}$ using learnable convolution parameters. Hereafter, we call the feature vector $\mathbf{v}_k$ and the convolution parameter as a weight vector and a basis signal, respectively. The number of the basis signals equals to the dimension of the $\mathbf{v}_k$, N. Here, each element of the weight vector should be non-negative, and hence, each mixed segment $\mathbf{x}_k$ is represented by a non-negative weighted sum of the basis signals. In the separator, the weight vector $\mathbf{v}_k$ is decomposed into a set of weight vectors for individual speakers' speech $\{\mathbf{d}_{i,k}\}_{i=1}^{C}$ using a set of estimated masks $\{\mathbf{m}_{i,k}\}_{i=1}^{C}$, where $\mathbf{d}_{i,k} \in \mathbb{R}^{1\times N}$ corresponds to the segment $k \in \{1,\ldots,K\}$ of a speaker $i \in \{1,\ldots,C\}$. The j-th element in the mask $\mathbf{m}_{i,k} \in \mathbb{R}^{1\times N}$, $\mathbf{m}_{i,k}(j)$, represents the rate of speech of speaker i in the j-th element in $\mathbf{v}_k$, $\mathbf{v}_k(j)$. Finally, in the decoder, a set of individual speakers' speech $\{\mathbf{s}_{i,k}\}_{i=1}^{C}$ is reconstructed based on $\{\mathbf{d}_{i,k}\}_{i=1}^{C}$ and learnable basis signals.

2) Encoder: The encoder converts the k-th segment $\mathbf{x}_k$ to a weight vector $\mathbf{v}_k$ as:

$$\mathbf{v}_k = \mathrm{ReLU}(\mathrm{Conv1D}_{\mathrm{ReLU}}(\mathbf{x}_k)) \odot \sigma(\mathrm{Conv1D}_{\sigma}(\mathbf{x}_k)), \quad (1)$$

where $\odot$ denotes the Hadamard product. $\mathrm{Conv1D}_{\mathrm{ReLU}}$ and $\mathrm{Conv1D}_{\sigma}$ are both concatenations of N 1D convolutional layers whose outputs are transferred to Rectified Linear Unit (ReLU) and sigmoid function (σ), respectively. The weight vectors of the 1D convolution layers are stacked to form the encoder basis signals $\mathbf{B}^{e(\mathrm{relu})}, \mathbf{B}^{e(\mathrm{sig})} \in \mathbb{R}^{N\times L}$, respectively. Based on the $\mathbf{B}^{e(\mathrm{relu})}$ and $\mathbf{B}^{e(\mathrm{sig})}$, both of the convolution are defined by $\mathrm{Conv1D}_{\mathrm{ReLU}}(\mathbf{x}_k) = \mathbf{x}_k \circledast \mathbf{B}^{e(\mathrm{relu})}$ and $\mathrm{Conv1D}_{\sigma}(\mathbf{x}_k) = \mathbf{x}_k \circledast \mathbf{B}^{e(\mathrm{sig})}$, respectively, where $\circledast$ is a convolution operator. This gated CNN approach empirically demonstrated better performance in language modeling than using only ReLU or sigmoid [8].

Given the weight vector $\mathbf{v}_k$, the separator outputs masks $\mathbf{M}_k = [\mathbf{m}_{1,k}, ..., \mathbf{m}_{C,k}]$ by using the LSTM, the subsequent fully connected layer, and the sigmoid function. The estimated mask $\{\mathbf{m}_{i,k}\}_{i=1}^{C}$ are used to perform the separation for $\mathbf{v}_k$. Then, $\mathbf{v}_k$ is normalized by layer normalization [9] to speed up and stabilized the training process [3].

$$\mathbf{d}_{i,k} = \mathbf{v}_k \odot \mathbf{m}_{i,k} \quad (2)$$

3) Decoder: The decoder converts the feature $\mathbf{d}_{i,k} \in \mathbb{R}^{1\times N}$ obtained from the separator into the waveform by

$$\mathbf{s}_{i,k} = \mathbf{d}_{i,k}\mathbf{B}^{d}, \quad (3)$$

where $\mathbf{B}^{d} \in \mathbb{R}^{N\times L}$ is a filter of the convolutional layer named the decoder basis signals. In this way, the waveform of the i-th speaker's k-th segment is obtained as $\mathbf{s}_{i,k} \in \mathbb{R}^{1\times L}$.

4) Objective function: The objective function of TasNet training is scale-invariant signal-to-distortion ratio (SI-SDR) [10] that has been widely used in single-channel speech separation. SI-SDR is defined in (6),

$$\mathbf{s}'_{i,k} = \frac{\langle \hat{\mathbf{s}}_{i,k}, \mathbf{s}_{i,k}\rangle \mathbf{s}_{i,k}}{\| \mathbf{s}_{i,k} \|^2}, \quad (4)$$

$$\mathbf{e}'_{i,k} = \hat{\mathbf{s}}_{i,k} - \mathbf{s}'_{i,k}, \quad (5)$$

$$\text{SI-SDR}(\mathbf{s}_{i,k}, \hat{\mathbf{s}}_{i,k}) := 10\log_{10}\frac{\| \mathbf{s}'_{i,k} \|^2}{\| \mathbf{e}'_{i,k} \|^2}, \quad (6)$$

where $\hat{\mathbf{s}}_{i,k}$ and $\mathbf{s}_{i,k}$ are the estimated and target signals of the i-th speaker's k-th segment, respectively. $\langle \alpha, \beta \rangle$ represents the inner product of α, β, and $\|\alpha\|$ represents the L2 norm. In the training process, permutation invariant training (PIT) [11] is used to deal with the uncertainty of the source order.

TasNet outperforms the conventional T-F systems in terms of reconstruction quality and latency. However, one problem of TasNet is that the separation accuracy is greatly degraded when the signal contains environmental noise.

B. Replacing the basis signals

In time-domain systems such as TasNet, the basis signals of the encoder and decoder are optimized in the end-to-end learning process. On the other hand, [6] demonstrated that replacing the trained basis signals with the multiphase gammatone filter bank not only improves the SI-SDR but also

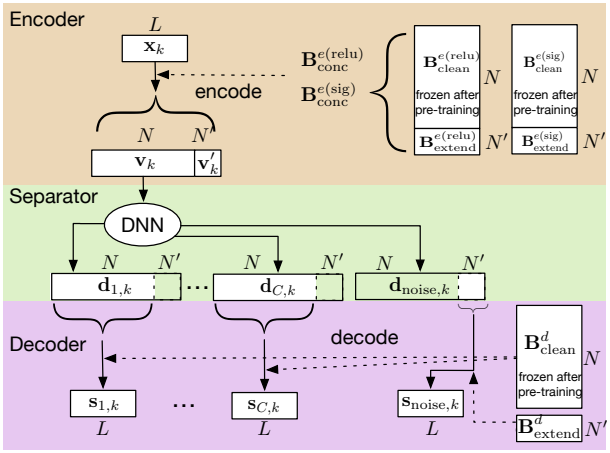

Fig. 2. Network of TasNet-NB.

keeps the performance the same even if the size of the basis signals is reduced.

C. Noise reconstruction loss

Methods of separating noise as well as speakers based on masks estimated by the reconstruction loss has been used in speech enhancement [12], [13]. It has also been described that noise estimation serves as a regularization for learning robust models [13]. Noise separation based on masked estimated by reconstruction loss has not been investigated in time-domain speech separation systems.

D. Curriculum learning

Curriculum learning [14] imitates the learning strategy of humans to gradually understand difficult concepts starting from easy ones. The goal is to improve the learning efficiency and performance through a curriculum in which students are gradually trained from easy to difficult tasks. For example, in speech and speaker recognition, multi-stage training in which the models are trained under step-by-step environments gradually increasing signal-to-noise ratio (SNR) are proved to be effective [15], [16], [17].

III. TasNet with noise basis signals

Inspired by the ideas of the post-training manipulation of basis signals [6] and the noise estimation based on noise reconstruction loss [12], [13] mentioned in Section II, we propose TasNet with noise basis signals (TasNet-NB, Fig. 2). It improves the separation performance between speakers in a noisy environment by introducing the noise basis signals in addition to the speaker basis signals. We further apply curriculum learning to improve the separation performance.

A. Noise basis signals

In the end-to-end learning process of TasNet, the basis signals of the encoder and decoder are updated in the same way as the other parameters. First, we pre-train the basis signals

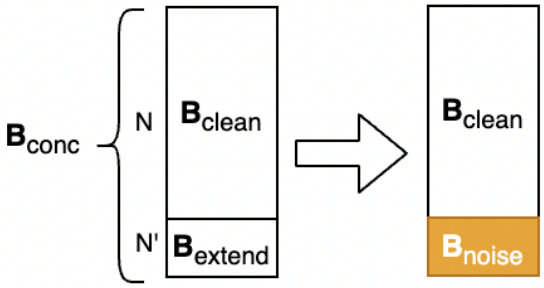

Fig. 3. Learning of noise basis signals.

$\mathbf{B}^{\gamma}$, $(\gamma = e(\mathrm{relu}), e(\mathrm{sig}), d)$ using the speaker-only dataset ("Clean"). At this point, $\mathbf{B}^{\gamma}$ is considered to be specialized for inter-speaker separation in the "Clean" environment.

$$\mathbf{B}^{\gamma} = \mathbf{B}^{\gamma}_{\mathrm{clean}} \tag{7}$$

Next, we introduce additional basis signals $\mathbf{B}^{\gamma}_{\mathrm{extend}} \in \mathbb{R}^{N' \times L}$ designed for decomposing noise in a noisy environment. N' is the given number of noise basis signals. The noise basis signals $\mathbf{B}^{\gamma}_{\mathrm{extend}}$ are combined with the original basis signals as follows,

$$\mathbf{B}^{\gamma}_{\mathrm{conc}} = \mathrm{Conc}(\mathbf{B}^{\gamma}_{\mathrm{clean}}, \mathbf{B}^{\gamma}_{\mathrm{extend}}) \in \mathbb{R}^{(N+N') \times L}, \tag{8}$$

where Conc denotes the union of vectors. During the training of $\mathbf{B}^{\gamma}_{\mathrm{conc}}$ for noisy speech separation ("Noisy"), only $\mathbf{B}^{\gamma}_{\mathrm{extend}}$ is updated while the "Clean" part $\mathbf{B}^{\gamma}_{\mathrm{clean}}$, is frozen (Fig. 3). By training in this way, $\mathbf{B}^{\gamma}_{\mathrm{conc}}$ is expected to acquire noise separation function while maintaining separation ability between speakers by preserving $\mathbf{B}^{\gamma}_{\mathrm{clean}}$, i.e.,

$$\mathbf{B}^{\gamma}_{\mathrm{extend}} = \mathbf{B}^{\gamma}_{\mathrm{noise}}, \tag{9}$$

where $\mathbf{B}^{\gamma}_{\mathrm{noise}}$ is the noise basis signals specialized for noise separation.

Given a segment $\mathbf{x}_k$, the encoder outputs an $(N + N')$-dimensional weight vector $\mathrm{Conc}(\mathbf{v}_k, \mathbf{v}'_k) \in \mathbb{R}^{1 \times (N+N')}$, where $\mathbf{v}_k \in \mathbb{R}^{1 \times N}$ and $\mathbf{v}'_k \in \mathbb{R}^{1 \times N'}$ are weight vectors for speakers and noise, respectively.

B. Noise reconstruction loss

Based on the concatenated weight vector $\mathrm{Conc}(\mathbf{v}_k, \mathbf{v}'_k)$, the separator computes masks for decomposing the concatenated weight vector into C speakers and noise, $\mathrm{Conc}(\mathbf{M}_k, \mathbf{m}_{\mathrm{noise},k}) = [\mathbf{m}_{1,k}, \ldots, \mathbf{m}_{C,k}, \mathbf{m}_{\mathrm{noise},k}]$, where $\mathbf{m}_{i,k}$ and $\mathbf{m}_{\mathrm{noise},k}$ are masks for i-th speaker and noise separations, respectively. In the decoder, we restrict $\mathbf{B}^{d}_{\mathrm{extend}} \in \mathbb{R}^{N' \times L}$ to be used only for noise reconstruction and $\mathbf{B}^{d}_{\mathrm{clean}} \in \mathbb{R}^{N \times L}$ for speaker reconstruction,

$$\begin{aligned} \mathbf{s}_{i,k} &= (\mathbf{d}_{i,k}\mathbf{P}_{\mathrm{speech}})\mathbf{B}^{d}_{\mathrm{clean}} \\ &= (\mathrm{Conc}(\mathbf{v}_k, \mathbf{v}'_k) \odot \mathbf{m}_{i,k})\,\mathbf{P}_{\mathrm{speech}}\mathbf{B}^{d}_{\mathrm{clean}}, \end{aligned} \tag{10}$$

$$\begin{aligned} \mathbf{s}_{\mathrm{noise},k} &= (\mathbf{d}_{\mathrm{noise},k}\mathbf{P}_{\mathrm{noise}})\mathbf{B}^{d}_{\mathrm{extend}} \\ &= (\mathrm{Conc}(\mathbf{v}_k, \mathbf{v}'_k) \odot \mathbf{m}_{\mathrm{noise},k})\,\mathbf{P}_{\mathrm{noise}}\mathbf{B}^{d}_{\mathrm{extend}}, \end{aligned} \tag{11}$$

where $\mathbf{d}_{\mathrm{noise},k} \in \mathbb{R}^{1\times(N+N')}$ is a weight vector of noise. $\mathbf{P}_{\mathrm{speech}}$ is an $(N+N')\times N$ orthogonal projection matrix whose (i,i) elements are 1 while the other elements are 0. $\mathbf{P}_{\mathrm{noise}}$ is an $(N+N')\times N'$ orthogonal projection matrix whose $(i+N,i)$ elements are 1 while the other elements are 0. In the end-to-end learning process of TasNet-NB, the basis signals of the encoder and decoder as well as the separator are updated so as to minimize a sum of the source and noise reconstruction loss defined by SI-SDR (6) with respect to $\mathbf{s}_{i,k}$ and $\mathbf{s}_{\mathrm{noise},k}$ defined in (10) and (11):

$$\mathcal{L} := \sum_{k=1}^{K}\sum_{i=1}^{C} \text{SI-SDR}(\mathbf{s}_{i,k}, \hat{\mathbf{s}}_{i,k}) + \sum_{k=1}^{K} \text{SI-SDR}(\mathbf{s}_{\mathrm{noise},k}, \hat{\mathbf{s}}_{\mathrm{noise},k}), \quad (12)$$

where $\text{SI-SDR}(\mathbf{s}_{i,k}, \hat{\mathbf{s}}_{i,k})$ and $\text{SI-SDR}(\mathbf{s}_{\mathrm{noise},k}, \hat{\mathbf{s}}_{\mathrm{noise},k})$ are SI-SDRs of the i-th speaker's speech signal and a noise in the k-th segment.

C. Curriculum learning

In our study, the curriculum consists of three steps in which the tasks are made progressively more difficult by decreasing the SNR in steps. In each step, we use the model trained in the previous step as the initial model.

Step 1. SNR of louder speaker to noise is 20dB
Step 2. SNR of louder speaker to noise is 10dB
Step 3. Original "Noisy" dataset

Referring to [16], SNRs are selected from a range of 8 dB and 20 dB. Considering the fact that one step (200 epochs) takes approximately 80 hours (in Section V), we restrict the maximum number of stages to 3. Because SNRs of the original "Noisy" dataset in the experimental setting (Section IV-A2) at the third step range from -6 to $+3$ dB, 10 dB and 20 dB are selected for SNRs at the first and second steps, respectively.

IV. Experiments

A. Datasets

We use a dataset wsj0-2mix [18] of two-speaker mixed speech and a dataset WHAM! [7] to evaluate the proposed method. The former is for normal speech separation without noise ("Clean") and the latter is for noisy speech separation ("Noisy").

1) wsj0-2mix: wsj0-2mix, which is widely used in single-channel speech separation, consists of two utterances randomly selected from the WSJ0 corpus [19] and mixed with a randomly selected signal-to-noise ratio in a range of -5 to +5 dB. The training set consists of 20,000 mixed utterances, and the validation set consists of 5,000 mixed utterances, which are selected from the WSJ0 corpus, si_tr_s, and some of the speakers are common. The evaluation set, on the other hand, consists of 3,000 mixtures of speech and uses only the speakers in the WSJ0 corpus, si_dt_05 and si_et_05, and does not have any speakers in common with the other sets. There are two versions, "min" and "max". "max" mixes the utterances without cropping, while "min" crops the utterances to the shorter length of the two selected utterances and then mixes them. In this paper, as in many other studies, we use the "min" version at 8 kHz.

2) WHAM!: In order to create a situation that is closer to the real environment than the wsj0-2mix dataset, we created a dataset that mixes the wsj0-2mix dataset with environmental sounds as noise. The noises were collected in coffee shops, restaurants, bars, office buildings, and parks in the San Francisco Bay area. Those containing speech above -6 dB, which is considered to interfere with speech separation, were not used. First, the SNR of the noise is randomly selected in the range of -6 to +3 dB, and the gain is set to achieve the selected signal-to-noise ratio for one of the two loud speakers to be mixed. Next, the same gain is used to mix the other speaker. The number of training sets, validation sets, and evaluation sets is the same as for the wsj0-2mix dataset. As for the verification based on the test sets, we assume that types and SNRs of noise are equivalent of the training sets.

B. Experimental settings

TasNet in the Asteroid Toolkit [20] was used as a baseline. The frame length L for speech segmentation is 40 that corresponds to a time window length of 5 ms. The number of basis signals for encoder and decoder N is 512, while the number of basis signals for noise N' is 128. The total number of basis signals is 640. A four-layer LSTM with 600 units in each direction is used as the separator, and a dropout of 0.3 is applied to the output of the LSTM except for the last layer. All implementations and parameter manipulations are done in the framework PyTorch [21]. The learning rate is halved when the validation loss does not improve for three consecutive epochs, with an initial value of $1.0e^{-3}$. In addition, early stopping is employed to stop learning when the validation loss does not improve for 10 epochs. A weighted decay of $1.0e^{-5}$ was used for normalization. Adam [22] was used as the optimization algorithm.

In this experimental setting, we assume that the number of speakers is known, and in particular that there are two speakers. However, when the number of speakers is unknown, separation can be achieved by recursively performing the operation of extracting the voices one by one in a time-domain system [23].

V. Results

Table I shows the results of noisy speech separation using TasNet and TasNet-NB, where TasNet_{512}, TasNet_{640}, NB_{128}, NL, CL represent TasNet with 512 basis signals, TasNet with 640 basis signals, additional 128 noise basis signals, training with noise reconstruction loss and training with curriculum learning, respectively. The scale-invariant signal-to-distortion ratio improvement (SI-SDRi) is the evaluation measure of the separation accuracy. For fairer comparison considering the difference in the number of basis signals between TasNet and TasNet-NB, results of TasNet with $N = 512$ and $N = 640$ are listed in the second (TasNet_{512}) and third (TasNet_{640}) rows,

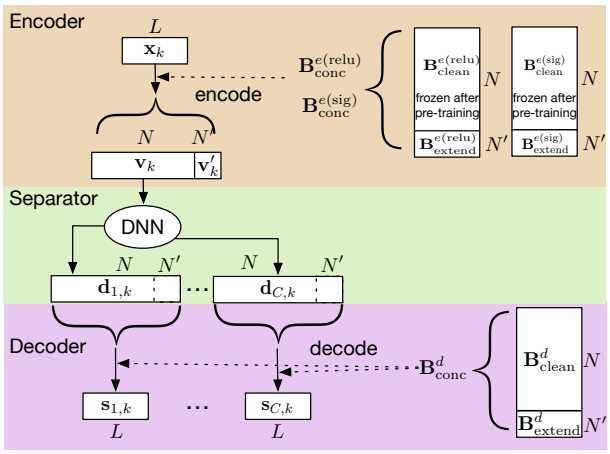

Fig. 4. Network of $\mathrm{TasNet}_{512} + \mathrm{NB}_{128}$.

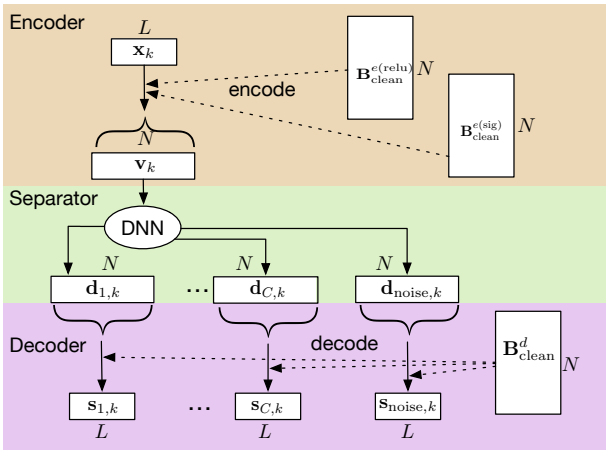

Fig. 5. Network of $\mathrm{TasNet}_{512} + \mathrm{NL}$.

respectively. The effect of the additional basis signals is a 0.1 dB increase in SI-SDRi. In the table, TasNet_{512} + NB_{128} + NL corresponds to the proposed TasNet-NB.

TasNet-NB (TasNet_{512} + NB_{128} + NL) outperforms the baseline TasNet (TasNet_{512}) by 0.5 dB in SI-SDRi. Further improvement is demonstrated when it is combined with curriculum learning (TasNet_{512} + NB_{128} + NL + CL), i.e., 14.6 dB in SI-SDRi.

Our ablation experiments on the effects of NB_{128}, NL and CL are also shown in Table I. After pre-training of TasNet_{512} based on "Clean" dataset, the training of TasNet_{512} + NB_{128} is done by updating basis signals other than $\mathbf{B}^{e(\mathrm{relu})}_{\mathrm{clean}}$ and $\mathbf{B}^{e(\mathrm{sig})}_{\mathrm{clean}}$ without using noise reconstruction loss (Fig. 4). In TasNet_{512} + NB_{128}, both $\mathbf{B}^{d}_{\mathrm{clean}}$ and $\mathbf{B}^{d}_{\mathrm{noise}}$ are used for the speaker reconstruction, i.e., $\mathbf{s}_{i,k} = \mathbf{d}_{i,k}\mathbf{B}^{d}_{\mathrm{conc}} = \mathbf{d}_{i,k}\mathrm{Conc}(\mathbf{B}^{d}_{\mathrm{clean}}, \mathbf{B}^{d}_{\mathrm{extend}})$. TasNet_{512} + NL updates $\mathbf{B}^{\gamma}_{\mathrm{clean}}$ by evaluating both signal and noise reconstruction loss while the number of basis signals equals to the baseline, i.e., $\mathbf{B}^{\gamma}_{\mathrm{extend}} = \varnothing$ (Fig. 5). Note that a weight vector of noise, $\mathbf{d}_{\mathrm{noise},k}$, is added in TasNet_{512} + NL for noise reconstruction.

The improvements in SI-SDRi of TasNet_{512} + NB_{128} compared with TasNet_{640} indicates that basis signals related to noise are implicitly stored in the additional NB_{128} without noise reconstruction loss. The results demonstrates that the two-step curriculum learning successfully restricts the noisy elements to the additional NB_{128}, while preserving the source separation accuracy of TasNet_{512}.

On the other hand, we observe that TasNet_{512} + NL is 0.5-dB superior to TasNet_{512} in SI-SDRi without additional basis signal. The results indicate that the basis signals trained by the "Clean" dataset is updated so as to learn basis signals of noise as well as sources in the fine-tuning phase by evaluating noise reconstruction loss.

Interestingly, we see no significant SI-SDRi improvement in TasNet_{512} + NB_{128} + NL compared with TasNet_{512} + NL. The results indicate that source separation accuracy is not deteriorated by updating the pre-trained $\mathbf{B}^{\gamma}_{\mathrm{clean}}$ using "Noisy" dataset. Based on the results of TasNet_{512} + NL, NB_{128} + NL and TasNet_{512} + NB_{128} + NL, we can conclude that 512 basis signals has sufficient capacity to model not only speakers' speeches but also background noise.

TABLE I
SI-SNRI ON WHAM! DATASET.
TasNet_{512}: TASNET WITH 512 BASIS SIGNALS.
TasNet_{640}: TASNET WITH 640 BASIS SIGNALS.
NB_{128}: 128 NOISE BASIS SIGNALS.
NL: NOISE RECONSTRUCTION LOSS.
CL: CURRICULUM LEARNING.

Method	SI-SDRi (dB)
TasNet_{512}	13.7
TasNet_{640}	13.8
TasNet_{512} + NB_{128}	14.0
TasNet_{512} + NL	14.2
TasNet_{512} + NB_{128} + NL (proposed TasNet-NB)	14.2
TasNet_{512} + NB_{128} + CL	14.6
TasNet_{512} + NB_{128} + NL + CL (proposed TasNet-NB + CL)	14.6

Curriculum learning contributes to further improvement in SI-SDRi, i.e., both TasNet_{512} + NB_{128} + NL + CL and TasNet_{512} + NB_{128} + CL demonstrate highest SI-SDRis of 14.6. Note that there is no difference in SI-SDRi due to the existence of noise reconstruction loss. An interpretation of the results is that gradual multi-step learning approach is beneficial to implicitly eliminate noisy elements, $\mathbf{B}^{d}_{\mathrm{extend}}$, after update of $\mathrm{Conc}(\mathbf{B}^{d}_{\mathrm{clean}}, \mathbf{B}^{d}_{\mathrm{extend}})$ without explicit guidance of noise reconstruction loss.

Training for 200 epochs takes approximately 80 hours with a NVIDIA Tesla P100. While effectiveness of curriculum learning is confirmed, learning 200 epochs at each stage requires a huge amount of time.

VI. Conclusions

In this paper, we have proposed TasNet-NB to improve the the separation accuracy of the original TasNet in noisy environment. SI-SDRi in the experimental setting using the WHAM! dataset was improved from 13.7 dB to 14.6 dB by TasNet-NB combined with curriculum learning. We also confirmed the effectiveness of using the reconstruction loss of noise as one of the learning criteria. In the future, we would verify the effectiveness using the dataset WHAMR![24], which includes not only additive noise but also reverberation.

References

[1] Y. Li and D. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Communication*, vol. 51, no. 3, pp. 230–239, 2009.

[2] Z. Wang, J. L. Roux and J. R. Hershey, "Alternative Objective Functions for Deep Clustering," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 686–690.

[3] Y. Luo and N. Mesgarani, "TasNet: time-domain audio separation network for real-time, single-channel speech separation," 2017.

[4] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," 2018.

[5] C. Xu, W. Rao, E. S. Chng and H. Li, "SpEx: Multi-Scale Time Domain Speaker Extraction Network," 2020.

[6] D. Ditter and T. Gerkmann, "A Multi-Phase Gammatone Filterbank for Speech Separation via TasNet," 2019.

[7] G. Wichern *et al*., "WHAM!: Extending Speech Separation to Noisy Environments," in *Interspeech*, 2019.

[8] Y. N. Dauphin, A. Fan, M. Auli and D. Grangier, "Language Modeling with Gated Convolutional Networks," 2016.

[9] J. L. Ba, J. R. Kiros and G. E. Hinton, "Layer Normalization," 2016.

[10] J. L. Roux, S. Wisdom, H. Erdogan and J. R. Hershey, "SDR - half-baked or well done?," 2018.

[11] M. Kolbæk, D. Yu, Z.-H. Tan and J. Jensen, "Multi-talker Speech Separation with Utterance-level Permutation Invariant Training of Deep Recurrent Neural Networks," 2017.

[12] H. Erdogan and T. Yoshioka, "Investigations on Data Augmentation and Loss Functions for Deep Learning Based Speech-Background Separation," in *Interspeech 2018*. 2018, ISCA.

[13] K. Kinoshita, T. Ochiai, M. Delcroix and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," 2020.

[14] Y. Bengio, J. Louradour, R. Collobert and J. Weston, "Curriculum Learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, pp. 41–48, Association for Computing Machinery.

[15] S. Braun, D. Neil and S.-C. Liu, "A Curriculum Learning Method for Improved Noise Robustness in Automatic Speech Recognition," 2016.

[16] S. Ranjan and J. H. L. Hansen, "Curriculum Learning Based Approaches for Noise Robust Speaker Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 197–210, 2018.

[17] R. K. Vuddagiri, H. K. Vydana and A. K. Vuppala, "Curriculum learning based approach for noise robust language identification using DNN with attention," *Expert Systems with Applications*, vol. 110, pp. 290–297, 2018.

[18] J. R. Hershey, Z. Chen, J. L. Roux and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," 2015.

[19] J. S. Garofolo, D. Graff, D. Paul and D. Pallett, "CSR-I (WSJ0) Sennheiser LDC93S6B," 1993.

[20] M. Pariente *et al*., "Asteroid: the PyTorch-based audio source separation toolkit for researchers," 2020.

[21] A. Paszke *et al*., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," 2019.

[22] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014.

[23] N. Takahashi, S. Parthasaarathy, N. Goswami and Y. Mitsufuji, "Recursive speech separation for unknown number of speakers," 2019.

[24] M. Maciejewski, G. Wichern, E. McQuinn and J. L. Roux, "WHAMR!: Noisy and Reverberant Single-Channel Speech Separation," 2019.