Journal of King Saud University – Computer and Information Sciences 35 (2023) 202–210

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Attention-based latent features for jointly trained end-to-end automatic speech recognition with modified speech enhancement

Da-Hee Yang, Joon-Hyuk Chang *

Department of Electronic Engineering, Hanyang University, Seoul 04763, Republic of Korea

ARTICLE INFO

Article history:
Received 27 December 2022
Revised 2 February 2023
Accepted 6 February 2023
Available online 15 February 2023

Keywords:
Time-domain speech enhancement
End-to-end automatic speech recognition
Attention-based latent feature
Joint training framework

ABSTRACT

In this paper, we propose a joint training framework that efficiently combines time-domain speech enhancement (SE) with an end-to-end (E2E) automatic speech recognition (ASR) system utilizing attention-based latent features. Using the latent feature to train E2E ASR implies that various time-domain SE models can be applied for noise-robust ASR and our modified framework is the first approach. We implement a fully E2E scheme pipelined from SE to ASR without domain knowledge and short-time Fourier transform (STFT) consistency constraints by applying a time-domain SE model. Therefore, using the latent feature of time-domain SE as appropriate features for ASR inputs is the main approach in our framework. Furthermore, we apply an attention algorithm to the time-domain SE model to selectively concentrate on certain latent features to achieve the better relevant feature for the task. Detailed experiments are conducted on the hybrid CTC/attention architecture for E2E ASR, and we demonstrate the superiority of our approach compared to baseline ASR systems trained with Mel filter bank coefficients features as input. Compared to the baseline ASR model trained only on clean data, the proposed joint training method achieves 63.6% and 86.8% relative error reductions on the TIMIT and WSJ "matched" test set, respectively.

1. Introduction

End-to-end (E2E) automatic speech recognition (ASR) systems have achieved significant advances (Graves and Jaitly, 2014; Chorowski et al., 2015; Chiu et al., 2018) with the advantage of directly predicting target sequences based on input speech. Nevertheless, the E2E ASR performance is still degraded under the influence of ambient background noise in real-world environments, which is an essential and challenging problem to address in ASR systems. There are two mainstream methods to achieve robustness against noise in E2E ASR systems. The first is the multi-condition training (MCT) that trains an ASR model using both clean and noisy data to improve ASR performance. The MCT method can improve the ASR performance against noise, but it still has limitations in that the performance improvement depends on the trained noisy environments and is affected by environmental distortion (Seltzer et al., 2013). The second method is to employ a speech enhancement (SE) module for the ASR model (Weninger et al., 2015; Wang et al., 2020; Gao et al., 2015; Wang and Wang, 2016). Depending on the application of the SE model, there are two approaches: front-end of ASR (Weninger et al., 2015; Wang et al., 2020) and joint training with ASR (Gao et al., 2015; Wang and Wang, 2016). In the front-end approach, SE modules enhance noisy speech and use enhanced speech for ASR systems. This approach can also improve the ASR performance to some extent but it cannot be fully optimized for higher ASR performance purpose (Seltzer, 2008) because the SE and ASR networks are trained separately, leading to a suboptimal problem. However, the joint training approach can simultaneously optimize the overall network (Mimura et al., 2016; Xu et al., 2019) to attain optimal performance and alleviate speech distortion (Narayanan and Wang, 2014; Menne et al., 2019).

In previous studies, SE and E2E ASR networks integrated with joint training have been widely applied for robust ASR (Wang and Wang, 2016; Liu et al., 2019; Fan et al., 2020; Li et al., 2021; Pandey et al., 2021; Kinoshita et al., 2020). Liu et al. (2019) jointly trained a mask-based SE network, attention-based

* Corresponding author.
E-mail addresses: douxi15@hanyang.ac.kr (D.-H. Yang), jchang@hanyang.ac.kr (J.-H. Chang).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.02.007
1319-1578/

encoder–decoder network, and discriminant network for noise-robust speech recognition. In addition, Fan et al. (2020) applied a gated recurrent fusion (GRF) algorithm to a joint network, suggesting that the joint training of SE and E2E ASR is better than that of the MCT method. In addition, Li et al. (2021) jointly trained a GAN-based SE network and an E2E ASR system, concluding that it is more robust to noise than the MCT method. However, previous studies (Wang and Wang, 2016; Liu et al., 2019; Fan et al., 2020; Li et al., 2021) have a limitation in that the domain of the SE models was limited to the time–frequency (TF)-domain. In contrast to previous studies, Kinoshita et al. (2020) proposed the use of a time-domain SE model for robust speech recognition, arguing that time-domain SE has more advantages than TF-domain SE and improves the ASR performance by adding noise-loss to the time-domain SE model. However, (Kinoshita et al., 2020) used the SE model as the front-end of the ASR. As mentioned in (Wang and Wang, 2016; Liu et al., 2019; Fan et al., 2020; Li et al., 2021), employing SE models as a front-end may be limited in obtaining a superior performance for the ASR system.

In this paper, we propose a joint training framework that efficiently integrates latent features of time-domain SE with E2E ASR. Recently, time-domain SE models have received considerable attention owing to their excellent performance. Unlike TF-domain SE models (Soni et al., 2018; Kim et al., 2020; Choi et al., 2018; Hu et al., 2020; Yin et al., 2020) that transform the input waveforms into spectral features via short-time Fourier transform (STFT), time-domain SE models (Luo and Mesgarani, 2019; Luo et al., 2020; Rethage et al., 2018; Pandey and Wang, 2019; Wang et al., 2021) operate mainly on raw waveforms. Therefore, time-domain SE models jointly enhance the magnitude and phase information without additional phase estimation algorithms. In addition, because time-domain SE is a fully E2E learning scheme that is entirely free from the constraints of STFT consistency (Wisdom et al., 2019; Nakaoka et al., 2021) and domain knowledge (Graves et al., 2013; Han et al., 2015), it can extract more appropriate features for the ASR task (Pandey and Wang, 2021; Kadıoğlu et al., 2020).

Nevertheless, time-domain SE models have rarely been utilized in joint training framework for robust E2E ASR. Unlike TF-domain SE models that can directly use enhanced spectral features for ASR training, time-domain SE models must extract spectral features for ASR training from enhanced waveforms. This process does not fully consider the advantages of time-domain SE models and cannot implement a complete E2E learning scheme. To overcome this limitation and implement a fully E2E learning scheme, we exploit a latent feature of the time-domain SE model without the need to reconstruct the waveform from a SE decoder module. Moreover, we apply an attention algorithm to a Conv-TasNet (Luo and Mesgarani, 2019) model to selectively concentrate on certain latent features to obtain better relevant features in noisy environments. Our contributions can be summarized as follows:

(1) For the first time, our framework provides an efficient approach for using latent features when performing joint training of time-domain SE and ASR networks.
(2) For better performance, we apply an attention algorithm to the time-domain SE model to extract more relevant latent features according to their relative importance.

The remainder of this paper is organized as follows. In Section 2, the original SE structure and E2E ASR network are reviewed and a general joint training method of TF-domain SE and E2E ASR is described. In Section 3, the attention-based SE structure and our joint training framework in the latent domain are introduced. In Section 4, various experimental setups and results are explored. Finally, in Section 5, the paper is concluded. In this paper, scalars are represented in lowercase, vectors in bold lowercase, and matrices in bold uppercase.

2. Related work

2.1. Original Conv-TasNet

Time-domain SE models have attracted considerable attention because they do not raise the phase-estimation issue (Luo and Mesgarani, 2019; Luo et al., 2020; Rethage et al., 2018; Pandey and Wang, 2019; Wang et al., 2021; Pascual et al., 2017). Among them, convolutional time-domain audio separation network (Conv-TasNet) is a popular model introduced by Luo and Mesgarani (2019) in the field of source separation and it specifically consists of encoder, mask estimation (separation), and decoder modules. The encoder module uses a one-dimensional (1-D) convolution operation to project input waveforms into a latent representation. The separation module then estimates a mask to suppress a particular interference at each time–frequency bin signal by taking the input latent representation obtained from a temporal convolutional network (TCN) comprising several 1-D convolutional blocks with dilation factors. The mask is estimated as follows:

$$\mathbf{M} = \mathcal{H}\left\{\sum_{r=1}^{R}\sum_{b=1}^{B}\mathcal{F}_{rb}(\mathbf{E}_{rb})\right\}, \tag{1}$$

where $\mathbf{E}_{rb}$ is the input latent representation of the $(r \times b)$-th block among several 1-D convolutional blocks comprising the TCN. B and R denote the numbers of consecutive convolutional blocks and repetitions, respectively. In addition, $\mathcal{F}$ and $\mathcal{H}$ are the convolution and nonlinear functions that constitute the estimation module. Consequently, the mask is multiplied by the input latent representation to generate a masked latent representation. The decoder module reconstructs the masked latent representations into waveforms using a transposed 1-D convolution operation. The corresponding structure for source separation can be applied to denoising tasks by estimating the noise mask (Koyama et al., 2020).

2.2. E2E ASR

E2E ASR models directly predict words or sequences from the input speech and are categorized into three main architectures: connectionist temporal classification (CTC) (Graves and Jaitly, 2014), attention-based encoder–decoder (AED) (Chan et al., 2016), and recurrent neural network transducer (Graves et al., 2012). In this study, we adopt the CTC/attention architecture widely used for E2E ASR (Watanabe et al., 2017), which utilizes both CTC and AED architectures. The CTC/attention architecture is designed by sharing the encoder module, and the total ASR model loss L_{ASR} is defined as follows:

$$L_{ASR} = \lambda L_{CTC} + (1-\lambda)L_{att}, \tag{2}$$

where λ is the weight that modulates the loss terms. In addition, L_{CTC} and L_{att} denote the CTC loss of CTC model and KL divergence loss of AED model, respectively.

2.2.1. Connectionist temporal classification

The CTC architecture (Graves and Jaitly, 2014) models a single output for each frame of the input sequence $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$, with T length input frames, without forced alignment between the input $\mathbf{X}$ and target $\mathbf{Y} = \{\mathbf{y}_u\}_{u=1}^{U}$ with U length output labels, where $\mathbf{y}_u \in \{1, \ldots, K\}$ represents the prediction label and K denotes the number of distinct labels. CTC obtains a possible set of output combinations $B(\mathbf{Y}, \mathbf{X})$ for each frame and eliminates redundant

sequences to predict the final output sequence. The log-likelihood function of all possible output sequences for the input feature $\mathbf{X}$ is as follows:

$$P(\mathbf{Y}|\mathbf{X}) = \sum_{\hat{\mathbf{y}} \in B(\mathbf{Y},\mathbf{X})} \prod_{t=1}^{T} P(\hat{\mathbf{y}}_t|\mathbf{X}), \tag{3}$$

where $\hat{\mathbf{y}}$ represents the predicted label sequence. Therefore, the CTC loss is expressed as follows:

$$L_{CTC} = -lnP(\mathbf{Y}|\mathbf{X}). \tag{4}$$

2.2.2. Attention-based encoder-decoder

The AED architecture (Chan et al., 2016) we use consists of encoder, decoder, and attention modules. The encoder module encodes the input feature $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{T}$ into $\mathbf{H} = \{\mathbf{h}_l\}_{l=1}^{L}$, where L is the number of frames in the encoder output. The attention module calculates the similarity between the output $\mathbf{H}$ of the encoder and decoder information $\mathbf{s}_{u-1}$, providing information on where the output sequence should focus on the input sequence. The following represents the computation of the attention mechanism at the u-th time step:

$$\mathbf{f}_u = F * a_{u-1}, \tag{5}$$

$$e_{u,l} = score(\mathbf{s}_{u-1}, \mathbf{h}_l) = w^T tanh\left(W\mathbf{s}_{u-1} + V\mathbf{h}_l + U\mathbf{f}_{u,l} + b\right), \tag{6}$$

$$a_{u,l} = softmax\left(e_{u,l}\right), \tag{7}$$

$$\mathbf{c}_u = \sum_{l=1}^{L} a_{u,l}\mathbf{h}_l, \tag{8}$$

where $\mathbf{f}_u$ is a convolutional feature vector of the previous attention weight a_{u-1} obtained by convolving with a trainable convolutional filter F. The location-based attention mechanism is employed to calculate the attention score $e_{u,l}$ from the decoder hidden state $\mathbf{s}_{u-1}$ at the previous output step of u and encoder hidden state $\mathbf{h}_l$. w, W, V, F, U, and b are the trainable parameters. Here, $a_{u,l}$ is computed by the softmax of the attention score $e_{u,l}$ and the attention context vector $\mathbf{c}_u$ is calculated by integrating all the inputs $\mathbf{h}_l$ based on the attention weight a_u over length L. The decoder generates the u-th output sequence $\hat{\mathbf{y}}_u$ using $\mathbf{y}_{u-1}, \mathbf{c}_u$, and $\mathbf{s}_{u-1}$. Each notation represents the previous output sequence, attention context vector, and the decoder hidden state.

$$\mathbf{y}_u = \mathrm{FFNN}(\mathbf{s}_{u-1}, \mathbf{c}_u), \tag{9}$$

$$\mathbf{s}_u = \mathrm{RNN}(\mathbf{s}_{u-1}, \mathbf{y}_u, \mathbf{c}_u), \tag{10}$$

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{u=1}^{U} P(\hat{\mathbf{y}}_u|\mathbf{X}, \mathbf{y}_{1:u-1}), \tag{11}$$

where FFNN represents the feed-forward neural network generating an output sequence $\mathbf{y}_u$ with the decoder hidden state and attention context vector. Subsequently, an RNN is used to produce a decoder hidden state $\mathbf{s}_u$. Finally, the probability distribution, given $\mathbf{X}$, is computed at each output step, conditioned on the previous outputs.

2.3. Conventional joint training method

The joint training of the TF-domain SE and E2E ASR networks has been utilized in various ways for noise-robust ASR. Previous studies (Wang and Wang, 2016; Liu et al., 2019; Fan et al., 2020; Li et al., 2021) integrated the TF-domain SE models for joint training with contemporary E2E ASR models (Liu et al., 2019; Fan et al., 2020; Li et al., 2021; Pandey et al., 2021) that primarily receive spectrogram-based features as an input. The conventional baseline framework for joint training is described as shown in Fig. 1(a): First, STFT is applied to the noisy signal $y(t) = x(t) + n(t)$ to produce $Y(t,f), X(t,f)$, and $N(t,f)$. The noisy input $\mathbf{Y}$ is fed into the estimation module to estimate the target clean $\mathbf{X}$ against $\mathbf{N}$, indicating the background noise. The estimation module yields the noise mask $\mathbf{M}_{TF}$ that eliminates noise from the noisy spectrogram and results in enhanced spectra.

$$|\widehat{\mathbf{X}}| = |\mathbf{Y}| \odot \mathbf{M}_{TF}, \tag{12}$$

where $\widehat{\mathbf{X}}$ is the enhanced spectrogram obtained from the TF domain SE network. Consecutively, a Mel filter bank coefficients (Fbank) are applied to the magnitude spectrogram $|\widehat{\mathbf{X}}|$ to extract the input features of E2E ASR. Since the output domain of the SE network is the same as the input domain of the ASR network, the two networks are jointly trained without additional modules as shown in Fig. 1(a).

3. Proposed methods

In this paper, we propose a joint training framework that efficiently integrates time-domain SE with an E2E ASR system utilizing latent representations, as shown in Fig. 1(b). Using the latent representation of the time-domain SE allows us to accomplish joint training directly in the latent domain for E2E ASR without any need to reconstruct a waveform. However, it is not possible to train an ASR system directly using latent representations as ASR inputs. A convolutional network with an absolute function allows the latent representation to be transformed into latent features for ASR training, instead of a Fbank component. In other words, the entire mechanism is jointly trained with the SE encoder, mask estimation module (without a decoder), convolutional network, and E2E ASR network. At this time, we modify the original Conv-TasNet into an attention-based Conv-TasNet to further improve performance.

3.1. Attention-based Conv-TasNet

As the original Conv-TasNet model equally adds the information of all 1-D convolution blocks, it has a limitation that it cannot utilize 1-D convolution blocks which are relatively important for denoising. To address this limitation, we apply learnable parameters to each 1-D convolutional block to provide different weights according to their relative importance while estimating a mask. Our proposed attention-based Conv-TasNet is shown in Fig. 2, where different weights are applied to the skip-connection paths of the 1-D convolutional blocks. The modified equation for the estimation module is described as follows:

$$\mathbf{L}_{rb} = \mathcal{F}_{rb}(\mathbf{E}_{rb}), \tag{13}$$

where $\mathbf{E}_{rb}$ is an input of the $(r \times b)$-th 1-D convolutional block in the TCN that comprises the estimation module. A successive 1-D convolutional block of B times with an increasing dilation factor is repeated R times. In addition, $\mathcal{F}_{rb}$ is the $(r \times b)$-th 1-D convolution function which has two outputs: residual and skip-connection paths.

$$\mathbf{M}_T = \mathcal{H}\left\{\sum_{r=1}^{R}\sum_{b=1}^{B} w_{rb} \cdot \mathbf{L}_{rb}\right\}, \tag{14}$$

The residual output is fed into the next $\mathcal{F}$ function and the skip-connection output $\mathbf{L}_{rb}$ is multiplied by w_{rb} to estimate the mask $\mathbf{M}_T$, where w_{rb} is the attention weight assigned to each $(r \times b)$-th output and is sequentially located in the attention block, as shown in Fig. 2. Since w_{rb} is a learnable parameter, it is determined by the relative importance of 1-D convolutional blocks during SE network

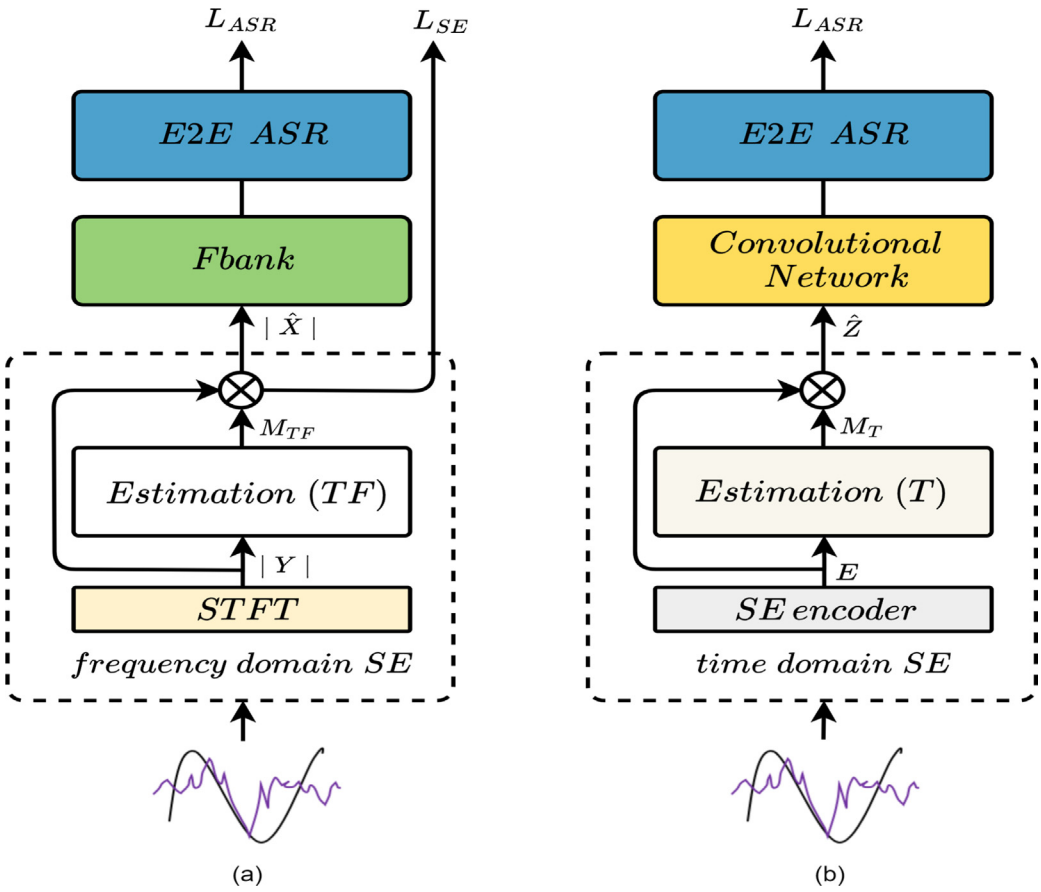

Fig. 1. Joint training framework of SE and E2E ASR. (a) Conventional joint training method with integrated TF-domain SE and E2E ASR with Fbank. (b) Proposed joint training method with integrated time-domain SE and E2E ASR with convolutional network.

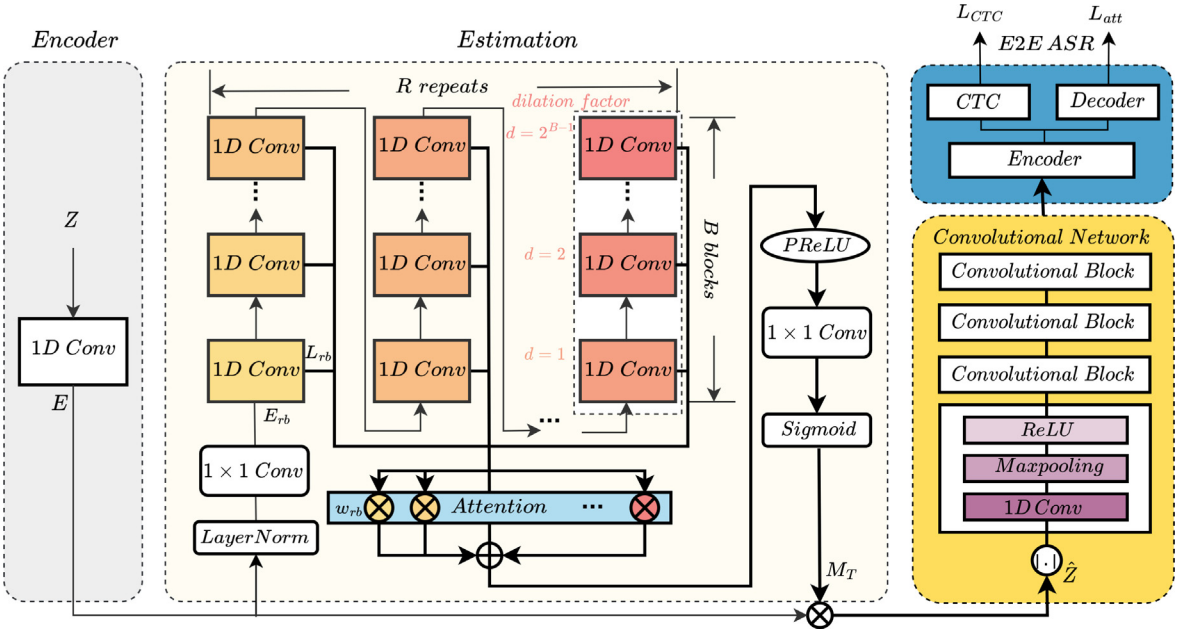

Fig. 2. The overall joint training framework with attention-based Conv-TasNet and E2E ASR network using the latent feature.

training. By assigning large parameters to relatively important blocks, it leads to optimal performance for denoising. In other words, the attention-based Conv-TasNet estimates the attention mask according to the relative importance of the 1D convolution blocks for more robust results. $\mathcal{H}\{.\}$ is a cascade of the PReLU, 1×1 convolution, and sigmoid function in the estimation module.

3.2. Convolutional network

In this paper, we introduce a convolutional network that subsamples the number of frames of the SE latent features. The convolutional network allows the latent representations of the time-domain SE to be trained on the ASR without being reconstructed into waveforms. The convolutional network consists of an absolute value function and four 1-D convolutional layers, as shown in Fig. 2. The absolute value function is applied, for which the ASR models use only magnitude spectra as the input. The absolute function is a key point in the joint training framework that allows training with latent features. The convolutional block consists of 1-D convolution, max-pooling and nonlinear activation function rectified linear units (ReLUs).

3.3. Joint training method on latent space

To implement our proposed method, we generate the noisy signal $z(t) = s(t) + n(t)$, where $s(t)$ and $n(t)$ denote the clean and noise signals, respectively. The input signal $z(t)$ can be divided into T overlapping frames of length L, denoted by $\mathbf{z}_t \in \mathbb{R}^L$, where $t = 1, \ldots, T$. The noisy signal is projected into a latent representation by the encoder module of the SE, as

$$\mathbf{E} = \mathbf{U} \cdot \mathbf{Z}, \tag{15}$$

where $\mathbf{Z} \in \mathbb{R}^{L \times T}$ is transformed into N-dimensional representations $\mathbf{E} \in \mathbb{R}^{N \times T}$ via multiplication through a trainable convolutional layer encoder $\mathbf{U} \in \mathbb{R}^{N \times L}$. The mask $\mathbf{M}_T \in \mathbb{R}^{N \times T}$ is the output of the estimation module and $\mathbf{M}_T$ is elementwise-multiplied to $\mathbf{E}$ to remove the noise, as expressed by

$$\hat{\mathbf{Z}} = \mathbf{E} \odot \mathbf{M}_T, \tag{16}$$

where $\hat{\mathbf{Z}} \in \mathbb{R}^{N \times T}$ becomes the enhanced latent-space attention representation that consists of N dimensions with T frames, and $\hat{\mathbf{Z}}$ is then fed into the convolutional network to be the trainable ASR input feature. The convolutional network takes a positive value from the input latent representation using an absolute value function.

The latent feature is extracted from the convolutional network and is used to train the E2E ASR system. Consequently, the overall framework, including an E2E ASR network, convolutional network, and attention-based Conv-TasNet, is jointly trained to optimize the entire network. Since we removed the SE decoder to efficiently integrate SE into ASR in the latent domain, the entire network is trained using the ASR loss (Heymann et al., 2017; Ochiai et al., 2017; Subramanian et al., 2019; Soni and Panda, 2019).

4. Experimental results

4.1. Dataset

We used two corpus datasets for our experiment: TIMIT (Garofolo, 1993) for a small dataset and wall street journal (WSJ) (Consortium et al., 1994) for a large dataset. To evaluate the performance in noisy environments, we generated noisy datasets by adding noise to clean TIMIT and WSJ datasets. For the noisy datasets, we prepared two types of noise data for evaluation in different noise conditions: CHiME-4 and NOISEX DB. The CHiME-4 noise dataset was recorded in the street, café, bus, and pedestrian environments. The NOISEX DB (Varga and Steeneken, 1993) contains babble, factory, and white noise.

TIMIT corpus: The TIMIT dataset consists of 10 sentences uttered by each of 630 speakers. The training set comprises 3696 utterances with eight sentences (the SA records of the training set were removed: i.e., identical sentences for all speakers in database) read by 462 speakers. The development set consists of 400 utterances and the core test set consists of 192 utterances. We used the phone error rate (PER) as an evaluation metric for phone recognition in the TIMIT data. PER is the number of phoneme errors (inserted, deleted, and changed phonemes) divided by the total number of phonemes. Lower values indicate better performance.

WSJ corpus: The WSJ dataset is a corpus of read English speech. It consists of train-si284 for training, test-dev93 for development, and test-eval92 for testing sets with 37416, 503, and 333 utterances, respectively. Word error rate (WER) was used as the evaluation metric for the WSJ dataset. It is the total number of word errors divided by the total number of words; lower values indicate better performance.

SE training noisy datasets: To pretrain the time-domain SE network, we mixed CHiME-4 noise with a clean dataset by selecting each utterance from TIMIT and WSJ. We added CHiME-4 noise to the clean data at signal-to-noise ratios (SNRs) randomly sampled between [0 dB and 20 dB]. Totally, 16,000 noisy utterances are generated in total by TIMIT and WSJ.

ASR training noisy datasets: We randomly mixed CHiME-4 noise from the training and development sets to generate a noisy dataset for the joint training of the network in which SNRs were randomly sampled between [0 dB and 20 dB].

Test noisy datasets: We generated the "matched" and "mismatched" test datasets. Test datasets were generated with SNRs of 0, 5, 10, 15, and 20 dB. The "matched" test dataset refers to the same environment as the training data and was generated using CHiME-4 noise with clean test sets of TIMIT and WSJ. The NOISEX DB was used for the "mismatched" test dataset, which means it is not the same as the training data environment.

4.2. Experimental setup

4.2.1. Conv-TasNet

The parameters of Conv-TasNet and attention-based Conv-TasNet are listed in Table 1 as follows: N = 512, L = 40, C = 128 Sc =128, H = 512, P = 3, B = 8, and $R = 3$. We adopted global layer normalization, Adam optimization algorithm (Kingma and Ba, 2015), and sigmoid function as the activation function. The sampling rate of the dataset was 16 kHz.

4.2.2. Convolutional network

We extracted latent features by using four 1-D convolutional layers, each with 512, 256, 128, and 128 filters. For the TIMIT data, the first layer adopted a filter size of 9 and the rest adopted a filter size of 3. In the WSJ, all the layers had a filter size of 3. These parameters were chosen so that the number of frames of the latent feature is similar to that of the original ASR input feature. We used max-pooling of length 2 to reduce the length of the latent feature and used ReLUs as an activation function. We also applied an absolute value function to the input latent representation, resulting in positive distributions. Because we considered the characteristic

Table 1
Hyperparameters of Conv-TasNet.

Symbol	Description
N	Number of filters in encoder and decoder
L	Length of the filters (in samples)
C	Number of channels in bottleneck and the residual paths' 1-D conv blocks
Sc	Number of channels in skip-connection paths' of 1-D conv blocks
H	Number of channels in convolutional blocks
P	Kernel size in convolutional blocks
B	Number of convolutional blocks in each repeat
R	Number of repeats

that the ASR model takes magnitude values as an input and produces a positive distribution. In addition to the absolute value function, we applied ReLUs and a square function to obtain a positive distribution. However, these activation functions did not work. We assumed that the activation functions did not play a positive role in training of the ASR model because they tend to oversmooth or distort the latent representation.

4.2.3. E2E ASR model

We adopted the hybrid CTC/attention architecture with an RNN structure for the E2E ASR system. Our method was implemented using the ESPnet toolkit (Watanabe et al., 2018). For comparison with the proposed latent features, the Fbank features were used to train the baseline model. For TIMIT and WSJ, 23 and 80 mel-scale filterbank coefficients with a window length of 25 and a window shift of 10 ms were used, respectively, as in (Parcollet et al., 2020). To train the ASR model, five- and six-layer Bi-GRUP with 512 units ASR encoder were used for the TIMIT and WSJ data, respectively. We trained the model for 20 and 15 epochs and applied CTC loss weights as 0.5 and 0.2, respectively.

4.3. Experimental results

4.3.1. Effect of attention algorithm applied to the Conv-TasNet model

We conducted experiments in two ways to show that the proposed attention-based Conv-TasNet is more effective for ASR systems, which consisted of:

(i) an experiment to train the ASR system using the Fbank feature; and
(ii) an experiment to train an ASR system using the latent feature.

The Fbank feature is a widely used feature type in E2E ASR systems that are trained by extracting pre-computed features from speech signals. To compare the ASR performance of Conv-TasNet and attention-based Conv-TasNet using Fbank features, STFT transformation was performed on the enhanced speech signal of each SE network, and then, log Mel filterbank coefficients were applied. In contrast, the latent feature was extracted from each time-domain SE network without waveform reconstruction to efficiently train the E2E ASR system. Each SE network was pretrained with SE training datasets and then frozen to extract latent features for training ASR systems. Experimental results indicate that applying an attention algorithm to Conv-TasNet is effective with only small parameter increase for ASR training. The results are listed in Tables 2 and 3. Each table represents the PER and WER results for the TIMIT and WSJ datasets, respectively. First, as shown in Tables 2 and 3, the attention-based Conv-TasNet improves the speech recognition performance for both Fbank and latent features. Therefore, in the following results, attention-based Conv-TasNet can acquire more robust results than Conv-TasNet, regardless of the feature type. An interesting fact from Tables 2 and 3 is that the performance improvement gap of the latent feature outweighs the performance improvement gap of the Fbank feature. This means that the attention algorithm is more effective on the latent feature than on the Fbank feature. Since we applied a learnable parameter to each 1-D convolutional block of Conv-TasNet, the parameter increases by the number of 1-D convolutional blocks. In other words, the attention-based Conv-TasNet increased the number of parameters by 24 compared to the Conv-TasNet because 24 blocks were used in our experiment.

In addition, we demonstrated that the latent feature is more robust to noise than the Fbank feature. In Tables 2 and 3, attention-based Conv-TasNet with latent results outperformed that of with Fbank results. And the same results were observed in the case of Conv-TasNet model. For this reason, we assumed that the latent features have more information and less distortion than the Fbank features extracted using filters on the speech signal. From these results, we demonstrated that the latent feature generated from attention-based Conv-TasNet is efficient for the integration of time-domain SE and E2E ASR.

4.3.2. Effect of joint training with attended latent feature

Our experiments demonstrated that extracting latent features by applying attention to Conv-TasNet is the most effective approach for E2E ASR performance. Based on the above results, we conducted an experiment to jointly train the SE and E2E ASR networks in the latent domain for noise-robust speech recognition. The annotations in Tables 4–7 are as follows:

(i) E2E_ASR-Clean: The ASR system was trained using only clean data. This baseline model shows three test results (None, Conv-TasNet, and attention-based Conv-TasNet) according to the SE networks.
(ii) E2E_ASR-MCT: The ASR system was trained using multi-condition data. This baseline model shows three test results (None, Conv-TasNet, and attention-based Conv-TasNet) according to the SE networks.
(iii) E2E_ASR-SE: The ASR system was trained using latent features. The proposed model shows three test results (Conv-TasNet without joint training, attention-based Conv-TasNet without joint training, and attention-based Conv-TasNet with joint training) according to the SE networks and joint training. All SE networks were used in addition to each pre-trained SE network.

Tables 4 and 5 present the experimental results for the TIMIT data and Tables 6 and 7 list the WSJ data. For the test dataset, "matched" and "mismatched" sets were used.

The baseline models with E2E_ASR-Clean and E2E_ASR-MCT are listed in Table 4. The MCT is one of the mainstream methods of noise-robust ASR systems and preprocessing through SE networks is also one of the method. E2E_ASR-Clean performs 60.5% and 64.4% in "matched" and "mismatched" noisy environments, respectively. These results indicate that ASR systems are highly susceptible to noise. In contrast, E2E_ASR-MCT improved the performance by 30.1% and 36.8%. In addition, we also applied enhancement networks, including Conv-TasNet and attention-based Conv-TasNet, as the front-end of ASR to slightly improve

Table 2
PER results (%) of the E2E ASR system trained by Fbank features and Latent features with speech enhancement including Conv-TasNet and Attention-based Conv-TasNet on TIMIT test sets.

Model	Fea.	PER (%)					
		0 dB	5 dB	10 dB	15 dB	20 dB	Average
Conv-TasNet	Fbank	31.3	26.2	23.9	22.8	22.6	25.36
Attention-based Conv-TasNet		31.1	26.0	23.4	22.8	22.5	25.16
Conv-TasNet	**Latent**	31.2	25.5	22.6	20.8	20.3	24.08
Attention-based Conv-TasNet		**30.0**	**24.8**	**21.7**	**20.3**	**20.2**	**23.4**

Table 3
WER results (%) of the E2E ASR system trained by Fbank features and Latent features with speech enhancement including Conv-TasNet and Attention-based Conv-TasNet on WSJ test sets.

Model	Fea.	WER (%)					
		0 dB	5 dB	10 dB	15 dB	20 dB	Average
Conv-TasNet	Fbank	16.1	10.0	8.5	8.0	7.7	10.06
Attention-based Conv-TasNet		16.0	10.4	8.1	7.6	7.5	9.92
Conv-TasNet	**Latent**	14.7	9.7	**7.1**	6.6	6.6	8.94
Attention-based Conv-TasNet		**14.2**	**8.9**	7.3	**6.5**	**6.3**	**8.66**

Table 4
PER results (%) of the E2E ASR system trained by clean and multi-condition data with and without the speech enhancement on TIMIT development (dev.) and test set. Fbank features were used as baseline models.

Model	Preprocessing	Fea.	PER (%)			
			Matched		Mismatched	
			Dev.	Test	Dev.	Test
E2E_ASR-Clean	None	Fbank	59.0	60.5	63.0	64.4
	Conv-TasNet		25.6	27.5	31.3	33.4
	Attention-based Conv-TasNet		25.4	26.8	30.7	32.2
E2E_ASR-MCT	None	Fbank	29.0	30.1	35.9	36.8
	Conv-TasNet		23.8	24.6	28.7	30.5
	Attention-based Conv-TasNet		23.7	24.6	28.4	28.9

Table 5
Impacts of using latent features with and without joint training on TIMIT development (dev.) and test set. Results are in PER (%).

Model	Preprocessing	Joint	Fea.	PER (%)			
				Matched		Mismatched	
				Dev.	Test	Dev.	Test
E2E_ASR-SE	Conv-TasNet	–	**Latent**	23.6	24.2	28.6	30.1
	Attention-based Conv-TasNet	–		22.8	23.4	27.1	28.8
	Attention-based Conv-TasNet	✓		**20.7**	**22.0**	**26.0**	**27.9**

Table 6
WER results (%) of the E2E ASR system trained by clean and multi-condition data with and without the speech enhancement on WSJ development (dev.) and test set. Fbank features were used as baseline models.

Model	Preprocessing	Fea.	WER (%)			
			Matched		Mismatched	
			Dev.	Test	Dev.	Test
E2E_ASR-Clean	None	Fbank	78.5	68.0	64.6	67.6
	Conv-TasNet		20.9	15.4	31.1	27.0
	Attention-based Conv-TasNet		21.4	15.7	30.9	25.8
E2E_ASR-MCT	None	Fbank	17.8	12.0	22.3	16.4
	Conv-TasNet		16.3	10.6	21.1	15.1
	Attention-based Conv-TasNet		16.5	11.0	20.4	15.0

Table 7
Impacts of using latent features with and without joint training on WSJ development (dev.) and test set. Results are in WER (%).

Model	Preprocessing	Joint	Fea.	WER (%)			
				Matched		Mismatched	
				Dev.	Test	Dev.	Test
E2E_ASR-SE	Conv-TasNet	–	**Latent**	13.6	9.7	19.7	14.9
	Attention-based Conv-TasNet	–		13.5	9.5	19.6	14.4
	Attention-based Conv-TasNet	✓		**13.3**	**9.0**	**17.7**	**12.9**

performance. As listed in Table 4, SE networks notably improves the performance of the ASR system. Especially, attention-based Conv-TasNet applied to E2E_ASR-MCT achieved good performance at 24.6% and 28.9% in "matched" and "mismatched", respectively.

Based on the above experimental results, we utilized the latent feature of SE networks for ASR training. As Table 5 presents the results, both Conv-TasNet and attention-based Conv-TasNet showed more robust results than the results of the baseline models just by altering the feature type to "Latent". Indeed, attention-based Conv-TasNet without joint training improved further by 0.8% and 1.3% on "matched" and "mismatched" test sets, respectively, compared to Conv-TasNet. Finally, we jointly train the entire

Table 8
Comparison of the number of parameters and relative inference time according to the E2E_ASR model.

Model	Joint	Fea.	Params (M)
E2E_ASR-Baseline	–	Fbank	20.63
E2E_ASR-SE	–	Latent	22.28
	✓	Latent	29.88

network according to the facts demonstrated above: attention-based Conv-TasNet is more robust than Conv-TasNet; and the latent feature is more robust than the Fbank feature. Moreover, our joint training framework shows the most robust results to noise with 22.0% and 27.9% on "matched" and "mismatched" test sets, respectively.

These processes were repeated for the WSJ data, and the results are presented in Tables 6 and 7. Similar to the above results, E2E_ASR-Clean performed poorly by 68.0% in the "matched" and 67.7% in the "mismatched" test sets. In addition, E2E_ASR-MCT shows significant improvements of 12.0% and 16.4%. Applying the SE network as a front-end of the ASR can further improve performance to some degree. However, this is still not an optimal performance result. To improve performance, we used latent features for training the E2E ASR system, as shown in Table 7. Furthermore, to implement a more optimal performance, we jointly trained the entire network using attended latent features. Our joint training framework shows 9.0% and 12.9% in the "matched" and "mismatched" test sets, respectively, showing the most robust results against noise in all environments.

In addition, in the TIMIT data, the number of parameters according to the E2E_ASR model were compared as shown in Table 8. E2E_ASR-Baseline includes E2E_ASR-Clean and E2E_ASR-MCT using the Fbank feature for training. The baseline model had 20.63 M parameters. Because it uses the pre-computed feature, Fbank. When a latent feature is used instead of the Fbank feature, the computational cost increases because a learnable feature is extracted using a convolutional network. Therefore, E2E_ASR-SE without joint training used 22.28 M learnable parameters. Further, since joint training with SE and ASR networks combines the two networks, more parameters were required to train the model.

5. Conclusions

In this study, we proposed a method for joint training with a time-domain SE network and an E2E ASR network by using latent features. For robust ASR, time-domain SE models have not been widely applied compared to TF-domain SE models because domain mismatch causes inefficiency. However, as time-domain SE models have received increasing attention, it is necessary to integrate them with ASR systems. Performing joint training with E2E ASR by simply extracting spectral features from enhanced waveforms of time-domain SE is an easy task, but this method does not fully utilize the advantages of time-domain SE. We proposed for the first time that the integration of the two networks in the latent domain will provide good guidance for joint training with the time-domain SE model and the E2E ASR system. Therefore, various time-domain SE model and ASR models can be integrated in the future. In addition to hybrid CTC/attention architecture for E2E ASR, we will conduct a study to integrate various E2E ASR systems and time-domain SE models using latent features via joint training.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University)), and in part by the Technology Innovation Program (20013726, Development of Industrial Intelligent Technology for Manufacturing, Process, and Logistics) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea).

References

Chan, W., Jaitly, N., Le, Q., Vinyals, O., 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4960–4964.
Chiu, C.-C., Sainath, T.N., Wu, Y., Prabhavalkar, R., Nguyen, P., Chen, Z., Kannan, A., Weiss, R.J., Rao, K., Gonina, E., et al., 2018. State-of-the-art speech recognition with sequence-to-sequence models. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4774–4778.
Choi, H.-S., Kim, J.-H., Huh, J., Kim, A., Ha, J.-W., Lee, K., 2018. Phase-aware speech enhancement with deep complex U-Net. In: International Conference on Learning Representations.
Chorowski, J.K., Bahdanau, D., Serdyuk, D., Cho, K., Bengio, Y., 2015. Attention-based models for speech recognition. Adv. Neural Informat. Process. Syst. 28.
Consortium, L.D., et al., 1994. CSR-II (WSJ1) complete, Linguistic Data Consortium, Philadelphia, vol. LDC94S13A.
Fan, C., Yi, J., Tao, J., Tian, Z., Liu, B., Wen, Z., 2020. Gated recurrent fusion with joint training framework for robust end-to-end speech recognition. IEEE/ACM Trans. Audio Speech Language Process. 29, 198–209.
Gao, T., Du, J., Dai, L.-R., Lee, C.-H., 2015. Joint training of front-end and back-end deep neural networks for robust speech recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4375–4379.
Garofolo, J.S., 1993. Timit acoustic phonetic continuous speech corpus. Linguis. Data Consort. 1993.
Graves, A., Jaitly, N., 2014. Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning (ICML), pp. 1764–1772.
Graves, A., 2012. Sequence transduction with recurrent neural networks. In: ICML Representation Learning Workshop.
Graves, A., Mohamed, A.-R., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6645–6649. https://doi.org/10.1109/ICASSP.2013.6638947.
Han, K, He, Y., Bagchi, D., Fosler-Lussier, E., Wang, D., 2015. Deep neural network based spectral feature mapping for robust speech recognition. In: INTERSPEECH.
Heymann, J., Drude, L., Boeddeker, C., Hanebrink, P., Haeb-Umbach, R., 2017. Beamnet: End-to-end training of a beamformer-supported multi-channel asr system. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5325–5329.
Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., Xie, L., 2020. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement. In: INTERSPEECH, pp. 2472–2476.
Kadıoğlu, B., Horgan, M., Liu, X., Pons, J., Darcy, D., Kumar, V., 2020. An empirical study of Conv-TasNet. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7264–7268.
Kim, J., El-Khamy, M., Lee, J., 2020. T-GSA: Transformer with gaussian-weighted self-attention for speech enhancement. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6649–6653.
Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR).
Kinoshita, K., Ochiai, T., Delcroix, M., Nakatani, T., 2020. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In: EEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7009–7013.
Koyama, Y., Vuong, T., Uhlich, S., Raj, B., 2020. Exploring the best loss function for DNN-based low-latency speech enhancement with temporal convolutional networks, arXiv preprint arXiv:2005.11611.
Li, L., Kang, Y., Shi, Y., Kürzinger, L., Watzel, T., Rigoll, G., 2021. Adversarial joint training with self-attention mechanism for robust end-to-end speech recognition. EURASIP J. Audio Speech Music Process. 2021 (1), 1–16.
Liu, B., Nie, S., Liang, S., Liu, W., Yu, M., Chen, L., Peng, S., Li, C., 2019. Jointly adversarial enhancement training for robust end-to-end speech recognition. In: INTERSPEECH, pp. 491–495.
Luo, Y., Mesgarani, N., 2019. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM Trans. Audio Speech Language Proces. 27 (8), 1256–1266.

Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 46–50.
Menne, T., Schlüter, R., Ney, H., 2019. Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6660–6664.
Mimura, M., Sakai, S., Kawahara, T., 2016. Joint optimization of denoising autoencoder and DNN acoustic model based on multi-target learning for noisy speech recognition. In: INTERSPEECH, pp. 3803–3807.
Nakaoka, S., Li, L., Inoue, S., Makino, S., 2021. Teacher-student learning for low-latency online speech enhancement using Wave-U-Net. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 661–665.
Narayanan, A., Wang, D., 2014. Joint noise adaptive training for robust automatic speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2504–2508.
Ochiai, T., Watanabe, S., Hori, T., Hershey, J.R., 2017. Multichannel end-to-end speech recognition. In: International Conference on Machine Learning, PMLR, pp. 2632–2641.
Pandey, A., Wang, D., 2019. A new framework for CNN-based speech enhancement in the time domain. IEEE/ACM Trans. Audio Speech Language Process. 27 (7), 1179–1188.
Pandey, A., Wang, D., 2021. Dense CNN with self-attention for time-domain speech enhancement. IEEE/ACM Trans. Audio Speech Language Process. 29, 1270–1279.
Pandey, A., Liu, C., Wang, Y., Saraf, Y., 2021. Dual application of speech enhancement for automatic speech recognition. In: IEEE Spoken Language Technology Workshop (SLT), pp. 223–228.
Parcollet, T., Morchid, M., Linares, G., 2020. E2E-SINCNET: Toward fully end-to-end speech recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7714–7718.
Pascual, S., Bonafonte, A., Serrà, J., 2017. SEGAN: Speech enhancement generative adversarial network. In: INTERSPEECH, pp. 3642–3646.
Rethage, D., Pons, J., Serra, X., 2018. A wavenet for speech denoising. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 5069–5073.
Seltzer, M.L., 2008. Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays. In: 2008 Hands-Free Speech Communication and Microphone Arrays, pp. 104–107. https://doi.org/10.1109/HSCMA.2008.4538698.
Seltzer, M.L., Yu, D., Wang, Y., 2013. An investigation of deep neural networks for noise robust speech recognition. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7398–7402.
Soni, M.H., Panda, A., 2019. Label driven time-frequency masking for robust continuous speech recognition. In: INTERSPEECH, pp. 426–430.
Soni, M.H., Shah, N., Patil, H.A., 2018. Time-frequency masking-based speech enhancement using generative adversarial network. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5039–5043.
Subramanian, A.S., Wang, X., Baskar, M.K., Watanabe, S., Taniguchi, T., Tran, D., Fujita, Y., 2019. Speech enhancement using end-to-end speech recognition objectives. 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, pp. 234–238.
Varga, A., Steeneken, H.J., 1993. Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. 12 (3), 247–251.
Wang, Z.-Q., Wang, D., 2016. A joint training framework for robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Language Process. 24 (4), 796–806. https://doi.org/10.1109/TASLP.2016.2528171.
Wang, Z.-Q., Wang, P., Wang, D., 2020. Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR. IEEE/ACM Trans. Audio Speech Language Process. 28, 1778–1787.
Wang, K., He, B., Zhu, W.-P., 2021. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 7098–7102.
Watanabe, S., Hori, T., Kim, S., Hershey, J.R., Hayashi, T., 2017. Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE J. Sel. Top. Signal Process. 11 (8), 1240–1253.
Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N.E.Y., Heymann, J., Wiesner, M., Chen, N., et al., 2018. Espnet: sEnd-to-end speech processing toolkit. In: INTERSPEECH, pp. 2207–2211.
Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J.R., Schuller, B., 2015. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR. In: International Conference on Latent Variable Analysis and Signal Separation. Springer, pp. 91–99.
Wisdom, S., Hershey, J.R., Wilson, K., Thorpe, J., Chinen, M., Patton, B., Saurous, R.A., 2019. Differentiable consistency constraints for improved deep speech enhancement. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 900–904.
Xu, Y., Weng, C., Hui, L., Liu, J., Yu, M., Su, D., Yu, D., 2019. Joint training of complex ratio mask based beamformer and acoustic model for noise robust ASR. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6745–6749.
Yin, D., Luo, C., Xiong, Z., Zeng, W., 2020. PHASEN: A phase-and-harmonics-aware speech enhancement network. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 9458–9465.