2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 979-8-3503-9796-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISCSLP57327.2022.10037802

Text-Informed Knowledge Distillation for Robust Speech Enhancement and Recognition

Wei Wang, Wangyou Zhang, Shaoxiong Lin, Yanmin Qian[†]

MoE Key Lab of Artificial Intelligence, AI Institute
X-LANCE Lab, Department of Computer Science and Engineering, Shanghai Jiao Tong University

wangwei.sjtu@sjtu.edu.cn, wyz-97@sjtu.edu.cn, Johnson-Lin@sjtu.edu.en,
yanminqian@sjtu.edu.cn

Abstract

Most existing speech enhancement (SE) approaches heavily depend on simulated data for training, leading to performance degradation on realistic data and subsequent speech recognition task. One of the main reasons is that SE models cannot be trained on real data due to the absence of reference signals. In this paper, we aim to tackle this problem by exploiting transcribed real data to mitigate the mismatch between training and evaluation. A text-informed SE teacher is first trained to provide "reference" signals for the transcribed real data. Then a SE student is trained on both simulated and real data, where the supervision comes from the simulated ground truth and the teacher, respectively. Finally, a speech recognition model is trained on enhanced signals from the SE student. Our experimental results show that the proposed method can not only improve the speech enhancement performance, but also reduce the word error rate on the downstream speech recognition task.

Index Terms: robust speech recognition, speech enhancement, knowledge distillation, multi-modality

1. Introduction

Many speech related applications, such as automatic speech recognition (ASR) and speaker verification, require speech enhancement (SE) as an indispensable front-end to improve the intelligibility and perceptual quality of degraded speech signals. Although many efforts have been made to build state-of-the-art speech enhancement models [1, 2, 3], single-channel speech enhancement remains challenging when dealing with real data, leading to performance degradation on the downstream speech recognition task.

One challenge of speech enhancement is the discrepancy between training and evaluation conditions [4]. Unlike speech recognition where the ground truth label can be easily annotated for real-world data, the parallel clean speech signal in speech enhancement is often unavailable when collecting real-world data. Therefore, most speech enhancement systems have to be built on simulated speech data. However, the simulation process usually covers only limited noise conditions and types [5], which can lead to performance degradation in unseen noise conditions. Moreover, since the signal level criteria used in speech enhancement are not directly correlated with speech recognition task, performance degradation on speech recognition is often observed with enhanced signals [6, 7].

Many attempts have been made to mitigate the mismatch between training and evaluation, and they can be divided into five main categories. (1) Data augmentation. One popular direction is to increase the noise diversity by involving as many noise conditions as possible during simulation. The motivation is that the evaluation condition can be probably covered by the training data and the model generalizability can be improved. Various data augmentation strategies have been explored, such as collecting large-scale real noise for training [8], noise perturbation [9], and generating noise based on a set of well-designed noise bases [5]. (2) Noise modeling. This approach takes into account the noise information explicitly in the model design or the training process, so that the model is guided to adapt to different types of noise. Various directions on noise modeling have been explored. For example, [10, 11] proposed to incorporate the predicted noise information into speech estimation. [12] proposed to train a noise-robust speech enhancement model via domain adversarial training (DAT). [13, 6] investigated improving speech enhancement with the multi-task loss by adding a noise related loss. (3) Generative adversarial networks (GANs) based approaches. Prior work [14, 15] has investigated the use of GANs to for speech enhancement with real data. The speech enhancement model, as the generator, separate the speech signal from the noisy input, while the discriminator tries to distinguish the enhanced signals from the true clean speech signal. (4) End-to-end training with downstream tasks. By training speech enhancement and downstream models as one single system in an end-to-end manner, the dependency on clean speech reference signals for training can be avoided by only using the final loss in the downstream task. This approach thus enables utilizing a large amount of real data for training, which implicitly reduces the mismatch between training and evaluation. Many prior studies have been conducted in this direction with different downstream tasks, such as speech recognition [16, 17] and speaker verification [18]. (5) Auxiliary Information. Instead of only exploiting audio information for speech enhancement, some research focuses on using auxiliary information from other modalities to improve speech enhancement. Different modalities have been explored in speech enhancement, including speaker identity [19, 20, 21], text information [22, 23], and visual clues [24, 25].

Although the aforementioned end-to-end training and GAN-based approaches can be applied, the speech enhancement performance in end-to-end training is not guaranteed to be comparable to separately trained SE models [16, 17], and GAN-based speech enhancement has a complicated training process that requires careful tuning of each component. In this paper, we propose a novel framework to utilize the real speech data for training speech enhancement models based on knowledge distillation [26, 27, 28]. The proposed method works with real speech transcription during training of the speech enhancement model and is more correlated with speech recognition task. First, a text-informed speech enhancement model is trained as the teacher on the simulated data, taking both noisy signals and

[†] corresponding author

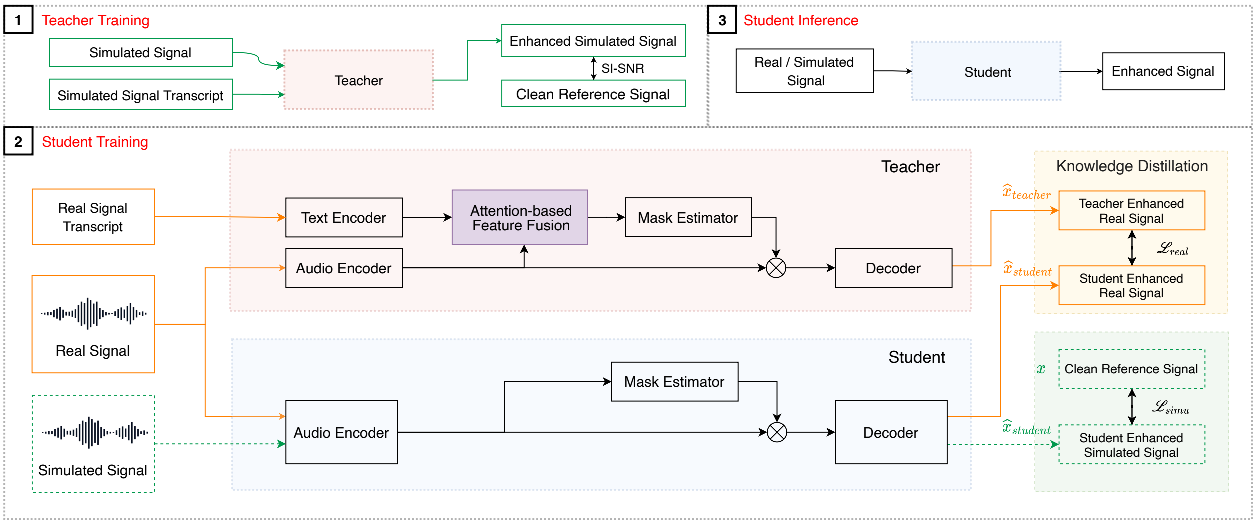

Figure 1: *Illustration of the proposed text-informed knowledge distillation framework;* $\mathcal{L}_{real} = \textit{SI-SNR}(\hat{x}_{teacher}, \hat{x}_{student})$, $\mathcal{L}_{simu} = \textit{SI-SNR}(x, \hat{x}_{student})$

transcriptions as input for estimating clean signals. It is then used to estimate clean "reference" signals on the real data. Then the audio-only speech enhancement model, as the student, is trained on both simulated and real data, with labels from the simulated ground truth and estimates of the teacher, respectively. Finally, the well-trained student model is used for inference, which does not rely on parallel text information as input. Experiments show that our proposed approach can improve the robustness of the speech enhancement model in terms of both speech enhancement and downstream ASR performance. Note that our proposed framework only requires auxiliary information from other modalities (e.g. text) during training, while previous guided speech enhancement works [19, 20, 22, 23, 24, 25] often assume access to such information for both training and inference and thus cannot be directly applied to downstream ASR task.

The remainder of this paper is organized as follows. In Section 2, we introduce the proposed knowledge distillation framework and the text-informed model architecture. The detailed experimental results and analysis are described in Section 3, and finally, we conclude the paper in Section 4.

2. Text-Informed Knowledge Distillation Framework

Our proposed framework involves the training of a teacher and a student model to exploit transcribed real speech data for more robust speech enhancement and recognition. The teacher model is trained to incorporate text and audio information and improve the quality of enhanced audios. In this way, the well-trained teacher model can estimate "reference" signals for transcribed real speech data and assist in training the student model. We refer to this framework as text-informed knowledge distillation.

By utilizing the real speech data for training under the above framework, the performance of the student model can be improved without increasing the number of parameters and computational cost. While our proposed framework is a general approach and can be applied to various types of speech enhancement models, we adopt the popular time-domain Conv-TasNet [29] structure for both teacher and student models in this paper.

2.1. Text-Informed Teacher Model

The structure of the text-informed teacher model is illustrated in the red blocks of Figure 1. The audio encoder, mask estimator and decoder adopts the same architecture as in [29].

The text encoder is a stack of transformer blocks, transforming the transcription into an intermediate textual feature space that can perform text and audio feature fusion. The feature fusion block combines encoded text features and audio features through the attention mechanism as illustrated in Figure 2. The feature fusion block takes the encoded audio features as the query, encoded text features as the key and value to produce a feature sequence with textual information. This also ensures that the output feature sequence has the same length as the encoded audio feature sequence to feed into the mask estimator. With the residual connection from the encoded audio features, the output feature sequence is embedded with text and audio information.

In Section 3, we show that the text-informed teacher outperforms the baseline Conv-TasNet by exploiting text input.

2.2. Knowledge Distillation on Real Speech Data

While most speech enhancement models cannot be trained on real speech data due to the absence of clean speech references, our proposed framework provides a workaround by estimating the reference signal with the teacher model. Considering the scenario where transcribed real speech data are available for training, the text-informed teacher model described in Section 2.1 can be used to provide "reference" signals on the real data. In this way, the real speech data can be used along with the simulated data to train a more robust speech enhancement model, i.e. the student model shown in Figure 1. To be more specific, we train the student speech enhancement model on both simulated and real speech data, whose reference signals come from the simulated ground truth x and the teacher estimation $\hat{x}_{teacher}$, respectively. In this paper, we adopt the SI-SNR

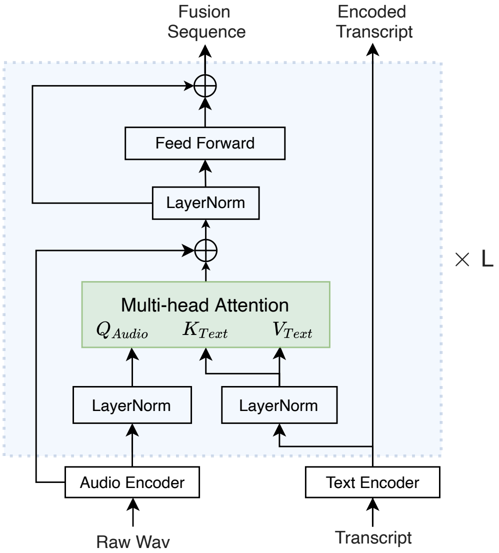

Figure 2: *Structure of the feature fusion block. The attention-based feature fusion module is repeated L times.*

between the enhanced output $\hat{\mathbf{x}}_{\text{student}}$ and the reference signal for loss calculation:

$$\text{loss}_{\text{student}} = \begin{cases} \text{SI-SNR}\left(\mathbf{x}, \hat{\mathbf{x}}_{\text{student}}\right), & \text{for simu data}, \\ \text{SI-SNR}\left(\hat{\mathbf{x}}_{\text{teacher}}, \hat{\mathbf{x}}_{\text{student}}\right), & \text{for real data}. \end{cases} \quad (1)$$

With the above training procedure, the student model learns to adapt to both simulated and realistic conditions, mitigating the mismatch between training and evaluation.

We show in Section 3 that the student model achieves better performance on both speech enhancement and ASR by applying the above knowledge distillation framework.

3. Experiments

3.1. Dataset

To evaluate the proposed framework, we conduct experiments on the single-channel track of the CHiME-4 datasets [4], which contains both simulated and recorded real data. The numbers of simulated samples in the training (`tr05_simu`), development (`dt05_simu`), and evaluation (`et05_simu`) sets are 42828, 1640, and 1320, respectively. The numbers of real recordings in the training (`tr05_real`), development (`dt05_real`), and evaluation (`et05_real`) sets are 9600, 1640, and 1320, respectively. For the ASR model, we further include the clean training data from the Wall Street Journal (WSJ) corpus [30] for training, which contains 37416 reading speech samples. The sample rate of all speech data is 16 kHz.

Speed perturbation with factors of 0.9, 1.0, and 1.1 is applied in both SE and ASR model training, as this technique has been shown effective in both tasks [31, 7]. All models are built based on the ESPnet toolkit [7]. The Adam optimizer is used in both speech enhancement and ASR model training.

3.2. Experiment Setup

3.2.1. SE Setup

The teacher and student models adopt the same setting for Conv-TasNet. We use 256 filters for 1D convolutional blocks in the audio encoder, each covering a length of 20 samples. For the mask estimator, we use 4 convolutional blocks, each consisting of 8 convolutional layers with 512 channels. The bottleneck layer has 256 channels for the 1×1-conv block.

The text encoder in the teacher model contains 4 transformer blocks with 4 heads of 64-dimensional self-attention layer in each block. Here, we adopt characters as the modeling unit for the input text. The attention-based feature fusion module is repeated $L = 6$ times, each with 4 heads of 64-dimensional self-attention layer.

Both teacher and student models are trained on 4 GPUs (each with 32gb memory) for 30 epochs with effective batch size 32. Initial learning rate of the Adam optimizer is set to 1e-3. Weight decay is set to 1e-7 for regularization. The parameters of 5 best checkpoints in terms of si-snr are averaged to get an ensemble model for inference.

3.2.2. ASR Setup

To validate the effectiveness of the SE models on real data, we evaluate the pretrained SE models on the downstream ASR task. The joint connectionist temporal classification (CTC)/attention based encoder-decoder network [32] is used as the ASR backend to evaluate the SE frontend. The loss function of joint CTC-attention network is defined as the weighted sum of CTC and S2S objective loss:

$$\mathcal{L}_{jca} = \lambda\mathcal{L}_{ctc} + (1-\lambda)\mathcal{L}_{s2s} \quad (2)$$

The input feature for ASR is the 80-dimensional log-Mel filterbank coefficients. The window length and hop length for feature extraction are 25 ms and 10 ms, respectively. The SpecAugment technique [33] is applied during training. We use 12 and 6 transformer layers with 2048 hidden units for the encoder and decoder, respectively. Each layer is a transformer block with 4 heads of 64-dimensional self-attention layers. For multitask learning (MTL), the weights for CTC and attention losses are set to 0.3 and 0.7, respectively (i.e. $\lambda = 0.3$ in Eq 2). An external character-based RNN language model is used for rescoring in the decoding stage. The ASR model is trained on both original and enhanced training data to mitigate the distortion caused by the enhancement model.

3.3. Performance Evaluation

3.3.1. Evaluation on speech enhancement

We first evaluate the speech enhancement models on the simulated data in CHiME4. To evaluate the effectiveness of the knowledge distillation framework, the baseline adopts the same structure as the Conv-TasNet student model and is trained on only simulated data.

Table 1 presents the speech enhancement performance of the proposed methods, including the signal-to-distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ) [34]. The teacher model, which is also trained on only simulated data, benefits from the text information and shows a consistent improvement over the baseline model. The student model, trained with the aforementioned knowledge distillation framework, also outperforms the baseline model.

Table 1: *Speech enhancement performance (PESQ/SDR [dB]) on the CHiME-4 single-channel track.*

SE Model	Dev (Simu)		Test (Simu)	
	PESQ	**SDR**	**PESQ**	**SDR**
Noisy Input	2.17	5.78	2.18	7.54
Conv-TasNet Baseline	2.59	11.14	2.50	11.79
Text-Informed Teacher	2.71	12.64	2.74	13.49
Conv-TasNet Student	2.63	11.97	2.66	13.47

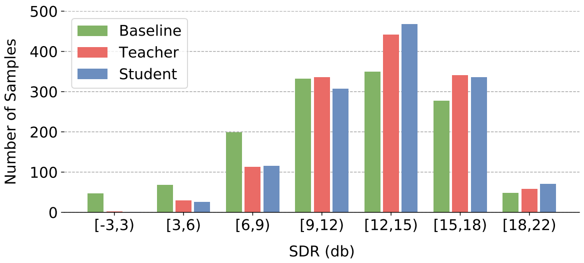

Figure 3: *SDR distribution of enhanced signals*

Figure 3 illustrates the SDR distribution of enhanced audios from different models. Both teacher and student models significantly reduce the amount of low-SDR samples, showing consistent conclusions with the results in Table 1.

3.3.2. Evaluation with clean-condition trained ASR

For ASR evaluation, we first perform an evaluation with the ASR model trained only on WSJ clean data, which is referred to as clean-condition trained ASR. Table 2 presents the evaluation result with clean-condition trained ASR. Speech recognition on Conv-Tasnet baseline enhanced signals shows consistent improvement over the unprocessed noisy signals. And the Conv-Tasnet student trained with text-informed knowledge distillation framework leads to further improvement on all development and test sets [1].

Table 2: *WER (%) on the CHiME-4 simulation and real data with clean-condition trained ASR model.*

SE	Dev		Test	
	real	**simu**	**real**	**simu**
No Process	39.8	44.0	62.5	55.3
Conv-TasNet Baseline	23.1	25.1	43.4	40.6
Conv-TasNet Student	**22.9**	**24.2**	**38.9**	**37.7**

3.3.3. Evaluation with multi-style trained ASR

Furthermore, we perform an evaluation with the ASR model trained on both WSJ clean data and CHiME4 noisy data, which is referred to as multi-style trained ASR.

The evaluation results with multi-style trained ASR are shown in Table 3. The first row shows the official baseline performance of the CHiME-4 challenge using the same ASR model structure. The Conv-TasNet baseline model in the third line shows degraded performance on test sets compared with unprocessed signals. This can be attributed to the distortion introduced by the Conv-TasNet model and the mismatch between training and evaluation conditions. Similar phenomena are also observed in some prior work [6, 7]. In the last line, the performance degradation disappeared after the text-informed knowledge distillation framework is applied. The Conv-TasNet student model shows consistent improvements over the unprocessed baseline.

Table 3: *WER (%) on the CHiME-4 simulation and real data with multi-style trained ASR model*

SE	Dev		Test	
	real	**simu**	**real**	**simu**
Baseline [4]	11.6	13.0	23.7	20.8
No Process	10.8	12.9	19.5	20.1
Conv-TasNet Baseline	10.8	12.3	23.4	23.9
Conv-TasNet Student	**9.7**	**11.6**	**18.8**	**19.4**

4. Conclusions

In this paper, we propose a text-informed knowledge distillation framework to utilize the transcribed real speech data in the training of speech enhancement models. We demonstrate that a Conv-TasNet based teacher model equipped with an extra audio-text fusion module can improve the quality of enhanced speech. Moreover, a student model can benefit from training on both simulated speech data with ground truth references and on real speech data with teacher estimated references. Evaluation on the student model shows 0.83 dB and 1.68 dB absolute SDR gains over the Conv-TasNet baseline on CHiME-4 simulated development and test sets, respectively. A consistent performance gain in terms of WER reduction is also observed on the downstream speech recognition task.

5. Acknowledgements

This work was supported in part by the China NSFC projects (Grant No. 62122050 and No. 62071288), and in part by Shanghai Municipal Science and Technology Major Project (Grant No. 2021SHZDZX0102).

[1] Since it's not reasonable to be text-informed in speech recognition task, teacher model is not evaluated in this experiment.

6. References

[1] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-attention dense U-Net for multichannel speech enhancement," in *Proc. IEEE ICASSP*, 2020, pp. 836–840.

[2] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[3] S. Zhao, T. H. Nguyen, and B. Ma, "Monaural speech enhancement with complex convolutional block attention module and joint time frequency losses," *arXiv preprint:2102.01993*, 2021.

[4] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

[5] S.-X. Wen, J. Du, and C.-H. Lee, "On generating mixing noise signals with basis functions for simulating noisy speech and learning DNN-based speech enhancement models," in *Proc. IEEE MLSP*, 2017, pp. 1–6.

[6] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *Proc. IEEE ICASSP*, 2020, pp. 7009–7013.

[7] C. Li, J. Shi, W. Zhang, A. S. Subramanian, X. Chang, N. Kamo, M. Hira, T. Hayashi, C. Boeddeker, Z. Chen, and S. Watanabe, "ESPnet-SE: end-to-end speech enhancement and separation toolkit designed for ASR integration," in *Proc. IEEE SLT*, 2021, pp. 785–792.

[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[9] J. Chen, Y. Wang, and D. Wang, "Noise perturbation for supervised speech separation," *Speech Communication*, vol. 78, pp. 1–10, 2016.

[10] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proc. Interspeech*, 2014, pp. 2670–2674.

[11] J. Lee, Y. Jung, M. Jung, and H. Kim, "Dynamic noise embedding: Noise aware training and adaptation for speech enhancement," in *Proc. APSIPA ASC*, 2020, pp. 739–746.

[12] C.-F. Liao, Y. Tsao, H.-Y. Lee, and H.-M. Wang, "Noise adaptive speech enhancement using domain adversarial training," in *Proc. Interspeech*, 2019, pp. 3148–3152.

[13] F. Deng, T. Jiang, X.-R. Wang, C. Zhang, and Y. Li, "NAAGN: Noise-aware attention-gated network for speech enhancement," in *Proc. Interspeech*, 2020, pp. 2457–2461.

[14] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[15] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, "Cycle-consistent speech enhancement," in *Proc. Interspeech*, 2018, pp. 1165–1169.

[16] T. Ochiai, S. Watanabe, and S. Katagiri, "Does speech enhancement work with end-to-end ASR objectives?: Experimental analysis of multichannel end-to-end ASR," in *Proc. IEEE MLSP*, 2017, pp. 1–6.

[17] A. S. Subramanian, X. Wang, M. K. Baskar, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "Speech enhancement using end-to-end speech recognition objectives," in *Proc. IEEE WASPAA*, 2019, pp. 234–238.

[18] S. Shon, H. Tang, and J. Glass, "VoiceID loss: Speech enhancement for speaker verification," in *Proc. Interspeech*, 2019, pp. 2888–2892.

[19] K. Žmolíková, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Černockỳ, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 800–814, 2019.

[20] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, 2019, pp. 2728–2732.

[21] Y. Koizumi, K. Yaiabe, M. Delcroix, Y. Maxuxama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. IEEE ICASSP*, 2020, pp. 181–185.

[22] K. Kinoshita, M. Delcroix, A. Ogawa, and T. Nakatani, "Text-informed speech enhancement with deep neural networks," in *Proc. Interspeech*, 2015, pp. 1760–1764.

[23] K. Schulze-Forster, C. S. Doire, G. Richard, and R. Badeau, "Joint phoneme alignment and text-informed speech separation on highly corrupted speech," in *Proc. IEEE ICASSP*, 2020, pp. 7274–7278.

[24] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech*, 2018, pp. 3244–3248.

[25] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *arXiv preprint:2008.09586*, 2020.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[27] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. Interspeech*, 2014, pp. 1910–1914.

[28] X. Hao, S. Wen, X. Su, Y. Liu, G. Gao, and X. Li, "Sub-band knowledge distillation framework for speech enhancement," in *Proc. Interspeech*, 2020, pp. 2687–2691.

[29] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[30] LDC, *LDC Catalog: CSR-I (WSJ0) Complete*, University of Pennsylvania, 1993.

[31] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. Interspeech*, 2015, pp. 3586–3589.

[32] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE ICASSP*, 2017, pp. 4835–4839.

[33] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.

[34] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE ICASSP*, vol. 2, 2001, pp. 749–752.