2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) | 979-8-3503-0689-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ASRU57964.2023.10389687

IMPROVING SPEECH ENHANCEMENT USING AUDIO TAGGING KNOWLEDGE FROM PRE-TRAINED REPRESENTATIONS AND MULTI-TASK LEARNING

Shaoxiong Lin[1], *Chao Zhang*[2], *Yanmin Qian*[1,3*]

[1] MoE Key Lab of Artificial Intelligence, AI Institute,
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Department of Electronic Engineering, Tsinghua University, Beijing, China
[3] Suzhou Institute of Artificial Intelligence, Shanghai Jiao Tong University, Suzhou 215000, China

ABSTRACT

In deep-learning-based speech enhancement (SE), an audio-knowledge-ignorant approach is often used, which estimates a denoising model to transform the noisy input speech into clean output speech without understanding the audio events that constitute the background noises. In this paper, an audio-knowledge-aware approach is proposed to improve SE, which explicitly leverages the knowledge from audio taggings to understand the background noises. Based on the recent progress in audio pattern analysis, the audio tagging knowledge is obtained using either additional input representations extracted by pre-trained audio tagging models, or from multi-task learning with extra audio event classification or regression tasks. Experimental results based on the DNS-2020 dataset and the pre-trained Wavegram-Logmel-CNN audio tagging model show that the proposed approach leads to considerable improvements in the STOI, SDR, and SI-SNR metrics.

Index Terms— Speech enhancement, pre-trained representations, audio tagging, multi-task learning

1. INTRODUCTION

Speech enhancement (SE), the task to improve the quality and intelligibility of speech signals corrupted by ambient noise, has many applications, such as automatic speech recognition, mobile communication, teleconference, and hearing aids *etc*. SE has been a prominent research direction in the field of audio signal processing for decades and remains challenging despite the enormous methods being proposed.

With the resurgence and rapid development of artificial neural networks (ANNs), deep-learning-based methods have quickly emerged as the mainstream SE approaches. These methods can be classified into time-frequency (T-F) domain and time-domain-based methods. In T-F domain methods, the input to the neural network is usually the magnitude spectrogram of the noisy speech obtained through a short-time Fourier transform. The network predicts either the clean magnitude spectrogram [1, 2] or a mask to filter the noisy magnitude spectrogram [3–5]. In contrast, the time-domain methods directly estimate the clean speech signal from the noisy speech signal without any spectral transformation [6–8].

In the existing literature [3–8], most ANN-based SE methods assume that background noise is meaningless interference and can be directly suppressed. This ignores the knowledge of the acoustic characteristics of each class of audio event that is not reserved in the output of SE, and thus is referred to as the *audio-knowledge-ignorant approach* in this paper. Although this approach allows ANNs to concentrate on the target of SE, it also deviates from the mechanism of the human auditory system since understanding any speech and non-speech sounds in a complex auditory scene is an inherent ability of human beings. In situations where multiple speakers coexist, such as a cocktail party scenario, humans possess a fundamental understanding of the characteristics of the background noise even if auditory attention is not applied to it. Research on human auditory perception strategies in noisy environments [9] further supports this notion, indicating that individuals with strong prior knowledge about the various components in the acoustic environment are better equipped to integrate and separate these components, enabling them to focus on the desired elements they wish to hear.

In this paper, by expanding on the insight of the need for understanding the acoustic characteristics of background noise, an *audio-knowledge-aware approach* is proposed for SE, wherein audio tagging knowledge is leveraged to provide the SE model with information pertaining to the background noise categories. More specifically, a latent representation encoding the information about the categories of the background noise can be extracted using an audio tagging model pre-trained on the AudioSet dataset [10], which can be fed into the SE model as an additional input source. To leverage this information more effectively, multi-task learning (MTL) can be used to include an extra training target in the SE model

*This work was supported in part by China STI 2030-Major Projects under Grant No. 2021ZD0201500, in part by China NSFC projects under Grants 62122050 and 62071288, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102, and in part by Jiangsu Technology Project (No.BE2022059-4).

that either separates the background noise from clean speech or classifies the noise according to the audio tagging. The experimental results show that the proposed method achieves an improvement of 0.38% in short-term objective intelligibility (STOI), 0.25 dB in signal-to-distortion ratio (SDR), and 0.19 dB in the scale-invariant (SI) source-to-noise ratio (SNR) on the test set with no reverberation. The increments are more considerable with reverberation, which are 2.45%, 1.92 dB, and 2.13 dB for STOI, SDR, and SI-SNR respectively.

The remainder of the paper is organised as follows: Section 2 provides the background knowledge of SE, audio tagging and multi-task learning. Section 3 describes the proposed methods. The experimental setup and results are presented in Sections 4 and 5, followed by a conclusion.

2. BACKGROUNDS

2.1. Speech enhancement with additional information

Recent SE models [11–21] have shown promising performance, yet can be further improved by incorporating additional input features extracted from audio or other modalities.

With the single audio modality, a prevalent method is to incorporate the speaker identity into the SE model [22–27], which facilitates the enhancement of the speech from a specific speaker. This technique is commonly referred to as personalised speech enhancement. Similarly, Xin *et al.* [28] demonstrate the usefulness of speaker gender information in improving SE performance. A pre-trained sound event detection model is used to extract the embeddings representing the gender of the speaker. Attention scores are computed between the noisy speech and the embeddings, and higher attention scores are used when combined with the noisy speech. Li *et al.* [29] introduces a technique called "noise token", which employs trainable noise templates to construct an embedding that represents the noise within the noisy speech. This embedding is then used by the SE model to enhance the performance. Additionally, there are also studies leveraging the text information [30–33] or facial [34, 35] and lip [36, 37] information from the visual modality.

2.2. Audio tagging task

Audio tagging is the task of predicting the presence or absence of sound classes within an audio clip. It is essentially a multi-label classification task. Traditional audio tagging approaches often use Gaussian mixture models, hidden Markov models, and discriminative support vector machines to address this task. In recent years, ANN models have become the prevailing paradigm in this domain. Training an ANN-based audio tagging model requires minimising the following binary cross-entropy loss function:

$$\mathcal{L} = -\sum_{n=1}^{N} \left(\mathbf{y}_n \cdot \ln f\left(\mathbf{x}_n\right) + \left(1-\mathbf{y}_n\right) \cdot \ln\left(1-f\left(\mathbf{x}_n\right)\right)\right),$$

where n refers to one of the N total number of audio clips, K is the number of audio classes, $y_n \in \{0,1\}^K$ is the target label and $f(\mathbf{x}_n) \in [0,1]^K$ is the output of the model.

This success of deep-learning-based audio tagging is not only attributed to the power of ANN models but also to the availability of the dataset, in particular AudioSet. AudioSet contains over 5,000 hours of audio recordings with 527 predefined sound classes, such as music, speech, and vehicle *etc*, which encompass the knowledge of real-world noise types. Consequently, the audio tagging knowledge can be leveraged in the SE model by either using latent representations extracted from a model pre-trained on AudioSet or using MTL with an extra audio tagging training target.

2.3. Multi-task learning

MTL is a machine learning paradigm that improves the training of a model by jointly learning multiple tasks. It offers several benefits such as improved generalisation ability, enhanced training efficiency, and the leverage of shared knowledge among the training tasks.

In MTL, the overall loss function is typically a linear combination of the individual task-specific losses:

$$\mathcal{L}_{\mathrm{MTL}} = \sum\nolimits_{i=1}^{T} \alpha_i \cdot \mathcal{L}_{\mathrm{task}_i},$$

where T is the number of tasks and the weight α_i is used to modulate the relative importance of task_i.

In addition to manual adjustment of α_i to control task relationships, the GradNorm method [38] provides a means of dynamically adjusting these relationships by scaling the losses with the gradient norms:

$$\mathcal{L}_{\mathrm{MTL}} = \sum_{i=1}^{T} \frac{1}{\left\|\nabla_\theta \mathcal{L}_{\mathrm{task}_i}\right\|_2} \cdot \mathcal{L}_{\mathrm{task}_i},$$

where $\left\|\nabla_\theta \mathcal{L}_{\mathrm{task}_i}\right\|_2$ represents the Frobenius norm of the gradient of $\mathcal{L}_{\mathrm{task}_i}$ with respect to the model parameters θ.

3. PROPOSED METHOD

3.1. Overview

The proposed audio-knowledge-aware SE system, as shown in Fig. 1(a), comprises a core SE model and an auxiliary audio tagging model with pre-trained parameters. During training, the noisy speech is initially processed by the audio tagging model to extract an embedding that encodes noise-type information. This embedding, along with the magnitude spectrum of the noisy speech, is then fed into the SE model to generate enhanced audio, and the loss function is computed by comparing the enhanced audio with the target clean audio.

In addition, MTL is used in training. The use of two auxiliary tasks, noise estimation and audio tagging, is investigated. The training targets for noise estimation are the noises used to synthesise noisy speech. As for the audio tagging task, training labels are pseudo-labels generated by a pre-trained audio tagging model.

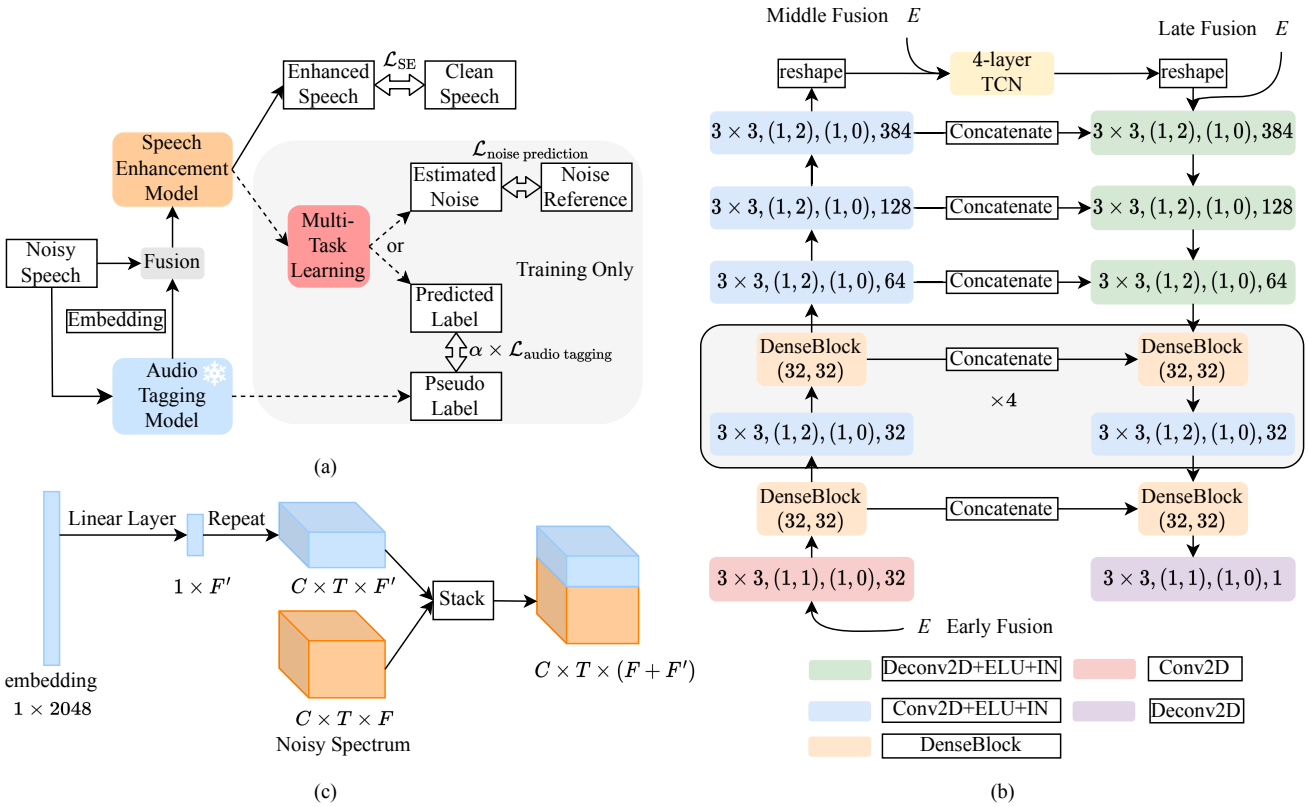

Fig. 1. (a) The overall diagram of the proposed system. The dashed lines mean the relevant connections only exist in training. The snowflake icon means the parameters of that component are frozen in training. (b) The detailed architecture of the backbone SE model and the three fusion positions of the tagging embedding are explored in this paper. (c) A simple concatenation-based fusion method is used in the system.

3.2. Audio tagging model

The pre-trained Wavegram-Logmel-CNN model [39], which achieves a mean average precision (mAP) of 0.439 on AudioSet tagging, is used as the audio tagging model. It is a 14-layer convolutional neural network (CNN) with 6 convolutional blocks. Each block includes 2 convolutional layers (3×3 kernel), with batch normalization and ReLU applied between them. Average pooling (2×2) is used for downsampling after each convolutional block. Global pooling is performed after the last convolutional layer to summarise the feature maps into a fixed-length vector. Two linear layers are then applied to the global pooling output, generating final outputs of dimensions 2048 and 527. In this paper, the output of the penultimate linear layer is referred to as the *tagging embedding*.

3.3. Speech enhancement model

The SE model in the proposed system is the TCN-DenseUNet [40], which achieved 1st place in task 1 of the L3DAS22 Challenge [41]. It takes the noisy magnitude spectrum as input and predicts the clean magnitude spectrum.

The model, shown in Fig. 1(b), is a modified U-Net architecture with a TCN network inserted between the encoder and decoder. The encoder consists of a 2D convolution layer and 7 convolutional blocks, while the decoder consists of 7 deconvolutional blocks and a 2D deconvolution layer. The TCN network has 4 layers, each containing 7 dilated convolutional blocks. The deconvolutional block in the decoder receives input from both the previous block and the corresponding convolutional block in the encoder.

3.4. Fusion of tagging embedding

In order to incorporate the audio tagging knowledge into the SE model, a simple concatenation method is used to fuse the tagging embedding with the original input feature map of the SE model, which has a shape of (C, T, F). The dimensionality of the tagging embedding is reduced to F', which is 256 in this paper, using a linear projection. Following this, the compressed tagging embedding vector is replicated along the time and channel dimensions, resulting in another tensor with a shape of (C, T, F'). Subsequently, the two tensors are concatenated, resulting in a tensor of shape $(C, T, F+F')$, which

serves as the new feature map for the SE model. Fig. 1(c) illustrates the fusion procedure described above. The tagging embedding captures information about the background noise in the entire audio clip. Concatenating the embedding with the original features frame by frame also enables each frame to have information from its contextual frames, leading to an improved SE performance. Furthermore, we examined the effect of embedding fusion at different positions in the SE system, as illustrated in Fig. 1(b).

3.5. Multi-task learning for speech enhancement

An additional output channel is incorporated into the final convolutional layer of the decoder in the SE model to facilitate the noise estimation (NE) task. Both the SE task and the noise prediction task share the same loss function and the total loss can be expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SE}} + \mathcal{L}_{\text{NE}}, \tag{1}$$

When audio tagging serves as the secondary task, the dimension representing speech probability is omitted from the labels generated by the pre-trained audio tagging model. This modification results in a 526-dimensional vector, which serves as the adjusted label, with a focus on the knowledge associated with background noise. Furthermore, to facilitate convergence, each element in the vector is normalised by dividing it by the maximum value within the vector.

To predict the audio tagging label, the output of the last DenseBlock in the TCN-DenseUNet decoder is used as input features for a prediction model. A shallow CNN followed by two linear layers is used as the predictor, which is only used in training and can be discarded in tests.

Since SE is a regression task and audio tagging is a classification task, there is an inherent need to balance between these two tasks. The two methods introduced in Section 2.3 are used to keep a balance between the SE and audio tagging (AT) tasks. Alternatively, a manual weight adjustment strategy can be used based on the following loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SE}} + \alpha \times \mathcal{L}_{\text{AT}}. \tag{2}$$

4. EXPERIMENTAL SETUP

4.1. Datasets and evaluation metrics

We synthesised 100 hours of noisy speech based on the DNS-2020 dataset, with each audio clip having a duration of 30 seconds. The SNR levels are uniformly sampled from a range of 0 to 40 dB. For training purposes, 90 hours of data are allocated, while 10 hours are reserved for validation. To facilitate the evaluation of the method's effectiveness and explore the impact of specific parameters on its performance, a subset of 36 hours of audio is randomly selected from the complete training set.

Two test sets denoted as "no_reverb" and "with_reverb" are used for evaluation, both contain 150 noisy-clean pairs. Within these two test sets, various everyday noises are included, such as fan, air conditioner, typing, and more. The audio clips in the "with_reverb" test set not only encompass background noise but also exhibit reverberation, and reverberation is not included in the training and validation sets.

Three commonly used metrics, namely STOI, SDR and SI-SNR, are used to assess the performance of SE models. Higher values indicate better performance for all metrics.

4.2. Model configurations

In Section 3.2, we provide an overview of the configuration for the audio tagging model. The detailed configuration of the SE model in the proposed system is shown in Fig. 1(b), where DenseBlocks is denoted as (g_1, g_2), and g_1 and g_2 are the growth rates. Other convolutional blocks are denoted as (k, s, p, o), where k, s, p, o are the kernel size, stride, padding, and output channels correspondingly. ELU and IN stand for exponential linear units nonlinearity layer and instance normalization layer.

In the experiments, a parameter-reduced variant of the model, referred to as TCN-DenseUNet$_{\text{small}}$, is introduced, featuring approximately 1/4 of the parameters compared to the original model. The encoder's convolutional modules utilize fewer channels (8, 8, 8, 8, 8, 16, 32, 384), and corresponding adjustments are made to the decoder. The TCN network is downscaled to 2 layers, each incorporating 4 dilated convolutional blocks. The experimental results in Section 5, unless otherwise stated, are obtained by training this model on the 36-hour training set.

4.3. Training details

- **Loss function**: For audio tagging task, the loss function is the binary cross entropy loss. For SE and noise estimation task, the loss function is the SI-SNR loss.
- **Data pre-processing**: During training, variance normalisation is applied to each noisy-clean data pair. Likewise, during the test, the noisy speech undergoes variance normalisation as well.
- **Optimisation setting**: The Adam optimizer is utilized with a learning rate of 1.0×10^{-3}. The batch size is set to 8. The maximum number of training epochs is 25, and training will halt if the loss does not decrease on the validation set for 5 consecutive epochs.

5. EXPERIMENTAL RESULTS

5.1. Comparisons with different auxiliary tasks

Table 1 presents the results with MTL. The tagging embeddings are not used in these systems. The results demonstrate that integrating audio knowledge through MTL can considerably improve the performance of the SE model.

SE with noise estimation performs better on the "no_reverb" test set compared to SE with audio tagging, possibly due to

Table 1. Performance with MTL with different auxiliary tasks. "AT" and "NE" stand for audio tagging and noise estimation respectively, and α is the weight parameter in Eqn. (2). The baseline model is the TCN-DenseUNet$_{\text{small}}$ trained on the smaller 36-hour dataset.

Model	no_reverb			with_reverb		
	STOI	SDR	SI-SNR	STOI	SDR	SI-SNR
Baseline	95.35	16.40	16.39	82.87	11.51	10.75
AT, α = 1	95.43	17.04	17.00	83.68	11.69	11.03
AT, α = 10	95.38	17.22	17.20	84.94	**12.66**	**12.00**
AT, GradNorm	95.28	16.52	16.43	**85.98**	12.48	11.85
NE	**95.56**	**17.29**	**17.27**	83.58	11.74	11.03
AT+NE	95.37	16.96	16.92	83.01	12.29	11.48

the fact that both SE and noise estimation are regression tasks and are easy to keep balanced during training, and therefore resulted in better performance. Meanwhile, SE with audio tagging outperforms SE with noise estimation on the "with_reverb" test set, showing its superior generalisation capability in handling training set mismatches. The approach of incorporating both tasks underperforms compared to selecting either task alone. This demonstrates the challenge in the simultaneous optimisation of such different tasks.

The results of using different task balance strategies are presented in Table 1. The GradNorm strategy exhibits lower overall performance compared to the manually weighted approach, as it treats both tasks equally during optimisation, compromising the more critical SE task.

In the manually weighted approach, compared to $\alpha = 10$, the performance is inferior when $\alpha = 1$. This difference can be attributed to the divergent scales of the loss functions. With $\alpha = 1$, the gradients from the SE loss dominate, prohibiting the impact of the audio tagging task. However, with $\alpha = 10$, more balanced training is achieved, resulting in substantial improvements in the final SE performance.

5.2. Comparisons with different fusion positions

Table 2 presents the SE results with different fusion positions. The MTL strategy is not used to train these systems. The terms "Early", "Middle", and "Late" in the table correspond to the three fusion positions shown in Fig. 1(b). The term "All" means to integrate the tagging embeddings at all three positions simultaneously.

The results presented in Table 2 demonstrate that integrating the tagging embedding at any of the three positions improves the SE performance. In particular, the later stage the fusion position is in the model, the smaller the improvements in SE performance on the "no_reverb" test set, whereas the improvements are more obvious on the "with_reverb" test set. A possible reason is that when the fusion position is in an earlier stage, the model is more effective in leveraging the audio knowledge from the tagging embedding, resulting in improved SE performance. However, this increases the risk of over-fitting and compromises the network's ability to generalise to unseen test data, such as reverberation. The simultaneous fusion at all three positions considerably enhances the performance in the "with_reverb" condition.

Table 2. Performance with different audio tagging embedding fusion positions. The baseline model is the TCN-DenseUNet$_{\text{small}}$ trained on the smaller 36-hour dataset.

Model	no_reverb			with_reverb		
	STOI	SDR	SI-SNR	STOI	SDR	SI-SNR
Baseline	95.35	16.40	16.39	82.87	11.51	10.75
Early	95.63	**16.89**	**16.85**	86.17	12.47	11.93
Middle	95.53	16.65	16.66	85.15	12.55	11.93
Late	95.06	16.44	16.48	85.61	12.65	12.09
All	**95.67**	16.53	16.56	**88.26**	**13.73**	**13.37**

5.3. Performance of the full-scale system

TCN-DenseUNet is used as the backbone SE model to train the full-scale system using the proposed method on the full 90-hour training set. The results are shown in Table 3. Based on preceding experiments, the audio tagging task is used as the auxiliary task in MTL, with the weight parameter α set to 10. The tagging embedding is fused at all three positions.

The proposed method is compared to Backbone and DCCRN [18], and the results show the superior performance of our proposed method on both test sets. It is worth noting that the "with_reverb" results in the last row of Table 3 suffer from a slight decline compared to the last row of Table 2. This is due to the overfitting with larger models, and the overfitted training data does not have any reverberant audio.

Table 3. Perfromance and computational burden of the full-scale systems trained on the complete 90-hour dataset. The backbone model is TCN-DenseUNet. The tagging embedding is fused at all positions, and MTL is used with the audio tagging task with $\alpha = 10$.

Model	no_reverb			with_reverb			MACs
	STOI	SDR	SI-SNR	STOI	SDR	SI-SNR	(G/s)
Noisy	91.52	9.09	9.23	86.62	9.16	9.19	—
DCCRN	96.00	17.69	17.50	81.00	10.45	9.72	**13.10**
Backbone	96.81	18.83	18.84	84.88	11.86	11.03	21.36
Proposed	**97.19**	**19.08**	**19.03**	**87.33**	**13.78**	**13.16**	22.60

6. CONCLUSIONS

In this paper, an audio-knowledge-aware speech enhancement approach is proposed, which uses the knowledge from audio tagging by fusing the embedding extracted using a pre-trained audio tagging model or using MTL. Experimental results on the DNS test sets show considerable improvements over the baselines. Future work includes exploring more efficient fusion methods and alternative loss functions.

7. REFERENCES

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[2] 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.

[3] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 708–712.

[5] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2015.

[6] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. ICASSP*, 2018, pp. 5069–5073.

[7] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1570–1584, 2018.

[8] 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.

[9] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*, MIT press, 1994.

[10] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017, pp. 776–780.

[11] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020, pp. 46–50.

[12] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-path convolution recurrent network for single channel speech enhancement," in *Proc. Interspeech*, 2021, pp. 2811–2815.

[13] Q. Hu, Z. Hou, X. Le, and J. Lu, "A light-weight full-band speech enhancement model," *arXiv preprint arXiv:2206.14524*, 2022.

[14] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-stage Transformer based neural network for speech enhancement in the time domain," in *Proc. ICASSP*, 2021, pp. 7098–7102.

[15] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI*, 2020, pp. 9458–9465.

[16] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," in *Proc. AAAI*, 2021, pp. 14549–14557.

[17] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *Proc. ICASSP*, 2022, pp. 7847–7851.

[18] 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.

[19] Y. Zhao, Z.-Q. Wang, and D.L. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 53–62, 2018.

[20] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *Proc. ICASSP*, 2021, pp. 6628–6632.

[21] M. Ge, C. Xu, L. Wang, E.S. Chng, J. Dang, and H. Li, "Multi-stage speaker extraction with utterance and frame-level reference signals," in *Proc. ICASSP*, 2021, pp. 6109–6113.

[22] R. Giri, S. Venkataramani, J.-M. Valin, U. Isik, and A. Krishnaswamy, "Personalized PercepNet: Real-Time, Low-Complexity Target Voice Separation and Enhancement," in *Proc. Interspeech*, 2021, pp. 1124–1128.

[23] S.E. Eskimez, T. Yoshioka, H. Wang, X. Wang, Z. Chen, and X. Huang, "Personalized speech enhancement: New models and comprehensive evaluation," in *Proc. ICASSP*, 2022, pp. 356–360.

[24] M. Thakker, S. E. Eskimez, T. Yoshioka, and H. Wang, "Fast Real-time Personalized Speech Enhancement: End-to-End Enhancement Network (E3Net) and Knowledge Distillation," in *Proc. Interspeech*, 2022, pp. 991–995.

[25] Y. Ju, W. Rao, X. Yan, Y. Fu, S. Lv, L. Cheng, Y. Wang, L. Xie, and S. Shang, "TEA-PSE: Tencent-Ethereal-Audio-Lab personalized speech enhancement system for ICASSP 2022 DNS-Challenge," in *Proc. ICASSP*, 2022, pp. 9291–9295.

[26] Y. Ju, S. Zhang, W. Rao, Y. Wang, T. Yu, L. Xie, and S. Shang, "TEA-PSE 2.0: Sub-band network for real-time personalized speech enhancement," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 472–479.

[27] Y. Ju, J. Chen, S. Zhang, S. He, W. Rao, W. Zhu, Y. Wang, T. Yu, and S. Shang, "TEA-PSE 3.0: Tencent-Ethereal-Audio-Lab personalized speech enhancement system for ICASSP 2023 DNS-Challenge," in *Proc. ICASSP*, 2023, pp. 1–2.

[28] Y. Xin, X. Peng, and Y. Lu, "Improving speech enhancement via event-based query," in *Proc. ICASSP*, 2023, pp. 1–5.

[29] H. Li and J. Yamagishi, "Noise Tokens: Learning Neural Noise Templates for Environment-Aware Speech Enhancement," in *Proc. Interspeech*, 2020, pp. 2452–2456.

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

[31] W. Wang, W. Zhang, S. Lin, and Y. Qian, "Text-informed knowledge distillation for robust speech enhancement and recognition," in *Proc. ISCSLP*, 2022, pp. 334–338.

[32] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "SoundBeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 121–136, 2022.

[33] C. Li, Y. Qian, Z. Chen, D. Wang, T. Yoshioka, S. Liu, Y. Qian, and M. Zeng, "Target Sound Extraction with Variable Cross-Modality Clues," in *Proc. ICASSP*, 2023, pp. 1–5.

[34] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W.-T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 112:1–112:11, 2018.

[35] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg, "Seeing through noise: Visually driven speaker separation and enhancement," in *Proc. ICASSP*, 2018, pp. 3051–3055.

[36] A. Gabbay, A. Shamir, and S. Peleg, "Visual Speech Enhancement," in *Proc. Interspeech*, 2018, pp. 1170–1174.

[37] Rui Lu, Zhiyao Duan, and Changshui Zhang, "Audio–Visual Deep Clustering for Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1697–1712, 2019.

[38] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. ICML*, 2018, pp. 794–803.

[39] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, w. Wang, and M.D. Plumbley, "PANNS: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[40] Z.-Q. Wang, G. Wichern, and J.L. Roux, "Leveraging low-distortion target estimates for improved speech enhancement," *arXiv preprint arXiv:2110.00570*, 2021.

[41] Y.-J. Lu, S. Cornell, X. Chang, W. Zhang, C. Li, Z. Ni, Z.-Q. Wang, and S. Watanabe, "Towards low-distortion multi-channel speech enhancement: The ESPnet-SE submission to the L3DAS22 challenge," in *Proc. ICASSP*, 2022, pp. 9201–9205.