Two-stage Framework for Robust Speech Emotion Recognition Using Target Speaker Extraction in Human Speech Noise Conditions

Jinyi Mi, Xiaohan Shi, Ding Ma, Jiajun He, Takuya Fujimura and Tomoki Toda
Nagoya University, Japan
E-mail: {mi.jinyi, xiaohan.shi, ding.ma, jiajun.he, fujimura.takuya}@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—Developing a robust speech emotion recognition (SER) system in noisy conditions faces challenges posed by different noise properties. Most previous studies have not considered the impact of human speech noise, thus limiting the application scope of SER. In this paper, we propose a novel two-stage framework for the problem by cascading target speaker extraction (TSE) method and SER. We first train a TSE model to extract the speech of target speaker from a mixture. Then, in the second stage, we utilize the extracted speech for SER training. Additionally, we explore a joint training of TSE and SER models in the second stage. Our developed system achieves a 14.33% improvement in unweighted accuracy (UA) compared to a baseline without using TSE method, demonstrating the effectiveness of our framework in mitigating the impact of human speech noise. Moreover, we conduct experiments considering speaker gender, showing that our framework performs particularly well in different-gender mixture.**

I. Introduction

Speech is a significant part of human communication. Besides linguistic information, it contains unique paralinguistic information such as gender, emotion, and age, which is essential to the normal communication. In certain instances, misunderstanding paralinguistic features would distort the correct information conveyed by speech, leading to an ineffective communication. Therefore, it is necessary to develop human-like communication machines that can comprehend paralinguistic data.

Speech emotion recognition (SER), as a branch of affective computing, has garnered growing attention over the past two decades because of its contribution to human-computer interactions [1], [2]. Generally, the mechanism of SER involves extracting and classifying effective emotional features from audio signals so that various emotions of a speaker can be captured, thanks to which SER has been applied in healthcare [3], [4], driver safety [5], [6], call center [7], [8], and online education [9], [10]. At present, research on SER systems on scenarios devoid of background noises, often referred to as clean scenarios, has shown good performance [11]–[13]. However, in real-world environments, the performance of SER significantly degrades, mainly due to the presence of various noises from different sources. These unknown noises severely affect the performance of SER systems, which poses major challenges for the widespread application of SER systems.

Several studies have focused on SER tasks in the environment affected by specific noise sources, including communication systems [14], [15], transportation [15], [16], and industrial activities [15]. In [14], Huang et al. studied SER from speech signals with additive white Gaussian noise, they proposed two speech enhancement methods based on spectral subtraction and masking properties, respectively. In [16], Chenchah et al. used power-normalized cepstral coefficients as acoustic features for improving the robustness of SER systems in noisy environment from cars and trains. In [15], Liu et al. proposed a multi-level knowledge distillation framework, which significantly reduced the affects of noises from channel, car, and factory.

However, the aforementioned studies mainly concentrate on addressing the noise sources associated with non-human activities, leaving a gap in addressing prevalent sources of noise in human-centric environments. While Shi et al. [17] did adopt ASR representations to filter out a specific category of noise related to human activities, which is typically from human physical actions like knocking on doors, a more common category of noise stemming from human speech itself remains underexplored. This type of noise called human speech noise, which is common in activities involving human interactions such as social gatherings, often becomes entangled with target speech data, forming a complex acoustic environment. Therefore, human speech noise becomes more unpredictable and more challenging to address.

On the other hand, humans have an extraordinary ability to selectively concentrate on a single speaker among a complex acoustic environment, commonly called cocktail party effect [18]. To replicate this specialized listening ability in machines, target speaker extraction (TSE) technique has been developed. This technique exploits an auxiliary information of the target speaker and extracts speech of that speaker from the mixture. Švec et al. [19] explored the potential of TSE for extracting target emotional speech. In light of this, we propose a novel two-stage framework by cascading TSE method and SER to mitigate the impact of human speech noise. In the experiments, we utilize the TSE–SER framework to ShiftCNN [12] that is a state-of-the-art SER model, and compare its performance to verify the effectiveness of our framework. Furthermore, we investigate different factors on the performance of SER, includ-

ing training methods and speaker genders. Our contributions to this work are as follows:

- We apply the TSE method on SER tasks to address the practical scenario in which target speech is interfered by human speech noises. To our knowledge, this study is the pioneering effort in exploring the integration of TSE with SER.
- We propose a two-stage framework using different training methods. According to comparative experiments against the baselines without using TSE technique, our framework significantly improves the accuracy of SER in human speech noise conditions.
- We investigate the impact of human speech noise on our framework, especially on same-gender mixture and different-gender mixture. Results indicate that our framework performs better in different-gender mixture.

II. Proposed Method

A. Framework overview

We assume that an observed time-domain mixture signal in human speech noise conditions is defined as

$$\boldsymbol{y} = \boldsymbol{s}_0 + \sum_{i=1}^{I-1} \boldsymbol{s}_i, \tag{1}$$

where I is the number of speakers in the mixture, $\boldsymbol{s}_i$ for $i = 0, ..., I-1$ is the speech signal of the ith speaker. In particular, $i = 0$ indicates the speech signal of the target speaker. Directly using the mixture signal $\boldsymbol{y}$ corrupted by overlapping speakers for SER task, would cause the poor SER performance, as the SER model cannot selectively focus on a single speaker. Therefore, the goal of the proposed framework is to extract the target speaker signal $\boldsymbol{s}_0$ from the mixture signal $\boldsymbol{y}$ for SER training.

As illustrated in Figure 1, the framework contains two stages: First, the TSE model is trained using a large-scale mixed-speech corpus. This ensures a well-trained TSE model that can extract high-quality speech of the target speaker given the input speech and the enrollment utterance of that speaker. Note that we use a neutral speech utterance for enrollment, which is a more convenient and realistic setting than using corresponding emotional speech. Then, we apply the TSE model as the form of data augmentation to extract the exclusive speech information of the target speaker from a mixed emotional speech corpus. This denoised corpus is used for both training and testing in SER tasks. We denote this training method as TSE-SER-base. In addition, we further propose another training method called TSE-SER-ft shown in Figure 2. We start from the same TSE pretraining as the first stage. In the second stage, we introduce mixed emotional speech as input, simultaneously fine-tuning the pretrained TSE model and training the SER model. This joint training process not only refines the TSE system by adjusting its parameters but also benefits the SER training.

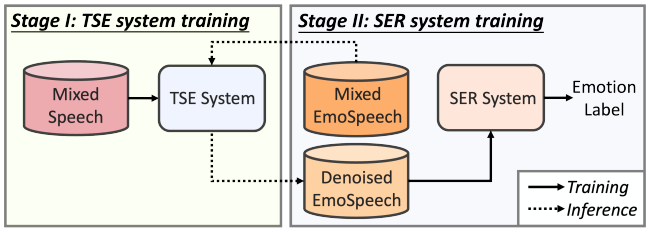

Fig. 1. Two-stage framework with TSE-SER-base.

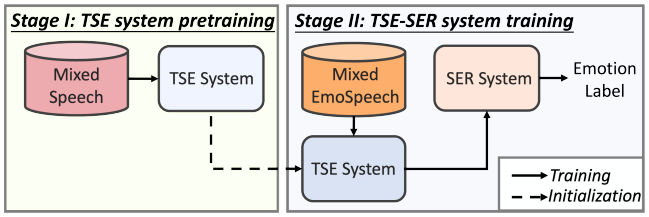

Fig. 2. Two-stage framework with TSE-SER-ft.

B. Target speaker extraction model

TSE refers to the task of reconstructing the speech signal of the target speaker from the mixture given auxiliary information of that speaker. This process can be formulated as

$$\hat{\boldsymbol{s}}_0 = g(\boldsymbol{y}, \boldsymbol{a}_0), \tag{2}$$

where $\hat{\boldsymbol{s}}_0$ is the estimated speech of the target speaker, $\boldsymbol{a}_0$ is the enrollment utterance of the target speaker, g represents the transformation carried out by the TSE system.

In this work, we adopt time-domain SpeakerBeam (TD-SpeakerBeam) [20], [21] as the TSE model. TD-SpeakerBeam consists of an auxiliary network and a speech extraction network, represented by h and f, respectively. The auxiliary network h accepts the enrollment utterance $\boldsymbol{a}_0$ and computes an embedding vector, denoted by $\boldsymbol{E}_0$, to represent the acoustic characteristics of the target speaker, i.e.,

$$\boldsymbol{E}_0 = h(\boldsymbol{a}_0). \tag{3}$$

Then, the speech extraction network f accepts the mixture signal $\boldsymbol{y}$ and the embedding vector $\boldsymbol{E}_0$ of the target speaker as inputs to predict the speech signal of the target speaker, i.e.,

$$\hat{\boldsymbol{s}}_0 = f(\boldsymbol{y}, \boldsymbol{E}_0), \tag{4}$$

where f comprises an encoder $\mathcal{E}$, a mask estimator $\mathcal{B}$, and a decoder $\mathcal{D}$. This process is formulated as

$$\boldsymbol{Y} = \mathcal{E}(\boldsymbol{y}), \tag{5}$$

$$\boldsymbol{M}_0 = \mathcal{B}(\boldsymbol{Y}, \boldsymbol{E}_0), \tag{6}$$

$$\hat{\boldsymbol{s}}_0 = \mathcal{D}(\boldsymbol{Y} \odot \boldsymbol{M}_0), \tag{7}$$

where $\odot$ denotes element-wise multiplication. The mixture signal $\boldsymbol{y}$ is fed into the encoder $\mathcal{E}$ that is represented by a 1D convolution layer. Then, the mask estimator $\mathcal{B}$ maps the output $\boldsymbol{Y}$ of the encoder $\mathcal{E}$ to a mask $\boldsymbol{M}_0$ for the target speaker, utilizing multiple convolution blocks. In particular, a

multiplicative adaptation layer, accepting the embedding vector $\boldsymbol{E}_0$ of the target speaker as auxiliary information, is inserted between the first and second blocks to drive the network towards extracting the target speaker. Finally, the mask $\boldsymbol{M}_0$ and the encoder features $\boldsymbol{Y}$ are fed into a 1D deconvolution layer-based decoder $\mathcal{D}$, to output the time-domain signal of the target speaker.

C. Loss functions

In the first stage, TSE-SER-base and TSE-SER-ft use scale-invariant source-to-noise ratio (SiSNR) [22] as the loss for TSE training. In the second stage, TSE-SER-base uses cross entropy (CE) loss for SER training, whereas TSE-SER-ft jointly trains the pretrained TSE model and SER model, considering both SiSNR and CE losses. The second stage losses of TSE-SER-base (L_{base}) and TSE-SER-ft (L_{ft}) are represented as

$$L_{base} = L_{CE}, \tag{8}$$

$$L_{ft} = L_{SiSNR} + L_{CE}. \tag{9}$$

III. Experimental evaluations

A. Datasets

In this work, we designed two kinds of datasets for comparable experiments: 1) the clean emotional dataset, and 2) the emotional dataset mixed with human speech noise. For the latter, two different human speech datasets were used as noise. All the mentioned datasets were sampled at 16 kHz.

IEMOCAP: The Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus [23], consisting of approximately 12 hours of recordings, includes five dyadic sessions, each with one English male speaker and one English female speaker. For our experiments, we used IEMOCAP as the clean emotional dataset, where we considered only four emotional categories of happy, angry, sad, and neutral. Note that "excited" was merged with "happy" to ensure category balance [13], [24]–[27].

LibriSpeech: The LibriSpeech corpus [28] contains about 1000 hours of read English speech. For our experiments, 105 hours of this corpus were chosen as a source of human speech noise.

ESD: The Emotional Speech Database (ESD) corpus [29] comprises about 29 hours of recordings from 10 English speakers and 10 Chinese speakers. For our experiments, we considered only the English part as another source of human speech noise.

In order to more accurately evaluate the performance of the proposed framework, we adopted leave-one-session-out 5-fold cross-validation to test all the models. Note that we used the following terms to represent different datasets designed: *Clean* means a clean set, *Noisy* means a dataset mixed with human speech noise, and *Denoised* indicates a dataset denoised from *Noisy* by the pretrained TSE model.

B. Experimental procedure

The first experiment investigated the impact of human speech noise on SER (see Section III-D1). The noisy dataset was generated by randomly selecting utterances from different speakers in LibriSpeech and mixing them with clean data of IEMOCAP. The number of speakers in the mixture of speech was limited to two.

The second and third experiments explored the effect of TSE method on SER and the proposed training methods on our framework (see Sections III-D2 and III-D3). We used the same dataset as the first experiment for the second stage of TSE-SER-base and TSE-SER-ft. We adopted LibriMix [30] where 100 hours of LibriSpeech were additionally used to generate mixtures for TSE pretraining. Furthermore, for the target speaker in the mixture, we randomly chose one neutral utterance of this speaker that does not belong to the mixture as the enrollment utterance.

The fourth experiment explored the impact of gender states of mixtures on SER (see Section III-D4). We used the same clean speech from the first experiment and used ESD as noise. We generated two types of mixtures: same-gender mixture and different-gender mixture, where the first type was stipulated that two speakers have the same genders, while the latter type required two speakers of opposite genders.

TABLE I
DEFINITION OF THE MODELS USED IN EXPERIMENTS.

Short Name	Model	Method
SB	TD-SpeakerBeam [21]	-
SC	ShiftCNN [12]	-
SB_SC	TD-SpeakerBeam + ShiftCNN	TSE-SER-base
SB+SC	TD-SpeakerBeam + ShiftCNN	TSE-SER-ft

C. Implementation and metrics

To build TSE model, we followed an open-source SpeakerBeam implementation[1]. For SER model, we used ShiftCNN [12], which has shown advanced performance in clean environments, adopting the same hyperparameters as [12]. All implemented models in experiments are shown in Table I.

For evaluation metrics of SER, we used the unweighted accuracy (UA) and weighted accuracy (WA). UA was the mean of the accuracies for each individual class while WA represented the ratio of correctly predicted samples to the total number of samples. In addition, we used scale-invariant signal-to-distortion ratio (SI-SDR) and scale-invariant signal-to-distortion ratio improvement (SI-SDRi) to evaluate the performance of TSE model.

D. Experimental results

To conduct a comparative study of all the experiments, aside from the proposed TSE-SER-base and TSE-SER-ft, we built two typical SER baselines using ShiftCNN, which were directly trained on *Clean* and *Noisy*, referred to as clean SER model and noisy SER model, respectively.

[1]https://github.com/BUTSpeechFIT/speakerbeam

TABLE II
PERFORMANCE OF SER IN CLEAN AND HUMAN SPEECH NOISE CONDITIONS.

Model	Train Set	Test Set	WA (%)	UA (%)
SC	Clean	Clean	**70.20**	**71.64**
	Clean	Noisy	47.11	46.02
	Noisy	Noisy	53.71	53.94

TABLE III
COMPARISON WITH TSE-SER-BASE AND THE SER BASELINES.

Model	Train Set	Test Set	WA (%)	UA (%)
SC	Noisy	Noisy	53.71	53.94
	Clean	Denoised	51.12	49.50
SB_SC	Denoised	Denoised	**62.20**	**63.42**

1) The impact of human speech noise on SER: To clarify the impact of human speech noise on SER, we compare the clean SER model and the noisy SER model. As shown in Table II, the clean SER model obtains an accuracy of over 70% for both UA and WA in the clean test set. But their performance drops significantly on the noisy test data, with a maximum decrease of up to 23.09% and 25.62% in terms of WA and UA. These results demonstrate that the clean SER model is fragile against human speech noise. Furthermore, when the SER model uses noisy data for training, the relatively better performance can be observed. Nonetheless, the results are still significantly worse than those of the clean SER model on the clean test set, suggesting deficient robustness. We argue that human speech noise severely hinders the SER model from establishing an effective mapping to the target emotional speech with the direct training.

2) The effect of TSE method on SER: To verify the effectiveness of the proposed framework on SER, we compare the performance of the SER model using the TSE method with those of the SER baselines. As presented in Table III, we first observe that the results of the clean SER model on denoised test set are not ideal, demonstrating that the clean model is unable to adapt to the denoised speech with distorted properties. We hence design the corresponding system trained on denoised data using TSE-SER-base, referred to as SB_SC in Table III. Our proposed system SB_SC performs significantly better than the other systems. Especially, the UA reaches 63.42%, which is a 9.48% increase compared to the noisy SER model in noisy conditions. Meanwhile, for the unbalanced training data from IEMOCAP corpus, the WA result is also competitive, closely resembling the UA results. The overall results demonstrate that the proposed framework with TSE-SER-base adapts well to human speech noise, showcasing effectiveness and robustness.

TABLE IV
COMPARISON WITH TSE-SER-BASE AND TSE-SER-FT.

Model	Method	WA (%)	UA (%)
SB_SC	TSE-SER-base	62.20	63.42
SB+SC	TSE-SER-ft	**67.02**	**68.27**

TABLE V
COMPARISON WITH TSE-SER-BASE AND THE SER BASELINES ON SAME- AND DIFFERENT-GENDER MIXTURES.

Model	Train Set	Test Set	Gender State	WA (%)	UA (%)
SC	Clean	Noisy	Same	48.00	46.50
	Noisy	Noisy		54.67	54.84
	Clean	Denoised		45.09	43.43
SB_SC	Denoised	Denoised	Same	**55.09**	**55.95**
SC	Clean	Noisy	Different	46.62	44.60
	Noisy	Noisy		54.29	54.20
	Clean	Denoised		48.87	47.35
SB_SC	Denoised	Denoised	Different	**59.75**	**61.32**

3) The effect of training methods on TSE–SER framework: We compare the performance of TSE-SER-base and TSE-SER-ft. As indicated in Table IV, TSE-SER-ft significantly outperforms TSE-SER-base. Moreover, the UA of TSE-SER-ft reaches a 14.33% improvement compared with the noisy SER model presented in Table III. The SI-SDR of the noisy speech before being processed by the TSE model is 0.09 dB. We calculate the SI-SDRi for the corresponding TSE models of TSE-SER-base and TSE-SER-ft to be 7.68 dB and 12.90 dB, respectively, verifying that TSE-SER-ft can improve TSE performance. Therefore, the TSE fine-tuning, applied to the noisy dataset containing the task-specific emotional data, enables the TSE model to extract purer emotional-related acoustic features to benefit the SER training.

4) The impact of gender states of mixtures on SER: Table V shows the results of TSE-SER-base, the SER baselines on same- and different-gender mixtures. First, the clean SER models unsurprisingly gain the lowest performance. Moreover, the noisy SER model shows a non-obvious trend on the noisy test set across both gender states, whereas the clean SER model in same-gender mixture clearly outperforms that in different-gender mixture, demonstrating better adaptability of SER models to same-gender mixture. In addition, besides the results of TSE-SER-base being all better than those of other systems for both same- and different-gender mixtures, we can see a significant performance gap of TSE-SER-base for dealing with same- and different-gender mixtures, which is opposite to the finding for the clean SER model in noisy conditions. Since our framework uses the TSE model, we conjecture that same-

TABLE VI
ABLATION STUDY FOR TSE-SER-BASE. "SAME" AND "DIFFERENT" INDICATE GENDER STATES.

Model	Train set	Test set	WA (%)	UA (%)
SB_SC	Denoised (Same)	Denoised (Same)	55.09	55.95
		Denoised (Different)	59.12	60.46
	Denoised (Different)	Denoised (Same)	55.32	56.40
		Denoised (Different)	59.75	61.32

gender mixture with the similar acoustic characteristics is more difficult for TSE model to separate. Before being processed by the TSE model, the SI-SDR of same- and different-gender mixtures are 0 dB and 0.02 dB. We calculate the SI-SDRi for TSE model to be 1.09 dB and 5.22 dB for same- and different-gender mixtures, respectively. This also explains why the clean SER model on the denoised test set gives better results in different-gender mixture.

We further conduct an ablation study for TSE-SER-base using two SB_SC systems from Table V on test sets denoised from the same- and different-gender mixtures. We have an interesting finding in Table VI that using training data denoised from different-gender mixture can enhance our model performance on both test sets denoised from same- and different-gender mixtures. We argue that the pretrained TSE model can extract higher-quality denoised data from different-gender mixture, thus finalizing a higher-performance SER model.

IV. CONCLUSION

In this work, we presented a two-stage framework to mitigate the impact of human speech noise on SER. Based on the framework, we designed two training methods, TSE-SER-base and TSE-SER-ft. The effectiveness and robustness of both methods have been verified. Moreover, we investigated the impact of human speech noise on SER, especially on same- and different-gender mixtures. In the future, we plan to explore how our proposed framework performs on human speech noise with different attributes, such as emotion classes and languages. Another possible direction involves multi-interference environments, such as noisy and reverberant environments.

ACKNOWLEDGMENT

This work was partly supported by JST CREST Grant Number JPMJCR19A3, Japan, and JSPS KAKENHI Grant Number 21H05054. In addition, this work was also financially supported by JST SPRING, Grant Number JPMJSP2125. The author would like to take this opportunity to thank the "THERS Make New Standards Program for the Next Generation Researchers."

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] X. Shi, S. Li, and J. Dang, "Dimensional emotion prediction based on interactive context in conversation.," in *INTERSPEECH*, 2020, pp. 4193–4197.

[3] M. Z. Uddin and E. G. Nilsson, "Emotion recognition using speech and neural structured learning to facilitate edge intelligence," *Engineering Applications of Artificial Intelligence*, vol. 94, p. 103 775, 2020.

[4] M. S. Hossain and G. Muhammad, "Cloud-assisted speech and face recognition framework for health monitoring," *Mobile Networks and Applications*, vol. 20, pp. 391–399, 2015.

[5] M. Grimm, K. Kroschel, H. Harris, *et al.*, "On the necessity and feasibility of detecting a driver's emotional state while driving," in *International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 126–138.

[6] A. Tawari and M. Trivedi, "Speech based emotion classification framework for driver assistance system," in *2010 IEEE Intelligent Vehicles Symposium*, 2010, pp. 174–178.

[7] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proceedings of artificial neural networks in engineering*, vol. 710, 1999, p. 22.

[8] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech communication*, vol. 49, no. 2, pp. 98–112, 2007.

[9] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in e-learning system based on affective computing," in *Third international conference on natural computation (ICNC 2007)*, vol. 5, 2007, pp. 809–813.

[10] A. Tickle, S. Raghu, and M. Elshaw, "Emotional recognition from the speech signal for a virtual education agent," in *Journal of Physics: Conference Series*, vol. 450, 2013, p. 012 053.

[11] Y. Liu, H. Sun, W. Guan, *et al.*, "A discriminative feature representation method based on cascaded attention network with adversarial strategy for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1063–1074, 2023.

[12] S. Shen, F. Liu, and A. Zhou, "Mingling or misalignment? temporal shift for speech emotion recognition with pre-trained representations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] L. W. Chen and A. Rudnicky, "Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[14] C. Huang, C. Guoming, Y. Hua, B. Yongqiang, and Z. Li, "Speech emotion recognition under white noise," *Archives of Acoustics*, vol. 38, no. 4, pp. 457–463, 2013.

[15] Y. Liu, H. Sun, G. Chen, *et al.*, "Multi-level knowledge distillation for speech emotion recognition in noisy conditions," *arXiv preprint arXiv:2312.13556*, 2023.

[16] F. Chenchah and Z. Lachiri, "A bio-inspired emotion recognition system under real-life conditions," *Applied Acoustics*, vol. 115, pp. 6–14, 2017.

[17] X. Shi, J. He, X. Li, and T. Toda, "On the effectiveness of asr representations in real-world noisy speech emotion recognition," *arXiv preprint arXiv:2311.07093*, 2023.

[18] S. Handel, *Listening: An introduction to the perception of auditory events*. MIT Press, 1993.

[19] J. Švec, K. Žmolíková, M. Kocour, *et al.*, "Analysis of impact of emotions on target speech extraction and speech separation," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, IEEE, 2022, pp. 1–5.

[20] K. Žmolíková, M. Delcroix, K. Kinoshita, *et al.*, "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.

[21] M. Delcroix, T. Ochiai, K. Zmolikova, *et al.*, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 691–695.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[23] C. Busso, M. Bulut, C.-C. Lee, *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[24] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7367–7371.

[25] L. Guo, L. Wang, C. Xu, J. Dang, E. S. Chng, and H. Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 6304–6308.

[26] X. Shi, X. Li, and T. Toda, "Emotion awareness in multi-utterance turn for improving emotion prediction in multi-speaker conversation," in *Proc. Interspeech*, 2023, pp. 765–769.

[27] X. Shi, X. Li, and T. Toda, "Multimodal fusion of music theory-inspired and self-supervised representations for improved emotion recognition," pp. 2024–2350, 2024.

[28] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015, pp. 5206–5210.

[29] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[30] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.