ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9747306

END-TO-END MULTI-MODAL SPEECH RECOGNITION WITH AIR AND BONE CONDUCTED SPEECH

Junqi Chen[1,2], Mou Wang[1,3], Xiao-Lei Zhang[1,2], Zhiyong Huang[3], Susanto Rahardja[1]

[1] CIAIC, School of Marine Science and Technology, Northwestern Polytechnical University, China
[2] Research & Development Institute of Northwestern Polytechnical University in Shenzhen, China
[3] School of Computing, National University of Singapore & NUS Research Institute in Chongqing

ABSTRACT

Improving the performance of automatic speech recognition (ASR) in adverse acoustic environments is a long-term tough task. Although many robust ASR systems based on conventional microphones have been developed, their performance with air-conducted (AC) speech is still far from satisfactory in low signal-to-noise-ratio (SNR) environments. Bone-conducted (BC) speech is relatively insensitive to ambient noise, and has a potential of promoting the ASR performance at such low SNR environments as an auxiliary source. In this paper, we propose a conformer-based multi-modal speech recognition system. It uses a conformer encoder and a transformer-based truncated decoder to extract the semantic information from AC and BC channels respectively. The semantic information of the two channels are re-weighted and integrated by a novel multi-modal transducer. Experimental results show the effectiveness of the proposed method. For example, given a 0 dB SNR environment, it yields a character error rate of over 59.0% lower than a noise-robust baseline conducted on AC channel only, and over 12.7% lower than a multi-modal baseline that takes the concatenated features of AC and BC speech as the input.

***Index Terms*—** Robust speech recognition, bone conduction, multi-modal transducer

1. INTRODUCTION

In recent years, robust automatic speech recognition (ASR) has been developed rapidly [1, 2], such as the sequence-to-sequence modeling method based on transformer [3]. Existing works on robust ASR can be mainly divided into two categories. One class attempts to remove the noise component of speech by speech enhancement front-ends [2, 4, 5]. The other one aims to build an adaptive ASR model with a properly designed training method, which is able to learn a noise-invariant speech representation [1, 6–8]. However, the above ASR systems are limited to air-conducted (AC) speech. Because AC speech is easily contaminated by noise, the performance of the ASR systems drop significantly in lower signal-to-noise-ratio (SNR) environments, especially in the presence of non-stationary noises [9].

To address the aforementioned issue, other modalities beyond AC speech have been introduced, such as the visual modality [10–13]. Results show that, when properly utilized, multi-modal joint processing leads to better performance than the single-modal processing based on AC speech only. However, because many phonemes share similar lip movements, the semantic information in video is limited, which in turn leads to limited performance improvement.

An alternative modality to the visual cues is bone-conducted (BC) speech which is inherently immune to the noise in AC speech [14]. BC speech is recorded by a BC microphone, which is a kind of skin-attached and non-audible sensor. It converts the vibration around a speaker skull into electrical signals. Therefore, BC speech is relatively insensitive to ambient noise, which makes it possible to promote the performance of speech related systems significantly, especially at low SNR environments. For example, BC speech has been studied in multi-modal speech enhancement [14–17].

However, BC modality seems far from explored in the modern ASR research, due to maybe the following shortcomings of BC speech. Because a BC microphone is insensitive to high frequency signals, BC speech suffers significant high-frequency loss. Besides, when BC microphone records speech, it rubs with skin, which generates unwanted self-noise to BC speech. These weaknesses bring new challenges to ASR. Fortunately, BC and AC speech is complementary, which provides an opportunity to boost their merits together while suppress their weaknesses simultaneously via multi-modal ASR.

In this paper, we propose a conformer based multi-modal ASR system. It consists of a conformer encoder and a truncated decoder. It takes both AC and BC speech as the input. It uses a novel multi-modal transducer (MMT) based on a scaling sparsemax operator to fuse the embedding representations of AC and BC speech. The contribution of this paper

This work was supported in part by National Science Foundation of China under Grant No. 62176211, in part by Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality under grant No. JCYJ20210324143006016.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on August 28,2022 at 14:53:41 UTC from IEEE Xplore. Restrictions apply.

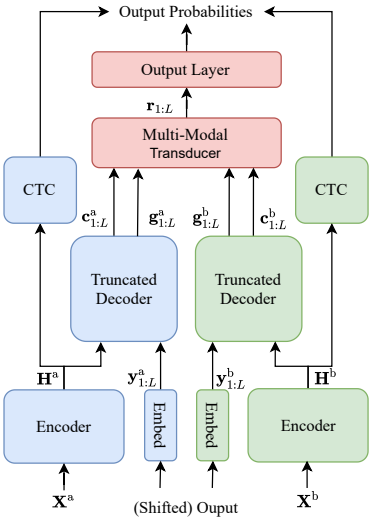

Fig. 1: The overview of the proposed system. The modules in blue color come from a pre-trained ASR system trained with AC speech. The modules in green color come form a pre-trained ASR system trained with BC speech. The modules in red color are fine-tuned with AC and BC parallel data.

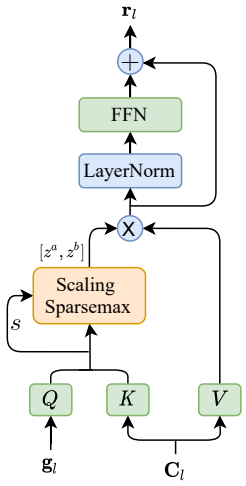

Fig. 2: The proposed multi-modal transducer module.

is summarized as follows. First, to the best of our knowledge, the proposed system is the first multi-modal end-to-end ASR work that deals with AC and BC speech jointly. Moreover, we apply the scaling sparsemax operator to the MMT module so that the module can adaptively adjust the fusion weights assigned to AC and BC channels respectively. In addition, a two-stage training method is proposed for the multi-modal ASR, where the parallel training data of AC and BC speech is used to fine-tune the MMT module only which contains only few parameters.

2. METHOD

2.1. System overview

The architecture of the proposed system is shown in Fig.1. It contains two branches, which takes AC and BC features respectively as the input. Each branch contains a conformer-based encoder, a transformer-based truncated decoder and a connectionist temporal classification (CTC) layer, which produces a context vector and a CTC-based output probability vector. Then, the proposed MMT takes the context vectors from the two branches as its input, and outputs a fused context vector. Finally, the fused context vector passes through the output layer, which produces the final attention-based output probability.

2.2. Parallel branch

Given the features $\mathbf{X}^{\mathrm{a}} \in \mathbb{R}^{T\times D_a}$ and $\mathbf{X}^{\mathrm{b}} \in \mathbb{R}^{T\times D_b}$ from AC and BC channels respectively where T denotes the number of frames and D_a and D_b denote the dimensions of the acoustic features, they first pass through a conformer-based encoder in parallel. The encoder has multiple blocks, each of which consists of two position-wise feed-forward (FFN) modules, a multi-head attention (MHA) module and a convolution module. The encoder produces a high level representation $\mathbf{H}^i \in \mathbb{R}^{\hat{T}\times D_m}$ by:

$$\mathbf{H}^i = \mathrm{Cenc}(\mathbf{X}^i), \quad \forall i \in \{a, b\} \tag{1}$$

where $\mathrm{Cenc}(\cdot)$ represents the conformer-based encoder, $\hat{T}$ represents the number of frames after down-sampling, and $i = a$ represents the encoder for AC channel while $i = b$ for BC channel. Then, given the high level representation $\mathbf{H}^i$ and the shifted output embedding vector $\mathbf{y}^i_{1:l}$, a truncated decoder extracts the context vector $\mathbf{c}^i_l$ in each time step l:

$$\mathbf{c}^i_l = \mathrm{Tdec}(\mathbf{H}^i, \mathbf{y}^i_{1:l}), \quad \forall i \in \{a, b\} \tag{2}$$

where $\mathrm{Tdec}(\cdot)$ represents the truncated decoder. It also contains multiple blocks, each of which consists of a MHA, a masked MHA and a FFN module except the last block. We removed the FFN module of the last block to retain more original semantic information. At the same time, we extract the output of the masked MHA module in the first decoder block, named the *guide vector* $\mathbf{g}^i_l \in \mathbb{R}^{D_m}$, by:

$$\mathbf{g}^i_l = \mathrm{MHA}(\mathbf{y}^i_l, \mathbf{y}^i_{1:l}, \mathbf{y}^i_{1:l}), \quad \forall i \in \{a, b\}. \tag{3}$$

Through the parallel branches, we can get a concatenated context matrix $\mathbf{C}_l = [\mathbf{c}^a_l, \mathbf{c}^b_l]^T \in \mathbb{R}^{2\times D_m}$, and a mean pooling guide vector $\mathbf{g}_l = \mathrm{Mean}(\mathbf{g}^a_l, \mathbf{g}^b_l) \in \mathbb{R}^{D_m}$.

2.3. Multi-modal transducer

Fig. 2 shows the architecture of the proposed MMT. It takes $\mathbf{C}_l$ and $\mathbf{g}_l$ as its input. It first conducts linear transformations

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on August 28,2022 at 14:53:41 UTC from IEEE Xplore. Restrictions apply.

on the context vectors produced from AC and BC channels, then uses the scaling sparsemax (SSP) [18] operator to assign weights to the two channels given the input guide vector, and finally obtains the fusion context vector $\mathbf{r}_l \in \mathbb{R}^{D_m}$:

$$\mathbf{z}_l = \mathrm{SSP}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_m}}, s), \tag{4}$$

$$\mathbf{r}_l = (\mathbf{z}_l\mathbf{V})^T + \mathrm{FFN}(\mathrm{LayerNorm}((\mathbf{z}_l\mathbf{V})^T)) \tag{5}$$

where

$$\mathbf{Q} = \mathbf{g}_l^T\mathbf{W}^Q,\ \mathbf{K} = \mathbf{C}_l\mathbf{W}^K,\ \mathbf{V} = \mathbf{C}_l\mathbf{W}^V$$

are the query, key, and value matrices respectively, $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ are learnable projection transformation matrices, $\mathrm{LayerNorm}(\cdot)$ and $\mathrm{FFN}(\cdot)$ denote the layer normalization and position-wise feedforward operation respectively, and $\mathrm{SSP}(\mathbf{x}, s)$ denotes the scaling sparsemax re-weighting operation with its scaling factor s computed as follows:

$$s = 1 + \mathrm{ReLU}(\mathrm{Linear}(||\mathbf{x}||, N)) \tag{6}$$

where Linear is a 1×2-dimensional learnable linear transform, $||\mathbf{x}||$ denotes the $\mathcal{L}_2$ norm of the input vector, N represents the number of channels which is always set to 2 in our task.

The output vector $\mathbf{z} = [z^a, z^b]^T \in [0, 1]$ from $\mathrm{SPP}(\cdot)$ represents the weight assigned to AC and BC channels. When s becomes small, the weight will be biased towards AC or BC channels. Based on this channel-reweighting property, MMT can fuse the air and bone conduction information effectively and flexibly.

After passing the fused context vector $\mathbf{r}_l$ through output layer, we obtain the final attention-based probability $p_{\mathrm{att}}(\mathbf{w})$, where $\mathbf{w} = \{w_1, w_2, \cdots, w_L\}$ denotes a predicted character sequence. At the same time, after passing the output of the encoder $\mathbf{H}^i$ through the CTC layer, we obtain the CTC -based probability $p^i_{\mathrm{ctc}}(\mathbf{w})$ where $i \in \{a, b\}$.

2.4. Training and decoding objectives

In the training phase, the objective function is to minimize:

$$\mathcal{L} = (1-\lambda)\log p_{\mathrm{att}}(\hat{\mathbf{w}}) + \frac{1}{2}\lambda(\log p^a_{\mathrm{ctc}}(\hat{\mathbf{w}}) + \log p^b_{\mathrm{ctc}}(\hat{\mathbf{w}})) \tag{7}$$

where $\hat{\mathbf{w}}$ is the target output sequence, $0 \le \lambda \le 1$ is a tunable CTC weight control factor.

In the decoding phase, we adopt the one-pass beam search [19]:

$$\tilde{\mathbf{w}} = \arg\max_{\mathbf{w}} \left\{(1-\lambda)\log p_{\mathrm{att}}(\mathbf{w}) + \lambda \log p^+_{\mathrm{ctc}}(\mathbf{w})\right\} \tag{8}$$

where $\tilde{\mathbf{w}}$ is the predicted output sequence, $p^+_{\mathrm{ctc}}(\mathbf{w})$ is the scaling-sparsemax-based CTC prefix probability computed by:

$$\log p^+_{\mathrm{ctc}}(\mathbf{w}) = z^a * \log p^a_{\mathrm{ctc}}(\mathbf{w}) + z^b * \log p^b_{\mathrm{ctc}}(\mathbf{w}) \tag{9}$$

where the channel weights z^a and z^b are calculated by (4).

3. EXPERIMENTS

3.1. Experimental settings

We collected a multi-modal corpus of synchronized AC and BC speech in an anechoic chamber, which contains 53 hours of Mandarin speech data from 100 speakers (50 males and 50 females). It is collected from a headset that integrates both AC and BC microphones. The text source for reading comes from over 30000 daily dialogues and RASC863 [20]. The duration of each utterance is in the range of $[1, 5]$ seconds. The speech was recorded at a sampling rate of 44.1kHz and further down-sampled to 16kHz. We divide this corpus into three subsets. The 'train' subset contains 84 speakers. The 'dev' and 'test' subsets contain 8 speakers respectively.

To simulate a complex noisy environment, we added additive noise to AC channel of the corpus. Because BC channel will not be contaminated by additive noise in real-world scenarios, we do not change BC channel. The noise source for the 'train' and 'dev' subsets is a large-scale noise library containing over 20000 noise segments [21]. The noise source for the 'test' subset is the non-stationary noise from the CHiME-3 dataset [22] and NOISEX-92 corpus [23]. For the 'train' and 'dev' subsets, we controlled the SNR in a range of $[0, 20]$ dB. For the 'test' subset, we set the SNR to six levels, which are $\{-5, 0, 5, 10, 15, 20\}$ dB, respectively.

We first perturbed the speech speed by 0.9 times and 1.1 times and extracted 80-dimensional Mel-banks as the acoustic feature. Then, we used SpecAugment [24] to augment the training data. The ground-truth labels were set at the character level. The size of the dictionary was set to 5209.

The kernel size of the convolutional layer in the conformer was set to 15. The block numbers of the conformer encoder and Truncated decoder were set to 12 and 6, respectively. The number of heads in the MHA module was set to 8. The number of units in the position-wise feedforward module was set to 2048. The model dimensions D_m is 256. We applied the relative position embedding to the encoder, and absolute position embedding to the decoder, respectively.

In the training phase, the parallel branches of the proposed system were initialized with the parameters of the pre-trained models at each modality. Then, MMT was fine-tuned with the multi-modal speech. The control factor λ was set to 0.3. In the decoding phase, λ was set to 0.5 and the beam size was set to 10.

To compare with the proposed system, we designed two baselines. The first one is a conformer based single-modal system trained with only the noisy AC speech. The second is a conformer based multi-modal system that concatenates the features of AC and BC speech as its input. It first takes the model of the first baseline as its pre-trained model, and then fine-tunes the system with the multi-modal speech. Note that the convolutional embedding layer was modified to fit the input dimension. The character error rate (CER) was used as the evaluation metric.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on August 28,2022 at 14:53:41 UTC from IEEE Xplore. Restrictions apply.

Table 2: CER (%) of the proposed method and two baselines on the noisy test sets.

System	Type of training data	Type of test data	SNR of test set					
			-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
Conformer	clean AC	noisy AC	100.3	91.6	72.8	47.3	29.2	20.4
	noisy AC		68.7	38.5	21.5	14.4	11.6	**10.4**
Conformer	Multi-modal	Multi-modal	21.6	18.1	15.2	13.1	11.7	11.0
Multi-modal transducer (proposed)			**18.1**	**15.8**	**13.6**	**12.1**	**11.2**	10.9

Table 1: CER (%) comparison of the conformer-based single-modal ASR using AC or BC speech modality.

System	Type of training data	Type of test data	Subset	
			Dev	Test
Conformer	clean AC	clean AC	6.4	9.9
	clean AC	BC	69.5	77.4
	BC	BC	11.1	18.1

3.2. Results and discussion

We first trained the standard conformer with the clean AC speech, and tested the model on clean AC and BC speech respectively. The result is shown in Table 1. By comparing results in the two test scenarios, we observe that the CER on BC speech is much higher than that on the clean AC speech, which means that the ASR model trained with the clean AC speech does not have a good generalization performance on BC speech. We also trained the same conformer with BC speech, and tested the model on BC speech. The result is listed in Table 1. From the table, we find that the performance of the system using only BC speech is not so bad. Due to the shortcomings of BC speech, the performance of the BC-based ASR system is worse than the AC-based onein their respective matching scenarios. Then, we tested the comparison methods on the multi-modal noisy data. Table 2 lists the results of the proposed system and the baselines. From the table, we see that, not only the proposed method but also the conformer baseline trained with multi-modal data are significantly better than the conformer baseline trained with only the noisy AC speech, when the SNR is lower than 15 dB, which demonstrates the importance of exploring BC speech for ASR. In addition, we observe that the proposed system achieves the best performance, which supports the advantage of the proposed MMT in fusing AC and BC speech over the conformer baseline where the multi-modal data is simply concatenated without channel reweighting.

To further analyze the effectiveness of the proposed MMT module, as well as to study how much different modality contributes to the performance, we analyzed the channel weights of AC and BC speech allocated by the MMT module in Fig. 3. From the figure, we see clearly that, when the SNR decreases, the weight of BC channel is gradually increased, which further indicates that BC speech contributes to the performance improvement of the multi-modal ASR over the single-modal ASR with AC speech only in the low SNR environments.

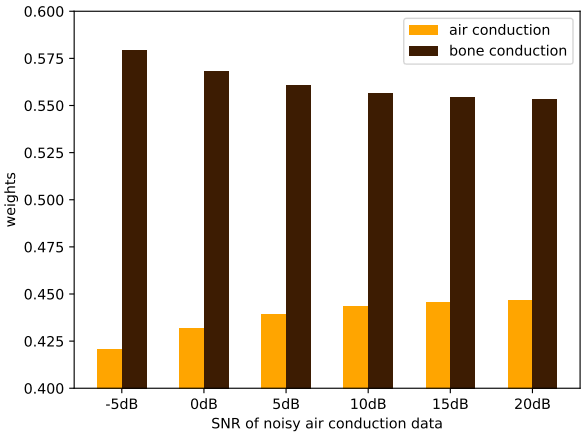

Fig. 3: Channel weights produced by the MMT module of the proposed method with respect to SNR. Note that, the weights are the average ones from all sentences at a SNR level.

4. CONCLUSION

In this paper, we propose a conformer-based multi-modal ASR system using AC and BC speech. In the proposed system, the conformer encoder and transformer-based truncated decoder are used to transform the semantic information of AC and BC channels respectively, then the MMT module applies the scaling sparsemax operator to re-weight and fuse the representations of AC and BC speech. Experimental results show that BC speech contains useful semantic information that is particularly helpful for robust ASR in adverse environments as an auxiliary source of AC speech. Moreover, the proposed system can effectively take advantage of both AC and BC speech, which leads to significant performance improvement in the low SNR environments over the single-modal system with only AC speech and the multi-modal system that simply concatenates AC and BC speech.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on August 28,2022 at 14:53:41 UTC from IEEE Xplore. Restrictions apply.

5. REFERENCES

[1] Tian Tan, Yanmin Qian, Hu Hu, Ying Zhou, Wen Ding, and Kai Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1393–1405, 2018.

[2] Li Chai, Jun Du, Qing-Feng Liu, and Chin-Hui Lee, "A cross-entropy-guided measure (cegm) for assessing speech recognition performance and optimizing dnn-based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 106–117, 2021.

[3] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 5036–5040.

[4] Takuya Yoshioka and Mark JF Gales, "Environmentally robust asr front-end for deep neural network acoustic models," *Computer Speech & Language*, vol. 31, no. 1, pp. 65–86, 2015.

[5] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[6] Dmitriy Serdyuk, Kartik Audhkhasi, Philémon Brakel, Bhuvana Ramabhadran, Samuel Thomas, and Yoshua Bengio, "Invariant representations for noisy speech recognition," *arXiv preprint arXiv:1612.01928*, 2016.

[7] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.

[8] Peidong Wang, Ke Tan, and De Liang Wang, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2019.

[9] Archiki Prasad, Preethi Jyothi, and Rajbabu Velmurugan, "An investigation of end-to-end models for robust speech recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6893–6897.

[10] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7613–7617.

[11] Rongfeng Su, Xunying Liu, Lan Wang, and Jingzhou Yang, "Cross-domain deep visual feature generation for mandarin audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 185–197, 2020.

[12] George Sterpu, Christian Saam, and Naomi Harte, "How to teach dnns to pay attention to the visual modality in speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1052–1064, 2020.

[13] Pan Zhou, Wenwen Yang, Wei Chen, Yanfeng Wang, and Jia Jia, "Modality attention for end-to-end audio-visual speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6565–6569.

[14] Cheng Yu, Kuo-Hsuan Hung, Syu-Siang Wang, Yu Tsao, and Jeih-weih Hung, "Time-domain multi-modal bone/air conducted speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 1035–1039, 2020.

[15] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek, "SEANet: A Multi-Modal Speech Enhancement Network," in *Proc. Interspeech 2020*, 2020, pp. 1126–1130.

[16] Tassadaq Hussain, Yu Tsao, Sabato Marco Siniscalchi, Jia-Ching Wang, Hsin-Min Wang, and Wen-Hung Liao, "Bone-conducted speech enhancement using hierarchical extreme learning machine," in *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pp. 153–162. Springer, 2021.

[17] Hung-Ping Liu, Yu Tsao, and Chiou-Shann Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," *Speech Communication*, vol. 104, pp. 106–112, 2018.

[18] Junqi Chen and Xiao-Lei Zhang, "Scaling sparsemax based channel selection for speech recognition with ad-hoc microphone arrays," *arXiv preprint arXiv:2103.15305*, 2021.

[19] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[20] Aijun Li, Zhigang Yin, Tianqing Wang, Qiang Fang, and Fang Hu, "Rasc863-a chinese speech corpus with four regional accents," *ICSLT-o-COCOSDA, New Delhi, India*, 2004.

[21] Xu Tan and Xiao-Lei Zhang, "Speech enhancement aided end-to-end multi-task learning for voice activity detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6823–6827.

[22] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third chimespeech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[23] Andrew Varga and Herman JM Steeneken, "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 communication*, vol. 12, no. 3, pp. 247–251, 1993.

[24] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

Authorized licensed use limited to: NORTHWESTERN POLYTECHNICAL UNIVERSITY. Downloaded on August 28,2022 at 14:53:41 UTC from IEEE Xplore. Restrictions apply.