End-to-End Multi-Modal Speech Recognition on an Air and Bone Conducted Speech Corpus

Mou Wang, *Graduate Student Member, IEEE*, Junqi Chen, *Graduate Student Member, IEEE*, Xiao-Lei Zhang, *Senior Member, IEEE*, and Susanto Rahardja, *Fellow, IEEE*

***Abstract*—Automatic speech recognition (ASR) has been significantly improved in the past years. However, most robust ASR systems are based on air-conducted (AC) speech, and their performances in low signal-to-noise-ratio (SNR) conditions are not satisfactory. Bone-conducted (BC) speech is intrinsically insensitive to environmental noise, and therefore can be used as an auxiliary source for improving the performance of an ASR at low SNR. In this paper, we first develop a multi-modal Mandarin corpus, which contains air- and bone-conducted synchronized speech (ABCS). The multi-modal speeches are recorded with a headset equipped with both AC and BC microphones. To our knowledge, it is by far the largest corpus for conducting bone conduction ASR research. Then, we propose a multi-modal conformer ASR system based on a novel multi-modal transducer (MMT). The proposed system extracts semantic embeddings from the AC and BC speech signals by a conformer-based encoder and a transformer-based truncated decoder. The semantic embeddings of the two speech sources are fused dynamically with adaptive weights by the MMT module. Experimental results demonstrate the proposed multi-modal system outperforms single-modal systems with either AC or BC modality and multi-modal baseline system by a large margin at various SNR levels. It also shows the two modalities complement with each other, and our method can effectively utilize the complementary information of different sources.**

***Index Terms*—Speech recognition, multi-modal speech processing, bone conduction, air- and bone-conducted speech corpus.**

I. Introduction

Automatic speech recognition (ASR) has achieved significant improvement in the deep learning era. In [1], deep neural networks (DNN) were introduced into large vocabulary continuous speech recognition, making the DNN-Hidden Markov Model (DNN-HMM) architecture predominating ASR. Subsequently the end-to-end deep models which essentially integrate HMM into neural network architectures overtook the popularity. There are mainly three categorises, i.e., the connectionist temporal classification (CTC) model [2], recurrent neural network transducer [3], and encoder-attention-decoder model [4]. Inspired by the success of the transformer architecture in other sequence-to-sequence tasks [5], a number of transformer-based ASR systems have achieved the state-of-the-art performance [6], [7].

Although huge efforts have been made, modern ASR systems still suffer from adverse acoustic conditions [8], such as reverberation, channel distortions, ambient noise, and multiple speakers. To address the issue of improving ASR robustness, many articles were published in the literature and they can be mainly divided into two categories: (i) feature-based front-ends, and (ii) model-based ASR back-ends. The former attempts to remove the noise component from the acoustic feature by adding a speech enhancement front-end or applying a feature transformation [8], [9], [10]. The latter relies on a model to realize robust ASR,which includes three types, i.e., learning the noise-invariant semantic representations [11], exploring the novel noise-robust acoustic models [12], [13], and designing the noise-aware adaptive training methods [14], [15]. However, the aforementioned methods consider speech signals only, leaving large amount of complementary information from other modalities unexplored.

Inspired by the multi-modal nature of human speech perception, there have been increasing interests in incorporating other modalities into ASR recently [16]. A typical multi-modal ASR is the audio-visual speech recognition, which collects video recordings of speakers' mouths jointly with their speech recordings [17], [18], [19], [20], [21]. The complementary information in the video recordings is that lip movements can provide an apparent place of articulation [22]. Although such complementary information is immune to acoustic noises, audio-visual speech recognition still faces several challenges.

Since the performance of multi-modal ASR depends crucially on how much complementary information the modalities provide each other [16], generally the lip movements of some phonemes are similar, which results in limited complementary information. In addition, the visual recordings themselves face many real-world problems that hinder the effectiveness of the

Manuscript received 30 January 2022; revised 1 August 2022 and 30 October 2022; accepted 7 November 2022. Date of publication 23 November 2022; date of current version 15 December 2022. The work of Susanto Rahardja was supported in part by the Oversea Expertise Introduction Project for Discipline Innovation, 111 Project, under Grant B18041. This work was supported in part by the National Science Foundation of China, under Grant 62176211, and in part by the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ozlem Kalinli. *(Mou Wang and Junqi Chen contributed equally to this work.) (Corresponding authors: Susanto Rahardja; Xiao-Lei Zhang.)*

Mou Wang, Junqi Chen, and Xiao-Lei Zhang are with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: wangmou21@mail.nwpu.edu.cn; jqchen@mail.nwpu.edu.cn; xiaolei.zhang@nwpu.edu.cn).

Susanto Rahardja is with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Singapore Institute of Technology, Singapore 138683 (e-mail: susantorahardja@ieee.org).

Digital Object Identifier 10.1109/TASLP.2022.3224305

audio-visual speech recognition, including missing of faces, occlusion, low resolution, and additional computational complexity.

All the aforementioned ASR methods were developed with air-conducted (AC) speech only. However, owing to the characteristics of AC microphones, AC speech is easily polluted by noise, which makes the performance of ASR systems drop significantly in low signal-to-noise ratio (SNR) environments, especially in the presence of non-stationary noises, such as babble noise and wind noise.

Bone-conducted (BC) microphone on the other hand is a kind of non-audible sensor. It is attached to the skin of a speaker, and records speech by converting the local vibration around the speaker's skull induced by the voice of the speaker into electrical signals. Therefore, the BC microphone is relatively insensitive to external sources, and has the intrinsic capability of suppressing background noise from environments, which makes it possible to promote the performance of speech related systems significantly, especially at low SNR levels. However, BC speech has many shortcomings. First and foremost, BC speech suffers severe loss of the high-frequency components because of the channel attenuation of human tissues. It is observed that only the spectral components below 1 kHz can be recorded effectively, which leads to the degradation of speech intelligibility. Moreover, BC speech contains self-noise due to the resonance and friction between the BC microphone and the skin of the speaker. Finally, the characteristics of BC speech are speaker-dependent, which decreases the generalization of traditional signal processing methods. These shortcomings bring new challenges to BC speech recognition.

As mentioned, BC speech is insensitive to external noise sources, which makes it worthy to be used as a complementary signal of AC speech for developing a robust ASR in low SNR environments. In the literature, BC speech has been used as an auxiliary source of AC speech for multi-modal speech enhancement [23], [24], [25], [26]. In real application, BC sensors has been integrated into some TWS headsets for voice activity detection, such as Apple Airpods Pro, Huawei Freebuds 3, etc. However, to our knowledge, the multi-modal ASR based on AC and BC speech signals and deep learning has not been explored yet. In addition, the datasets of the BC speech used in those works are too small-scale to train modern ASR systems effectively since ASR models often require large amounts of high quality data. In AC speech, there are many high quality free corpora, such as Librispeech [27], AISHELL-2 [28], THCHS30 [29]. However, there is no large-scale high-quality corpus of AC and BC synchronized speeches for the ASR research so far.

To address the above issues, in this paper, we first develop a multi-modal speech corpus, named *air- and bone-conducted speech corpus* (ABCS), which contains 47,182 AC and BC synchronized utterances with a total valid duration of 42 hours. Based on this corpus, we analyze and compare the characteristics of the AC and BC speech signals from the perspective of acoustic features and mutual information. Then, we propose a multi-modal conformer ASR system based on a novel multi-modal transducer. The proposed system first extracts Mel-spectrograms from the AC and BC speech signals respectively, then feeds the Mel-spectrograms into the conformer network. A multi-modal transducer (MMT) is proposed to assign adaptive weights to the AC and BC branches by a scaling sparsemax operator, which fuses the two branches dynamically according to the instantaneous SNR levels along the time axis. The performance of the multi-modal ASR can be further improved by initializing the models pre-trained with other AC speech corpora. There are two main contributions in this paper. First is the development of a combined AC and BC synchronized large-scale speech corpus, which is essentially a data set useful for the bone conducted ASR research. Second, an end-to-end multi-modal ASR system is developed specially for the AC and BC joint speech recognition.

The remainder of the paper is organized as follows. Section II describes the multi-modal speech database. Section III introduces the proposed multi-modal speech recognition system. Experimental setups are presented in Section IV, and results and discussion are presented in Section V. Finally, Section VI concludes this paper.

II. Air- and Bone-Conducted Speech Corpus

The proposed ABCS corpus is composed of 47,182 AC and BC synchronized utterances from 100 speakers, which amounts to 42 hours of speech with each speaker contributing approximately 25 minutes. The duration of each utterance varies from 1 to 5 seconds. Table I compares the proposed ABCS corpus with existing BC speech corpora. From the table, we see that our corpus is much larger than the other corpora. Moreover, our corpus is multi-modal with synchronized AC and BC channels. In the following, the data collection process is described in details.

A. Data Collection

1) Text Transcription: The comprehensiveness and diversity of syllables, phonemes and cadences should be considered in text transcription. Thus, we used the transcriptions of two corpora, which are the transcription of the RASC863 corpus made by the Chinese Academy of Social Sciences and a 30 k daily dialogue corpus provided by the iFLYTEK corporation. For the transcription of RASC863, the content was collected from broadcast news and interview. Most syllables, phonemes and tones are covered. However, the number of the sentences in RASC863 is limited. By contrast, the 30 k daily dialogue corpus does not consider the balance problem of syllables. The total number of prompts is around 30,000. The content covers many topics in daily life, such as finance, sports, study, entertainment, etc.

2) Collection Process: To collect synchronised AC and BC speech signals, we used a headset where air and bone conducted microphones are integrated. The headset is designed by the SabineTek corporation. The two-channel signals from the headset were recorded by a Zoom H1n handy audio recorder. The AC and BC speech signals were recorded at a sampling rate of 44.1 kHz, and further downsampled to 16 kHz.

As shown in Fig. 1, the corpus was collected in an anechoic chamber, where the speakers wore the headset and read the text transcription. The anechoic chamber was built in accordance

TABLE I
SUMMARY OF AIR- AND BONE-CONDUCTED SPEECH CORPORA

Name	Year	Location	Language	Speakers	Utterances	Duration (h)	Publicly available
Erzin et al. [30]	2009	throat	Turkish	21	8,400	10.4	no
Turan et al. [31]	2015	throat	Turkish	1	799	-	no
Shan et al. [32]	2018	throat	Mandarin	16	3,200	-	partially
Yu et al. [24]	2020	throat	Mandarin	1	320	-	yes
Tagliasacchi et al. [26]	2020	earbud	English	25	-	1.25	no
Zheng et al. [33]	2022	throat	Mandarin	6	5,330	5.2	no
ABCS (proposed)	2022	ear canal	Mandarin	100	47,182	42	yes

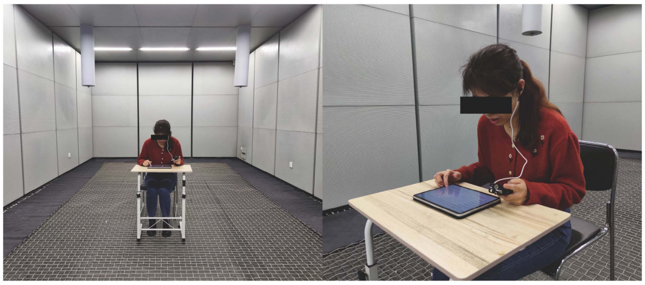
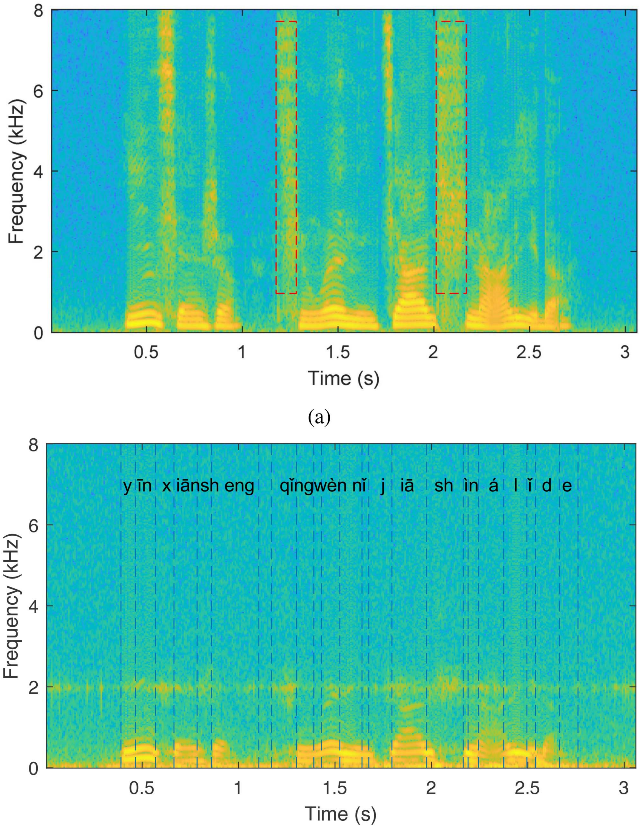


Fig. 1. A snapshot of the recording process of the ABCS corpus. The left figure shows the recording environment. The right figure shows the headset, handy audio recorder and prompter.

with the ISO 3745 [34]. The size of the anechoic chamber is $11.8 \times 4.2 \times 7.6$ m^3. One hundred native Chinese speakers from 20 to 35 years old participated in the recording, including 50 males and 50 females. They spoke standard Mandarin. To encourage the diversity of the data, we have no requirement on the volume and emotion. To avoid the noise generated by the friction between the BC microphone and the skin of the speaker, the speaker was required to tightly wear the headset and movement was strictly controlled and monitored.

3) Data Processing and Labeling: Before labeling, the raw recording were split into sentences with little silence segments. Because the BC speech lost many components of speech such as the consonant, aspirate and fricative, we could not apply voice activity detection to split the BC speech. Therefore, the time stamps of the speech were first obtained from the clean AC speech only, and then used to segment the BC speech recordings.

To further improve the quality of our corpus, data cleaning was carried out manually to remove incomplete, misread, or repeated utterances. After data cleaning, 42 hours of valid utterances were collected. Finally, the utterances were labeled manually with the text transcription serving as references. The database is available online at https://github.com/wangmou21/abcs.



(a)

(b)

Fig. 2. Spectrograms of AC and BC speech signals, where the text transcription of this example is "Mr. Yin, where is your home?" (in Chinese). (a) Spectrogram of AC speech. (b) Spectrogram of BC speech. The string "y īn x ian sh eng q ǐng w èn n ǐ j iā sh ì n á l ǐ d e" is phoneme annotations. The blue lines are the stamps for phoneme annotations.

B. Acoustic Characteristics

The magnitude spectrogram of short-time Fourier transform (STFT) is used as the visualized representation of speech. Fig. 2 shows the spectrograms of the AC and BC speech signals of an utterance. Comparing the 2 figures we can see that the high frequency components of the BC speech spectrogram (bottom) are severely attenuated. Only the frequency components less than 1 kHz can be partially recorded. Particularly, the BC speech spectrogram lost most unvoiced consonants and fricatives, as shown in the red-dashed box of Fig. 2(a). In addition, the BC speech spectrogram contains self-noise in the low-frequency components and sustained narrow-band noise at approximately 2 kHz. The low-frequency self-noise was essentially generated by the accidental friction between the BC microphone and the skin of the speaker. The sustained narrow-band noise was generated by the resonance of the BC microphone and the contact skin of the speaker.

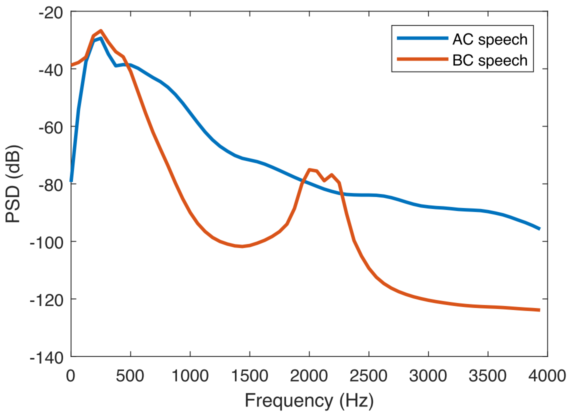


Fig. 3. Comparison of the power spectral density of AC and BC speech signals.

To further analyze the bandwidth and frequency characteristics, we computed the power spectral density (PSD) of the entire corpus. The statistics of the PSD are shown in Fig. 3. It can be seen that the AC and BC speech has similar responses in the frequency band of 100 to 500 Hz. However, the BC speech has higher power and slower frequency attenuation than the AC speech in this frequency band. In the frequency band of 500 to 1200 Hz, the AC and BC speech signals degrade by about 3.4 dB and 9.7 dB per 100 Hz, respectively. At the 1 kHz, the power of the BC speech has decreased approximately 60 dB, which causes the BC speech sound to be dull. In addition, the sustained narrow-band noise is in the range of 1.8 to 2.4 kHz, which does not overlap with the main frequency bins of the BC speech. We did not filter out this noise, because filtering will unfortunately remove the residual speech in this frequency band as well. Therefore, we retained all noise of the BC speech without applying any filter.

C. Mutual Information Analysis

The ambient noise in the AC speech and the information loss in the medium and high frequency bins of the BC speech have different influences on the human speech intelligibility. To explore how much semantic information is contained in the BC speech and noisy AC speech, we analyzed the mutual information of the noisy AC speech or BC speech with its corresponding synchronized clean AC speech.

Specifically, we used the Gaussian mixture model (GMM) based mutual information approach [35], [36] for the frequency analysis of speech. To analyze how much information the BC speech had lost in the high-frequency band, we computed the mutual information of the low-frequency and high-frequency components respectively. First, Mel-frequency ceptral coefficients (MFCC) which mimic human speech perception, were extracted from each utterance. According to the critical bands of MFCC at a sampling frequency of 16 kHz, 29 triangular filters were used, where the first 12 filters covered the frequency range of 0 to 1.2 kHz and the remaining 17 filters covered the range of 1.2 to 8 kHz. Then, a GMM was built to model the joint probability density function (PDF) for any speech type or any particular frequency range in interests:

$$P_{\mathrm{GMM}}(x, y) = \sum_{m=1}^{M} \omega_m P_G\left(x, y|\theta_m\right), \tag{1}$$

where x and y represent acoustic features from two types of speech sources respectively, M refers to the number of mixture components, ω_m is the mixture weight of the m-th mixture component, and $P_G(\cdot)$ is the multivariate Gaussian distribution defined by the parameter $\theta_m = \{\mu_m, C_m\}$ with μ_m and C_m as the mean vector and diagonal covariance matrix of the distribution respectively. Finally, the mutual information is calculated by:

$$MI\left(\widehat{X;Y}\right) = \frac{1}{N}\sum_{n=1}^{N}\left(\log 2\left(\frac{P_{\mathrm{GMM}}\left(x_n, y_n\right)}{P_{\mathrm{GMM}}(x_n)P_{\mathrm{GMM}}(y_n)}\right)\right), \tag{2}$$

where N is the number of speech frames. The larger the value of the mutual information is, the more the speech information in the recordings contains.

We first added white Gaussian noise to AC speech to produce noisy AC speech at different SNR. We then computed the mutual information of the noisy AC speech and BC speech with the corresponding clean AC speech in the ranges of 0 to 1.2 kHz and 1.2 to 8 kHz, respectively. The results in Fig. 4 show that the mutual information between the noisy AC speech and its clean speech counterpart decreases with the decrease of SNR. From Fig. 4(a) and (b), we observe that the information contained in the BC speech is close to the noisy AC speech at an SNR level of 5 dB in the ranges of both 0 to 1.2 kHz and 1.2 to 8 kHz. It indicates that the BC speech can provide abundant information for multi-modal ASR. It is interesting that the high frequency band of the BC speech also contains certain semantic information, which is intuitively unobservable from the spectrogram.

III. Multi-Modal Speech Recognition

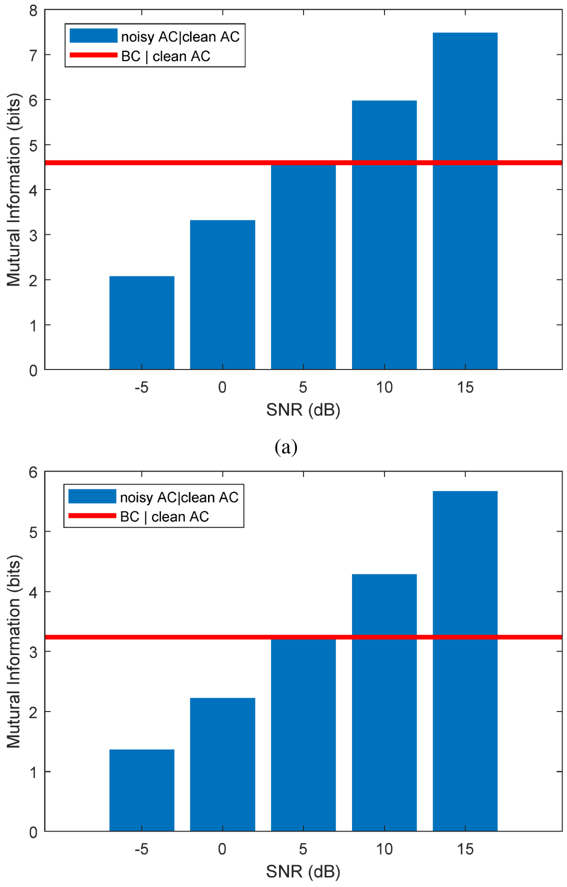
In this section, we first present the overall architecture of a conformer-based multi-modal ASR system in Section III-A, and then present the components of the system from Section III-B to Section III-E.

A. System Overview

Fig. 5 shows the architecture of the proposed system. Mel-spectrograms are first extracted from the AC and BC speech signals, and then passed through two parallel branches respectively. Each branch comprises of a conformer-based encoder, a transformer-based truncated decoder, and a CTC layer, and 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 to produce a fused context vector. Finally, the fused context vector passes through the output layer, which produces the final attention-based output probability.

B. Parallel Branches

The proposed system consists of two parallel branches, which are used to extract high-level semantic representations from the



(a)

(b)

Fig. 4. Mutual information of the noisy AC speech and BC speech with the clean counterpart. (a) Mutual information in the frequency band of 0 to 1.2 kHz. (b) Mutual information in the frequency band of 1.2 to 8 kHz.

AC and BC speech signals respectively. The architecture of each branch is described in Fig. 6. Specifically, the Mel-spectrogram features $\mathbf{X}^{\mathrm{a}} \in \mathbb{R}^{T\times D_a}$ and $\mathbf{X}^{\mathrm{b}} \in \mathbb{R}^{T\times D_b}$ are first extracted from the AC and BC speech signals respectively, where T refers to the number of frames, a and b denote the AC and BC branches respectively, and D_a and D_b refer to the dimensions of the Mel-spectrograms. Then, each branch first extracts a high level representation $\mathbf{H}^i, \forall i \in \{a, b\}$ from the Mel-spectrogram feature by a conformer-based encoder shown in Fig. 6(a). Subsequently, the high level representation $\mathbf{H}^i$ passes through a transformer-based truncated decoder shown in Fig. 6(b), which generates a context vector $\mathbf{c}_l^i$ and a guide vector $\mathbf{g}_l^i$ at each time step l. Finally, they are stacked to form a context matrix $\mathbf{C}_l = [\mathbf{c}_l^a, \mathbf{c}_l^b]^T \in \mathbb{R}^{2\times D_m}$, and a mean pooling guide vector $\mathbf{g}_l = \mathrm{Mean}(\mathbf{g}_l^a, \mathbf{g}_l^b) \in \mathbb{R}^{D_m}$ for the subsequent MMT module. The details of the encoder and decoder are described as follows.

1) Conformer-Based Encoder: As shown in Fig. 6(a), the encoder has a convolution subsampling layer and multiple conformer blocks. The convolution sub-sampling layer downsamples the number of input feature frames by 4 times and convert the feature dimension to the model dimension D_m.

Multiple conformer encoder blocks are stacked to extract local and global dependencies of the feature sequence. Each block consists of two position-wise feed-forward (FFN) modules, a multi-head attention (MHA) module, and a convolution module. The MHA module is used to model the long-term context and capture the global information of the feature sequence while the convolution module containing a 1-D depthwise convolution layer is to exploit and capture the local information. The output of the conformer-based encoder $\mathbf{H}^i \in \mathbb{R}^{T//4\times D_m}$ is calculated as follows:

$$\mathbf{H}^i = \mathrm{Enc}\left(\mathrm{Sub}\left(\mathbf{X}^i\right)\right), \quad \forall i \in \{a, b\}, \tag{3}$$

where $\mathrm{Sub}(\cdot)$ refers to the convolution sub-sampling layer, $\mathrm{Enc}(\cdot)$ refers to the multiple conformer-based encoder blocks.

2) Transformer-Based Truncated Decoder: The decoder extracts the corresponding context vector $\mathbf{c}_l^i \in \mathbb{R}^{D_m}$ in each time step l as follows

$$\mathbf{c}_l^i = \mathrm{Tdec}\left(\mathbf{H}^i, \mathbf{y}_{1:l}^i\right), \quad \forall i \in \{a, b\}, \tag{4}$$

where $\mathbf{y}_{1:l}^i \in \mathbb{R}^{l\times D_m}$ represents a sequence of shifted output embedding vector, and $\mathrm{Tdec}(\cdot)$ represents the transformer truncated decoder. As shown in Fig. 6(b), it also consists of multiple blocks, where each block except the last one comprises of an MHA, a masked MHA and an FFN module. In the last block, FFN module is removed to retain more original semantic information. In addition, in the first decoder block, the output of the masked MHA module in each time step, named the *guide vector* $\mathbf{g}_l^i \in \mathbb{R}^{D_m}$, is extracted by:

$$\mathbf{g}_l^i = \mathrm{MHA}\left(\mathbf{y}_l^i, \mathbf{y}_{1:l}^i, \mathbf{y}_{1:l}^i\right), \quad \forall i \in \{a, b\}. \tag{5}$$

C. Multi-Modal Transducer

The architecture of the proposed MMT is shown in Fig. 7. The MMT module takes the context matrix $\mathbf{C}_l$ and guide vector $\mathbf{g}_l$ as its input. The context vectors $\mathbf{c}_l^a$ and $\mathbf{c}_l^b$ produced from the AC and BC speech signals are linearly transformed, followed by the scaling sparsemax (SSP) [37] operator to adaptively assign weights to the two branches given the two guide vectors of the AC and BC speech signals. Finally, the fused context vector $\mathbf{r}_l \in \mathbb{R}^{D_m}$ is obtained as follows:

$$\mathbf{z}_l = \mathrm{SSP}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_m}}, s\right), \tag{6}$$

$$\mathbf{r}_l = (\mathbf{z}_l\mathbf{V})^T + \mathrm{FFN}\left(\mathrm{LayerNorm}\left((\mathbf{z}_l\mathbf{V})^T\right)\right), \tag{7}$$

where $\mathrm{FFN}(\cdot)$ and $\mathrm{LayerNorm}(\cdot)$ refer to the position-wise feedforward operation and layer normalization respectively, and

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

are the query, key, and value matrices respectively, $\mathbf{W}^Q$, $\mathbf{W}^K$, and $\mathbf{W}^V$ are learnable projection transformation matrices, and $\mathrm{SSP}(\mathbf{x}, s)$ refers to the scaling sparsemax re-weighting operation with a learnable scaling factor s. The formualtion of SSP is expressed as follows:

$$\mathrm{SSP}_i(\mathbf{x}, s) = \max\left(x_i - \tau(\mathbf{x}, s), 0\right)/s, \ i = 1, \ldots, K \tag{8}$$

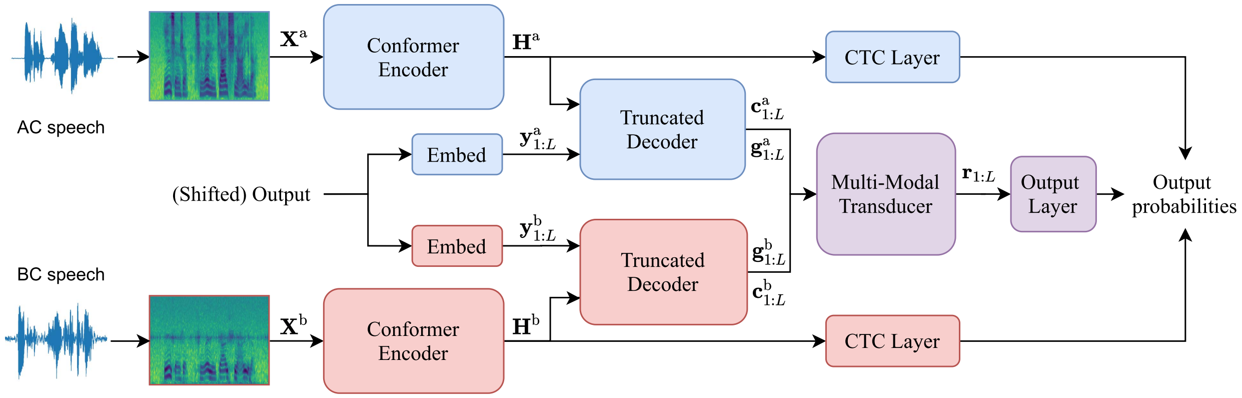
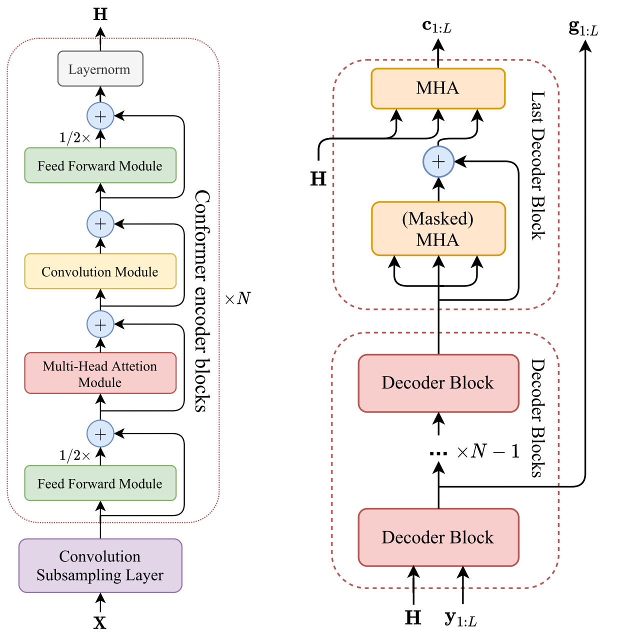

AC speech
$\mathbf{X}^{\mathrm{a}}$
Conformer Encoder
$\mathbf{H}^{\mathrm{a}}$
CTC Layer
Truncated Decoder
$\mathbf{c}^{\mathrm{a}}_{1:L}$
$\mathbf{g}^{\mathrm{a}}_{1:L}$
Embed
$\mathbf{y}^{\mathrm{a}}_{1:L}$
(Shifted) Output
Multi-Modal Transducer
$\mathbf{r}_{1:L}$
Output Layer
Output probabilities
Embed
$\mathbf{y}^{\mathrm{b}}_{1:L}$
Truncated Decoder
$\mathbf{g}^{\mathrm{b}}_{1:L}$
$\mathbf{c}^{\mathrm{b}}_{1:L}$
BC speech
$\mathbf{X}^{\mathrm{b}}$
Conformer Encoder
$\mathbf{H}^{\mathrm{b}}$
CTC Layer

Fig. 5. Architecture of the proposed end-to-end multi-modal ASR system. The parameters of the modules in blue color are from a single-modal ASR that is pre-trained with the AC speech, the parameters of the modules in red color are from a single-modal ASR that is pre-trained with the BC speech, and the parameters of the modules in purple color are those that require fine-tuning.



(a)

(b)

Fig. 6. Architecture of each of the two parallel branches, which consists of (a) a conformer-based encoder and (b) a transformer-based truncated decoder.

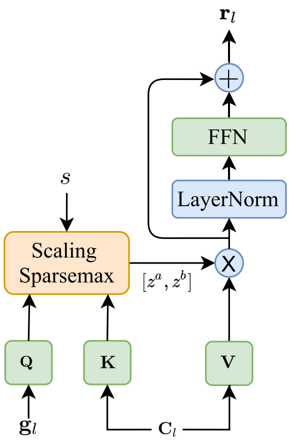


Fig. 7. The proposed multi-modal transducer.

where $\tau(\cdot)$ is a threshold computed by

$$\tau(\mathbf{x}, s) = \left(\sum_{i=1}^{k} x_{(i)} - s\right) / k \tag{9}$$

with the parameter k obtained by:

$$k = \arg\max_k x_{(k)} \geq \left(\sum_{i=1}^{k} x_{(i)} - s\right) / k, \ x_{(1)} \geq \cdots \geq x_{(K)} \tag{10}$$

The scaling factor s can be computed as follows:

$$s = 1 + \mathrm{ReLU}\left(\mathrm{Linear}(||\mathbf{x}||, K)\right), \tag{11}$$

where $||\mathbf{x}||$ denotes the $\mathcal{L}_2$ norm of the input vector, $\mathrm{Linear}(\cdot)$ is a 1×2-dimensional learnable linear transformation, and K is the dimension of the input vector which represents the number of channels and is always set to 2 in this work.

The output vector $\mathbf{z}_l = [z_l^a, z_l^b]^T \in [0, 1]$ of the scaling sparsemax represents the weights assigned to the AC and BC branches. The weight vector can be adjusted adaptively at each time step by controlling s. For example, a small s will make the weight biased towards one of the AC and BC modalities. The channel-reweighting strategy of MMT makes the multi-modal system fuse the semantic information of the AC and BC speech signals flexibly and effectively.

D. Attention and CTC Based Probability

After passing the fused context vector $\mathbf{r}_l$ through the output layer, we can get the attention-based probability $p_{\mathrm{att}}(w_l|w_1, \ldots, w_{l-1})$ at each time step, where $w_l \in \mathcal{V}$ denotes a predicted character at time step l and $\mathcal{V}$ denotes the vocabulary set. Then, the final attention-based probability is computed by:

$$p_{\mathrm{att}}(w_{1:L}) = \prod_{l=1}^{L} p_{\mathrm{att}}\left(w_l|w_1, \ldots, w_{l-1}\right), \tag{12}$$

where $w_{1:L} = w_1, w_2, \ldots, w_L$ denotes a predicted character sequence.

At the same time, each of the BC and AC branches has a CTC layer. The high level representation $\mathbf{H}^i$ produced by the conformer-based encoder is fed into the CTC layer at each branch respectively. The CTC layer produces the output symbol probability $p(s_t|\mathbf{h}_t^i)$ at each frame, where $s_t \in \mathcal{V}'$ denotes the CTC-based output symbol, $\mathcal{V}'$ denotes an extended vocabulary set with a blank symbol, and the subscript t represents the t-th frame. Based on the independence assumption, the CTC-based probability is computed by:

$$p_{\text{ctc}}^i(w_{1:L}) = \sum_{\pi(s_{1:T})=w_{1:L}} \prod_{t=1}^{T} p(s_t|\mathbf{h}_t^i), \forall i \in \{a, b\} \tag{13}$$

where $\pi(\cdot)$ denotes the mapping function from the symbol sequence to the character sequence.

E. Training and Decoding Objectives

In the training stage, the objective function to be optimized [2], [3], [4] is as follows:

$$\mathcal{L} = (1 - \lambda) \log p_{\text{att}}(w_{1:L}^*) + \frac{1}{2}\lambda \left(\sum_{i=a,b} \log p_{\text{ctc}}^i(w_{1:L}^*) \right) \tag{14}$$

where $w_{1:L}^*$ represents the target output sequence, and $\lambda \in [0, 1]$ is a tunable parameter.

In the decoding stage, we use the one-pass beam search [38] to find the character with the highest probability at each time step l, i.e.,

$$\hat{w}_l = \arg\max_{w_l \in \mathcal{V}} \{(1 - \lambda)\alpha_{\text{att}}(w_l) + \lambda\alpha_{\text{ctc}}(w_l)\} \tag{15}$$

where $\hat{w}_l$ denotes a predicted character at time step l, and $\alpha_{\text{att}}(w_l)$ and $\alpha_{\text{ctc}}(w_l)$ are the attention score and CTC-based score respectively:

$$\left\{ \begin{matrix} \alpha_{\text{att}}(w_l) \triangleq \log p_{\text{att}}(w_{1:l}) \\ \alpha_{\text{ctc}}(w_l) \triangleq \sum_{i=a,b} z_l^i \cdot \log p_{\text{ctc}}^i(w_{1:l}, \ldots \end{matrix} \right\} \tag{16}$$

where the channel weights z_l^a and z_l^b are calculated by (6), and $p_{\text{ctc}}^i(w_{1:l}, \ldots)$ denotes the CTC prefix probability [38].

IV. Experimental Setups

In this section, we first describe the setting of the ABCS data set, and then present the experimental setup of the proposed end-to-end multi-modal system, three single-modal systems and a multi-modal conformer system.

A. Dataset

We conducted the proposed multi-modal ASR system on our ABCS corpus. We divided the ABCS corpus into three subsets, named 'train,' 'dev' and 'test'. The 'train' subset contains 84 speakers. The 'dev' and 'test' subsets contain 8 speakers respectively. The numbers of male and female speakers in each subset are balanced. The detailed information is listed in Table II.

TABLE II
Description of the ABCS Corpus for ASR

Subset	Speakers		Duration (h)	Sentences
	Male	Female		
train	42	42	35.85	39692
dev	4	4	3.23	3750
test	4	4	2.95	3740

To simulate a complex noisy environment, we added additive noise to the AC speech of the corpus but no multiplicative noise was added into the noisy AC speech. Because the BC channel was intrinsically not contaminated by additive noise in real-world scenarios, we did not change the BC speech. The noise source for the 'train' and 'dev' subsets is a large-scale noise database containing more than 20,000 noise segments [39], [40]. And the noise source for the 'test' subset is the non-stationary noise from NOISEX-92 corpus [41] and CHiME-3 dataset [42]. 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 five levels, which are $\{0, 5, 10, 15, 20\}$ dB, respectively.

Although the duration of our ABCS corpus is up to 42 hours, there are much more utterances in other large-scale AC speech corpus. To utilize more speech information, we pre-trained partial modules of the proposed system with the AISHELL-2 corpus [28], which contains about 1,000 hours of clean reading-speech data from more than 1900 speakers. The data was recorded via three parallel acoustic channels—an Android smartphone (Android), an iPhone (iOS) and a high fidelity microphone (Mic).

B. Experimental Settings

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 [43] to augment the training data. The ground-truth labels were set at the character level. The size of the dictionary $\mathcal{V}$ was set to 5209.

The kernel size of the convolutional layer in the conformer encoder 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 FFN module was set to 2048. The model dimension 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, the MMT module and output layer were 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. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ and applied a transformer learning schedule with 10 k warm-up steps [5]. The mini-batch size was set to 64. The model was trained for 50 epochs. We use character error rate (CER) as evaluation metric.

TABLE III
RESULTS OF THE SINGLE-MODAL AC-BASED ASR SYSTEM THAT IS PRE-TRAINED WITH AISHELL-2

Dataset	Channel	Subset	CER(%)
AISHELL-2	AC	Dev	8.3
		Test-Android	9.8
		Test-iOS	8.5
		Test-Mic	9.8
ABCS	AC	dev	7.7
		test	12.9
	BC	dev	54.5
		test	67.8

C. Baseline Systems

To compare with the proposed system, we designed multiple baseline systems. There were three baselines trained with the AC speech. The first one was a conformer-based single-modal system trained with the AISHELL-2 corpus. The other two baselines were conformer-based single-modal systems that were first initialized from the first baseline, and then trained with the clean and noisy AC data of the ABCS corpus, respectively. In addition, we designed a multi-modal baseline which took the concatenated acoustic feature of the AC and BC speech of ABCS as its input. Specifically, we used the first baseline to initialize the conformer-based single-modal system. Then, we randomly initialized the convolution subsampling layer to fit the dimension of the concatenated input feature. Finally, we trained this system with the concatenated features from the BC speech and noisy AC speech. This system is referred to as 'multi-modal Conformer'.

V. EXPERIMENTAL RESULTS

In this section, we first demonstrate the effectiveness of the proposed end-to-end multi-modal system over three single-modal systems, and then evaluate a multi-modal conformer system on the proposed ABCS corpus, which in turn supports the importance of the BC speech as a complementary source, for improving the ASR performance.

A. Results of Single-Modal ASR on the ABCS Corpus

1) Single-Modal AC-Based ASR on the AC Test Sets: First of all, we need to validate the correctness of the AC speech and its labels of the ABCS corpus. We first trained a standard single-modal conformer system with the AISHELL-2 corpus, and tested the model on the clean AC speech of the AISHELL-2 and ABCS corpora respectively. The results are shown in Table III. We observe that the pre-trained model has similar performance on the test sets of the AISHELL-2 and ABCS corpora, which demonstrate the effectiveness of the AC speech of the ABCS corpus.

However, when the AC speech is noisy, an ASR system trained with only the AC modality is difficult to obtain satisfactory performance. To evaluate this claim, we trained a conformer standard single-modal system with various training strategies as summarized in Table IV. As we can see from the table, the performance of the ASR system in all noisy test scenarios drop significantly when the SNR decreases, no matter which training strategy is adopted. For instance, when the SNR is below 5 dB, the CER of the single-modal ASR is up to 21%. This phenomenon indicates that additional information, such as the BC modality, may be a solution to the problem.

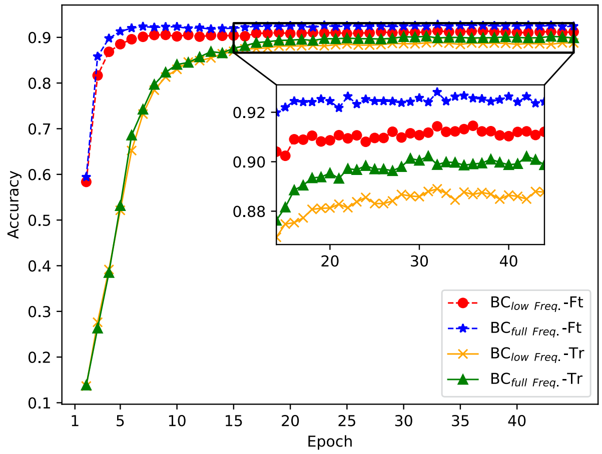


Fig. 8. Convergence curve comparison of BC speech recognition between different strategies. The subscript 'low Freq.' denotes low frequency band while 'full Freq.' denotes full frequency band. The term 'Ft' means applying pre-training while 'Tr' means training from scratch.

2) Single-Modal AC-Based ASR on the BC Test Sets: In this subsection, we evaluate how the single-modal AC-based ASR performs on the BC speech.

As we know, the BC speech loses information at the high frequency band. To evaluate how this weakness affect the performance of ASR, we tested the single-modal AC-based ASR on the BC speech. The results are shown in Table III. By comparing the results on the AC and BC speech of ABCS, we observe that the CER on the BC speech is much higher than that on the clean AC speech. It manifests that the information at the high frequency band is important to the ASR system. Moreover, simply applying the ASR system trained with the AC speech does not generalize well to the BC speech.

3) Single-Modal BC-Based ASR on the BC Test Sets: In this subsection, we evaluate how the single-modal ASR that is trained with the BC speech behaves on the BC test sets when different frequency bands of the BC speech are adopted. The results are summarized in Table V. The convergence curves of the methods are shown in Fig. 8.

Specifically, we first trained a single-modal conformer-based ASR system with the full frequency band of the BC speech of ABCS, and tested the system on the same full frequency band of the BC speech, where the system is not pre-trained with AISHELL-2. The result is listed in the first line of Table V. From the table, we find that although the performance is not too bad, it is still much worse than the performance of the AC-based system on the clean AC speech.

As analyzed in Section II-B, the high frequency components of BC speech fade quickly. To study whether the high-frequency components of BC speech are useful for ASR, we fed only the low frequency band of the BC speech into the ASR system

TABLE IV
CER (%) OF THE SINGLE-MODAL AC-BASED ASR SYSTEM ON THE NOISY AC SPEECH TEST SETS OF ABCS

Methods	Pre-training with AISHELL2	Training	Testing	SNR of test set					
				0 dB	5 dB	10 dB	15 dB	20 dB	clean
Single-modal AC-based conformer [7]	✗	AISHELL-2	Noisy AC	81.3	56.2	34.8	23.5	18.4	12.9
	✓	Clean AC		77.3	49.6	27.0	15.7	11.2	6.8
	✗	Noisy AC		38.5	21.5	14.4	11.6	10.4	11.1
	✓			41.2	21.0	12.2	8.7	7.2	6.9

TABLE V
CER (%) OF THE SINGLE-MODAL BC-BASED ASR SYSTEM THAT IS EVALUATED WITH DIFFERENT FREQUENCY BANDS OF THE BC SPEECH OF ABCS

Method	Pre-training with AISHELL2	Frequency band of BC speech	Subset	
			Dev	Test
Single-modal BC-based conformer [7]	✗	full frequency band	11.1	18.1
		low frequency band	12.9	19.1
	✓	full frequency band	**8.4**	**15.3**
		low frequency band	10.4	16.9

TABLE VI
CER(%) OF THE PROPOSED MMT METHOD AND THE MULTI-MODAL CONFORMER BASELINE

Method	Pre-training with AISHELL2	Frequency band of BC speech	SNR of test set						
			-5dB	0dB	5dB	10dB	15dB	20dB	clean
Multi-modal conformer baseline [7]	✗	full frequency band	21.6	18.1	15.2	13.1	11.7	11.0	11.0
	✓	low frequency band	23.3	17.5	13.2	10.6	9.2	8.2	8.2
	✓	full frequency band	19.1	15.0	11.8	9.6	8.2	7.5	7.6
MMT (proposed)	✗	full frequency band	18.1	15.8	13.6	12.1	11.2	10.9	10.7
	✓	low frequency band	20.4	16.9	12.5	9.5	**7.8**	**7.0**	6.7
	✓	full frequency band	**17.5**	**14.9**	**11.8**	**9.4**	7.9	7.1	**6.7**

for training and test. The low frequency band, which is below 1.5 kHz, corresponds to the first 37 dimensions of Mel-banks features. As shown in Table V and Fig. 8, we find that discarding the high-frequency band yields an increase of CER. That is to say, even if the high frequency components of BC are severely attenuated, their information is still useful to ASR.

Finally, considering that the semantic information of the AC and BC speech signals are strongly related, we pre-trained the BC-based ASR model with AISHELL-2. Compared with the aforementioned cases without pre-training, the CER of the pre-trained model is reduced by an average of relatively 10%. Moreover, as shown in Fig. 8, applying the AC-based pre-training can not only improve the accuracy, but also speed up the convergence of the model training.

B. *Results of Multi-Modal ASR on the ABCS Corpus*

In this subsection, we first evaluate the effectiveness of the proposed multi-modal ASR system on the ABCS corpus by comparing it with the four baseline systems mentioned in Section IV-C, and then analyze how different modalities complement with each other by evaluating the adaptive weights produced by MMT.

1) Main Results: Table VI shows the comparison result between the proposed multi-modal MMT ASR system and the multi-modal conformer baseline that simply concatenates the features of the AC and BC speech. From Tables IV and VI, we can observe that both of the multi-modal systems outperform the single-modal AC-based ASR systems by a large margin in low SNR conditions, which demonstrates that the BC speech can provide sufficient complementary information to AC speech for the performance improvement of ASR in the low SNR levels. However, when the SNR is higher than 15 dB, the multi-modal conformer baseline is slightly worse than the single-modal AC-based systems, while the proposed MMT-based system obtained the best performance in all SNR conditions. The reason why the multi-modal baseline system behaves poorly in the high SNR levels is that it treats the AC and BC speech as equivalently important without distinction. However, when SNR is high, the AC speech is good enough, which makes the BC speech behaves like an “interference”. On the contrary, because the proposed MMT-based system adaptively assigns weights to the BC and AC speech signals according to the instantaneous noisy conditions, it integrates the merits of the AC and BC speech signals together. In addition, we can also find that the pre-training strategy can improve the performances of both multi-modal ASR systems, especially when the SNR is high. However, the gap between the MMT and multi-modal Conformer baseline is also reduced. Nevertheless, MMT still outperforms or equals to the baseline system.

Similar to the study in Section V-A3, we also studied whether the high frequency components of BC speech are useful to the multi-modal ASR. Specifically, we fed either the low frequency band or full frequency band of the BC speech into the

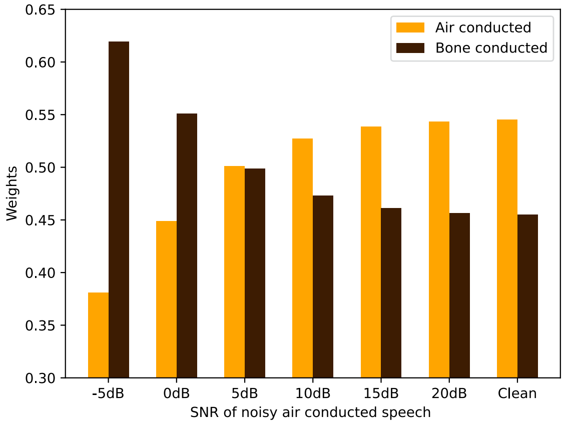
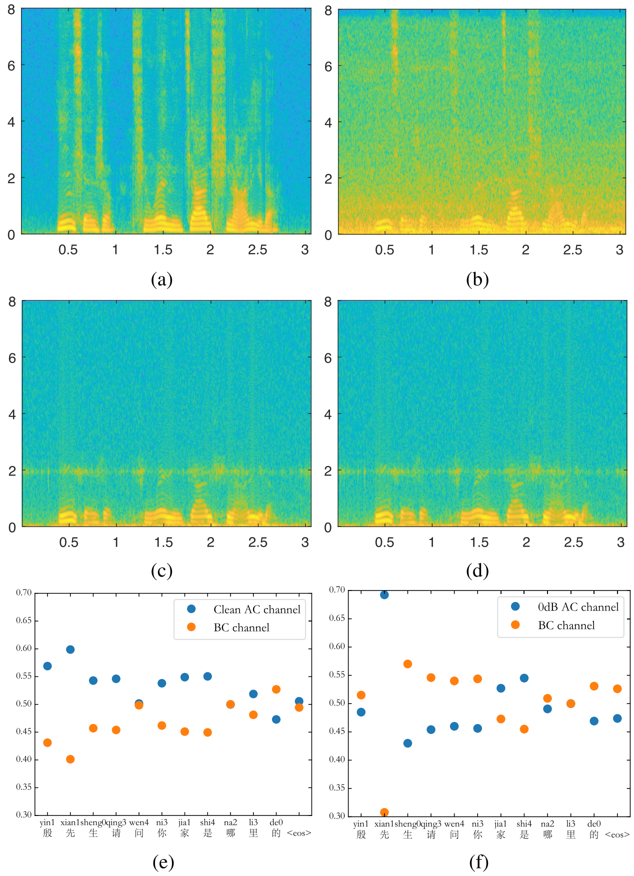


Fig. 9. Statistical mean of the multi-modal attention weights over all test sentences with respect to different SNR of the noisy AC speech.

multi-modal ASR systems. The result is shown in Table VI. We observe that the performance of the systems using the full frequency band is better than that using only the low frequency band. It further manifests that the high frequency components of the BC speech contain useful information for ASR. Interestingly, the proposed MMT method with either the low frequency band or full frequency band of the BC speech, has nearly the same performance when SNR is higher than 10 dB. This phenomenon is caused by the fact that when SNR is high, the high frequency band of the AC speech is clean enough to substitute the contribution of the high frequency band of the BC speech. To contrast with the above phenomenon, we also see that, when the SNR is low, the high-frequency band of the BC speech contributes more to the performance improvement than the low-frequency band of the BC speech. It also demonstrates that the proposed MMT with the channel re-weighting operation can make full use of the information from each modality.

2) Adaptive Weights Produced by MMT: To study how much different speech modalities contribute to the multi-modal ASR, we analyzed the channel weights produced by the MMT module statistically in Fig. 9. From the figure, we observe that the AC speech is the dominant source when the SNR is higher than 5 dB. When the SNR further decreases, the weight of the BC speech gradually increases, which indicates the noise-resisting property of the BC speech. When the SNR equals to 5 dB, the weights of the AC speech and BC speech are similar, which demonstrates that the information in the BC speech is similar to the noisy AC speech at the SNR of 5 dB. This further demonstrates that the BC speech contributes more to the performance improvement of the multi-modal ASR in the low SNR environments than the AC speech.

Besides the statistical analysis, we analyzed the instantaneous channel weights of the AC and BC speech signals of an example sentence in Fig. 10. We observe that the weight of the AC speech is higher than that of the BC speech for most of the words in the clean condition. However, when SNR drops to 0 dB, the weight of the BC speech is higher than the weight of the AC speech. It should be noted that in both clean and noisy conditions, the weight of the AC speech is higher than the weight of the BC speech for the words containing fricative sounds. This is because most of the fricative sounds in BC speech are missed out. This instantaneous example demonstrates that the proposed MMT can adaptively assign the weights according to not only the SNR but also the phoneme characteristics of the speech.



(a) (b) (c) (d) (e) (f)

Fig. 10. Visualized instantaneous weight allocation of the proposed MMT method on a clean utterance and its noisy counterpart at 0 dB. (a) and (b) are spectrograms of the clean and noisy AC speech respectively, where the SNR of the noisy AC speech is 0 dB. (c) and (d) are the spectrograms of the corresponding BC speech respectively. (e) and (f) are the attention weights produced by MMT. The text transcription of this example is“Mr. Yin, where is your home?”. The pronunciations of the corresponding Chinese syllables are shown in the x-axis.

VI. Conclusion

In this paper, a multi-modal Mandarin corpus that contains air- and bone-conducted synchronized speech (ABCS) is built. It is by far the largest corpus for the ASR research based on BC speech. Then, a multi-modal conformer ASR system based on a novel multi-modal transducer is proposed. Specially, the semantic representations of AC and BC speech signals are reweighted and fused together by the MMT module, where the weights are calculated dynamically along the time according to the SNR levels and characteristics of the AC and BC speech. Experimental results show that the BC speech can promote the performance of the proposed ASR system in adverse environments as an auxiliary source of the AC speech. Moreover, the proposed system can effectively integrate the advantages of both the AC and BC speech signals, which leads to significant performance improvement in the low SNR environments over the single-modal system trained with only the AC speech as well as the multi-modal system that simply concatenates the AC and BC speech signals. In addition, it was found that the information in the BC speech is close to the noisy AC speech at the SNR of 5 dB, according to the mutual information analysis and the results of the multi-modal ASR on the ABCS corpus. In the future, we will continue to explore the impact of pre-training strategy on the performance of multi-modal systems.

Acknowledgment

The authors would like to thank SabineTek for providing the air and bone conducted integrated headset.

References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[2] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[3] A. Graves, “Sequence transduction with recurrent neural networks,” *Comput. Sci.*, vol. 58, no. 3, pp. 235–242, 2012.

[4] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4960–4964.

[5] A. Vaswani et al., “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[6] L. Dong, S. Xu, and B. Xu, “Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5884–5888.

[7] A. Gulati et al., “Conformer: Convolution-augmented transformer for speech recognition,” in *Proc. INTERSPEECH*, 2020, pp. 5036–5040.

[8] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, “A cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing DNN-based speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 106–117, 2021.

[9] T. Yoshioka and M. J. Gales, “Environmentally robust ASR front-end for deep neural network acoustic models,” *Comput. Speech Lang.*, vol. 31, no. 1, pp. 65–86, 2015.

[10] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An overview of noise-robust automatic speech recognition,” *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[11] D. Serdyuk, K. Audhkhasi, P. Brakel, B. Ramabhadran, S. Thomas, and Y. Bengio, “Invariant representations for noisy speech recognition,” in *Proc. Neural Inf. Process. Syst. Workshop*, 2016, pp. 1–5.

[12] M. L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7398–7402.

[13] P. Wang, K. Tan, and D. L. Wang, “Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 39–48, 2020.

[14] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, “Adaptive very deep convolutional residual network for noise robust speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 8, pp. 1393–1405, Aug. 2018.

[15] B. Liu, S. Nie, Y. Zhang, D. Ke, S. Liang, and W. Liu, “Boosting noise robustness of acoustic model via deep adversarial training,” in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2018, pp. 5034–5038.

[16] J. Yu et al., “Audio-visual multi-channel integration and recognition of overlapped speech,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2067–2082, 2021.

[17] P. Ma, S. Petridis, and M. Pantic, “End-to-end audio-visual speech recognition with conformers,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 7613–7617.

[18] R. Su, X. Liu, L. Wang, and J. Yang, “Cross-domain deep visual feature generation for mandarin audio–visual speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 185–197, 2020.

[19] G. Sterpu, C. Saam, and N. Harte, “How to teach DNNs to pay attention to the visual modality in speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1052–1064, 2020.

[20] F. Tao and C. Busso, “Gating neural network for large vocabulary audiovisual speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1290–1302, Jul. 2018.

[21] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, “Modality attention for end-to-end audio-visual speech recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6565–6569.

[22] A. H. Abdelaziz, “Comparing fusion models for DNN-based audiovisual continuous speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 475–484, Mar. 2018.

[23] Z. Zhang et al., “Multi-sensory microphones for robust speech detection, enhancement and recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 3, pp. iii—781.

[24] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, “Time-domain multi-modal bone/air conducted speech enhancement,” *IEEE Signal Process. Lett.*, vol. 27, pp. 1035–1039, 2020.

[25] M. Takada, S. Saki, P. L. Tobing, and T. Toda, “Semi-supervised enhancement and suppression of self-produced speech using correspondence between air- and body-conducted signals,” in *Proc. 28th Eur. Signal Process. Conf.*, 2021, pp. 456–460.

[26] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, “SEANet: A multi-modal speech enhancement network,” in *Proc. Interspeech*, 2020, pp. 1126–1130. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1563

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*2015, pp. 5206–5210.

[28] J. Du, X. Na, X. Liu, and H. Bu, “Aishell-2: Transforming mandarin ASR research into industrial scale,” 2018, *arXiv:1808.10583*.

[29] D. Wang and X. Zhang, “THCHS-30: A free chinese speech corpus,” 2015, *arXiv:1512.01882*.

[30] E. Erzin, “Improving throat microphone speech recognition by joint analysis of throat and acoustic microphone recordings,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1316–1324, Sep. 2009.

[31] M. A. T. Turan and E. Erzin, “Source and filter estimation for throat-microphone speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 2, pp. 265–275, Feb. 2016.

[32] D. Shan, X. Zhang, C. Zhang, and L. Li, “A novel encoder-decoder model via NS-LSTM used for bone-conducted speech enhancement,” *IEEE Access*, vol. 6, pp. 62638–62644, 2018.

[33] C. Zheng, L. Xu, X. Fan, J. Yang, J. Fan, and X. Huang, “Dual-path transformer-based network with equalization-generation components prediction for flexible vibrational sensor speech enhancement in the time domain,” *J. Acoust. Soc. Amer.*, vol. 151, no. 5, pp. 2814–2825, 2022.

[34] *Acoustics - Determination of Sound Power Levels and Sound Energy Levels of Noise Sources Using Sound Pressure - Precision Methods for Anechoic Rooms and Hemi-Anechoic Rooms*, ISO Standard 3745:2012, 2012.

[35] M. Nilsson, H. Gustaftson, S. V. Andersen, and W. B. Kleijn, “Gaussian mixture model based mutual information estimation between frequency bands in speech,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2002, vol. 1, pp. I525–I528.

[36] R. E. Bouserhal, T. H. Falk, and J. Voix, “On the potential for artificial bandwidth extension of bone and tissue conducted speech: A mutual information study,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5108–5112.

[37] J. Chen and X.-L. Zhang, “Scaling sparsemax based channel selection for speech recognition with ad-hoc microphone arrays,” in *Proc. INTERSPEECH*, 2021, pp. 291–295.

[38] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.

[39] X. Tan and X.-L. Zhang, “Speech enhancement aided end-to-end multi-task learning for voice activity detection,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2021, pp. 6823–6827.

[40] [Online]. Available: http://www.sound-ideas.com/sound-effects/series-6000-combo-sound-effects.html

[41] A. Varga and H. J. 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 Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[42] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘chime’speech separation and recognition challenge: Dataset, task and baselines,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 504–511.

[43] D. S. Park et al., “SpecAugment: A simple data augmentation method for automatic speech recognition,” in *Proc. Interspeech*, 2019, pp. 2613–2617.

Mou Wang (Graduate Student Member, IEEE) received the B.S. degree in electronics and information engineering in 2016 from Northwestern Polytechnical University, Xi’an, China, where he is currently working the Ph.D. degree in information and communication engineering. From 2021 to 2022, he was a Visiting Ph.D. Student with the National University of Singapore, Singapore. His research interests include machine learning and speech signal processing. He was the recipient of the Excellent Paper Award from International Conference on Ubi-Media Computing and Workshops in 2019. He was awarded Outstanding Reviewer of IEEE Transactions on Multimedia in 2022.

Junqi Chen (Graduate Student Member, IEEE) received the B.S. degree in detection guidance and control technology in 2020 from Northwestern Polytechnical University, Xi’an, China, where he is currently working toward the M.S. degree in signal and information processing. His research interests include deep learning and speech recognition.

Xiao-Lei Zhang (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Tsinghua University, Beijing, China, in 2012. He is currently a Full Professor with the School of Marine Science and Technology, Northwestern Polytechnical University, Xi’an, China. He was a Postdoctoral Researcher with Perception and Neurodynamics Laboratory, The Ohio State University, Columbus, OH, USA. His research interests include speech processing, machine learning, statistical signal processing, and artificial intelligence. He is a Member of SPS and ISCA.

Susanto Rahardja (Fellow, IEEE) received the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore. He is currently a Professor with the Singapore Institute of Technology, Singapore, and a Ph.D. Advisor with Northwestern Polytechnical University, Xi’an, China. He also held other Visiting Professor appointments with several universities including University of Malaya, Kuala Lumpur, Malaysia, University of Eastern Finland, Kuopio, Finland, Zhejiang University, Zhejiang University. He has more than 350 papers and 70 patents worldwide out of which 15 are U.S. patents. His research interests include multimedia, signal processing, wireless communications, discrete transforms, machine learning and signal processing algorithms and implementation. He contributed to the development of a series of audio compression technologies such as Audio Video Standards AVS-L, AVS-2 and ISO/IEC 14496-3:2005/Amd.2:2006, ISO/IEC 14496-3:2005/Amd.3:2006 in which some have been licensed worldwide. He was past Associate Editor for IEEE Transactions on Audio, Speech and Language Processing and IEEE Transactions on Multimedia, past Senior Editor of theIEEE Journal of Selected Topics in Signal Processing, and is currently an Associate Editor for the *Elsevier Journal of Visual Communication and Image Representation*, IEEE Transactions on Multimedia and IEEE Transactions on Industrial Electronics. He was the Conference Chair of 5th ACM SIGGRAPHASIA in 2012 and APSIPA 2nd Summit and Conference in 2010 and 2018 and other conferences in ACM, SPIE and IEEE. Dr Rahardja was the recipient of several honors including the IEE Hartree Premium Award, the Tan Kah Kee Young Inventors’ Open Category Gold award, the Singapore National Technology Award, A*STAR Most Inspiring Mentor Award, Finalist of the 2010 World Technology & Summit Award, the Nokia Foundation Visiting Professor Award, the ACM Recognition of Service Award and the Thousand Talent Plan of People’s Republic of China under Foreign Expert category. Professor Rahardja is a Fellow of the Academy of Engineering, Singapore.