Applied Acoustics 200 (2022) 109058

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Multi-modal speech enhancement with bone-conducted speech in time domain

Mou Wang [a,b], Junqi Chen [a], Xiaolei Zhang [a], Zhiyong Huang [b], Susanto Rahardja [a,c,*]

[a] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China
[b] School of Computing, National University of Singapore, 117417, Singapore
[c] Infocomm Technology Cluster, Singapore Institute of Technology, 138683, Singapore

ARTICLE INFO

Article history:
Received 18 March 2022
Received in revised form 5 September 2022
Accepted 1 October 2022
Available online 13 October 2022

Keywords:
Speech enhancement
Bone conduction
Deep learning
Multi-modal speech

ABSTRACT

Bone-conducted (BC) speech captures speech signals based on the vibrations of a speaker's skull. It is thus not affected by noise sources from environments and hence exhibits better noise-resistance capabilities than air-conducted (AC) speech. Although the quality and intelligibility of the BC speech degrade due to the nature of the solid vibration, BC speech can be utilized as an auxiliary source to jointly improve the performance of speech enhancement. In this paper, we propose an end-to-end multi-modal model for time-domain speech enhancement at low signal-to-noise ratios. The model utilizes both noisy AC speech and synchronized BC speech as the input. It takes an encoder-decoder architecture, where an involution network is used to estimate the mask of clean speech component, and the mask is then applied to remove the noise component. We compared the proposed method with several state-of-the-art multi-modal and single-modal methods on an air- and bone-conducted multi-modal corpus. Experimental results demonstrate that the proposed approach outperforms the comparison methods in terms of the speech quality and intelligibility of the enhanced speech. When applied to speech recognition, the enhanced speech significantly reduces the error rate.

1. Introduction

Speech enhancement aims to improve the intelligibility and quality of a noisy speech signal [1,2]. It is commonly used as a front-end of many speech processing systems, such as robust speech recognition and speech communication. Speech enhancement, especially the deep learning based approaches, has attracted much attention in the past decades. Early deep learning based speech enhancement methods are implemented in the time–frequency domain. Frequently adopted deep models include feedforward neural network [3–5], convolutional neural network [6], and long-short term memory [7]. They estimate the magnitude spectrum of clean speech [5] or its corresponding time–frequency mask [4]. Recently, end-to-end architectures have been proposed to directly estimate the clean speech in the time-domain [8–12].

All of the aforementioned methods were developed with air-conducted (AC) speech. As AC speech is easily corrupted by ambient noise because of the characteristics of the air conduction, AC speech enhancement generally does not perform well in low signal-to-noise ratio (SNR) and non-stationary noise environments. Therefore, several other modalities such as video [6], accelerometer [13] and bone-conducted speech (BC) speech [14–16] had been explored as an additional resource to further improve the target speech. In this paper, we focus on using bone-conducted (BC) speech as an additional modality.

A BC microphone is a kind of skin-attached sensor [17]. It records speech by converting the vibration around a speaker's skull into electrical signals [18]. Thus, the BC microphone has the intrinsic capability of suppressing environmental noise, which is a good physical property for speech enhancement in low SNR conditions. However, BC speech still has the following shortcomings. First and foremost, due to the channel attenuation suffered as the speech propagates through the human tissues, the high-frequency bands of BC speech would experience severe information loss compared to clean AC speech. Although the bandwidth characteristics of different types of BC microphones may be different, only the spectral components below 1 kHz can be recorded effectively in most cases, which significantly degrades speech intelligibility. In addition, due to the friction and resonance between the speaker's skin and the BC microphone, BC speech

* Corresponding author at: School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an710072, Shaanxi, China.
E-mail addresses: wangmou21@mail.nwpu.edu.cn (M. Wang), jqchen@mail.nwpu.edu.cn (J. Chen), xiaolei.zhang@nwpu.edu.cn (X. Zhang), huangzy@comp.nus.edu.sg (Z. Huang), susantorahardja@ieee.org (S. Rahardja).

https://doi.org/10.1016/j.apacoust.2022.109058
0003-682X/

inevitably contains self-generated noise. Finally, the characteristics of BC speech vary with many factors, such as the characteristics differences of speakers like the inherent human tissues, which is a property difficult to be addressed by conventional signal processing.

To utilize the above advantages of BC speech and simultaneously overcome its limitations, there are a few articles and currently the techniques can be categorized into two classes. The first class, named *blind enhancement of BC speech*, replaces the AC microphone with the BC microphone, and then extends the bandwidth of the BC speech by learning a projection from the BC speech to AC speech. The early works are mainly based on transformation filters. According to the type of filtering methods, these methods are categorized into three classes [19], i.e., equalization [20,21], analysis-and-synthesis [22,23], and probabilistic approaches [24,25]. Later, machine learning models were introduced, such as Gaussian mixture models [14,26], deep neural networks [27], and deep denoising autoencoder [28], In the above-mentioned works, the acoustic features are mostly cepstral coefficients [14] or spectrogram [27]. A drawback of the methods is that the BC speech is assumed to be noise-free, which cannot be overcome by the bandwidth extension. The second class is multi-modal speech enhancement where BC speech is considered as an auxiliary resource of AC speech. Representative multi-modal models include Gaussian mixture model [15] and fully convolutional network (FCN) [16]. To our knowledge, this class is still far from being well explored, leaving much room for improvement.

Here we focus on exploiting the deep models for the second class of BC speech processing. Convolution neural network is a class of commonly used deep models, in which temporal convolutional network (TCN) shows good performance in speech separation and enhancement. However, the convolution operator has two limitations. The first limitation is that the convolution operator is difficult to capture long-term information. The other limitation is the inter-channel redundancy between the convolution filters. Recently, *involution* was proposed [29] to overcome the limitations. It was applied to 2-dimensional visual recognition tasks, such as image classification and object detection. However, its effectiveness has not been studied in speech processing yet, where the first limitation of the convolution operator on speech processing may be more challenging than that on image processing.

In this paper, we propose an end-to-end multi-modal involution neural networks (MMINet) for AC and BC joint speech enhancement. It takes AC and BC speech signals in the time domain as the input, and outputs the enhanced speech in the time domain directly. It consists of three modules, i.e. an encoder, a mask estimator, and a decoder. The encoder fuses the AC speech and BC speech to form a feature map. The mask estimator takes the feature map as the input, and outputs an estimated mask of the clean speech. Applying the estimated mask to the feature map produces a new feature map of the enhanced speech. Finally, the decoder is used to convert the feature map of the enhanced speech to the enhanced speech in the time domain.

The novelties and main contributions of this paper are summarized as follows. First, we propose an end-to-end multi-modal speech enhancement method that learns a joint representation of the noisy AC speech and BC speech. Moreover, we develop a novel involution neural network to model the context of speech. Finally, experimental results on a large multi-modal speech database show that the proposed method achieves significant improvement over its single-modal components, state-of-the-art speech enhancement methods [30,11], and a recent multi-modal method [16], in terms of both objective evaluation metrics for speech enhancement and character error rate (CER) for speech recognition.

2. Method

2.1. Problem formulation

Suppose a clean AC speech signal x_a is corrupted by ambient noise n_a, which results in a noisy AC speech signal $\hat{x}_a = x_a + n_a$. A synchronized BC speech x_b is corrupted by self-noise n_b, and is insensitive to n_a. The noisy AC speech and BC speech are paralleled as a binary channel signal $\mathbf{x} = (\hat{x}_a, x_b)$. The end-to-end multi-modal model takes the original speech waves in time domain $\hat{x}_a$ and x_b, and produces the enhanced speech y by a deep neural network $y = f(\hat{x}_a, x_b)$. The model is trained in a supervised way. The training loss function is denoted as $L(x_a, y)$.

2.2. Multi-modal model

As shown in Fig. 1, the proposed model consists of three modules—an encoder, a mask estimator, and a decoder.

2.2.1. Encoder

The encoder E linearly merges the noisy AC speech and BC speech into a single feature map $\mathbf{z}$, which is implemented by 1-D convolution with N kernels.

The input signals of the two channels are segmented synchronically into K overlapped frames $\mathbf{x}_k \in \mathbb{R}^{L\times 2},\ k = 1, 2, \ldots, K$, where L is the dimension of each frame and the frame shift is $L/2$. Zero padding is used to ensure the number of dimensions of all frames to be the same. The encoder is a linear transform, i.e.:

$$\mathbf{z} = E(\mathbf{x}) = \mathbf{U}\mathbf{X}, \tag{1}$$

where $\mathbf{U} \in \mathbb{R}^{N\times 2\times L}$ consists of N learnable kernels, and $\mathbf{X} \in \mathbb{R}^{L\times 2\times K}$ contains all input frames. In this work, N is set to 256, L is set to 16 with frame shift of 8.

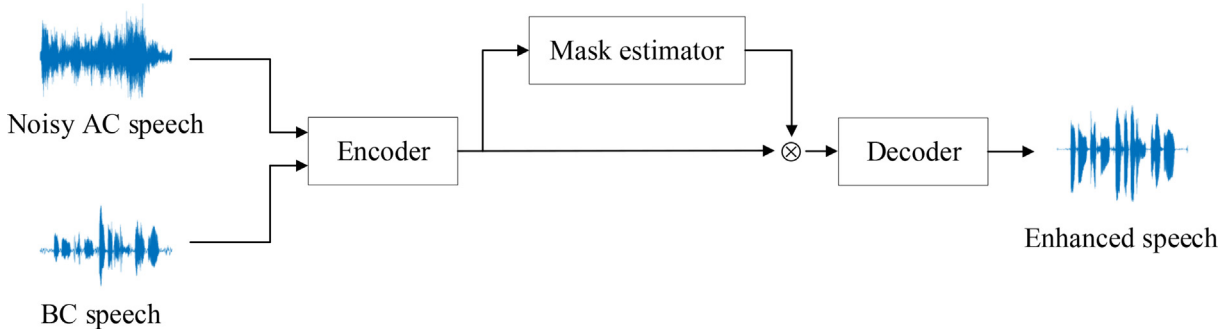

Fig. 1. The block diagram of the proposed MMINet method.

2.2.2. Mask estimator

The mask estimator M learns a ratio mask from the input feature map $\mathbf{z}$ from the encoder, i.e. $\mathbf{m} = M(\mathbf{z})$ whose elements m are non-negative. The estimated clean speech component $\mathbf{c}$ can be calculated by:

$$\mathbf{c} = \mathbf{z} \odot \mathbf{m}, \tag{2}$$

where $\odot$ is the element-wise multiplication, and $\mathbf{c} \in \mathbb{R}^{N \times K}$.

The detailed architecture of the mask estimator is shown in Fig. 2. The mask estimator is implemented by an involution network. The original involution network was proposed for visual tasks, therefore its involution operator is 2-dimensional. In this paper, we developed 1-D involution operation for speech tasks.

The procedure of 1-D involution is shown in Fig. 3. Unlike the convolution operator, involution operation generates its kernels conditioned on the input tensor. If we denote the kernel generation function is ϕ, then the involution kernel ψ can be described as

$$\psi = \phi(v), \tag{3}$$

where v is the local sequence. In this paper, we use the 1-D depthwise separable convolution [31] as the kernel generation function. Then, the involution kernel is applied to the input tensor through Multiply–Add operation which consists of the following two steps. First, the input tensor is multiplied by the involution kernel across the channel dimension; then, the output of each kernel is aggregated along the sequence dimension. After the above involution operation, the input tensor is transformed into new feature maps.

As shown in Fig. 2, the mask estimator is built on the involution operator. The structure of the mask estimator is similar to the temporal convolutional network (TCN) architecture in [32,33]. It consists of Q stacked 1-D involution blocks, where each block generates H learnable kernels. The structure of each 1-D involution block is shown in Fig. 4. It contains a 1-D involution operator followed by the PReLU activation and layer normalization. In addition, we apply a residual path and a skip-connection path in each block. In order to make the mask estimator having a sufficiently large temporal context window, we introduce dilation into the involution operator and increase the dilation factors exponentially. The whole mask estimator repeats the stacked blocks for R times, followed by a convolution layer and ReLU activation function. In this paper, Q is set to 8, H is set to 256, and R is set to 3.

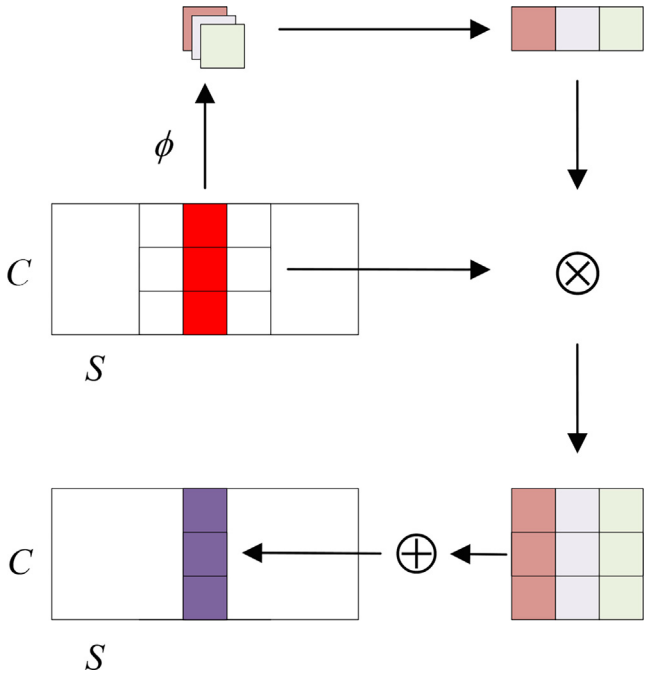

Fig. 3. Illustration of 1-D involution. The symbol ϕ indicates the kernel generation function. The symbol $\otimes$ indicates multiplication broadcast across channels C. The symbol $\oplus$ indicates the aggregation within local sequence.

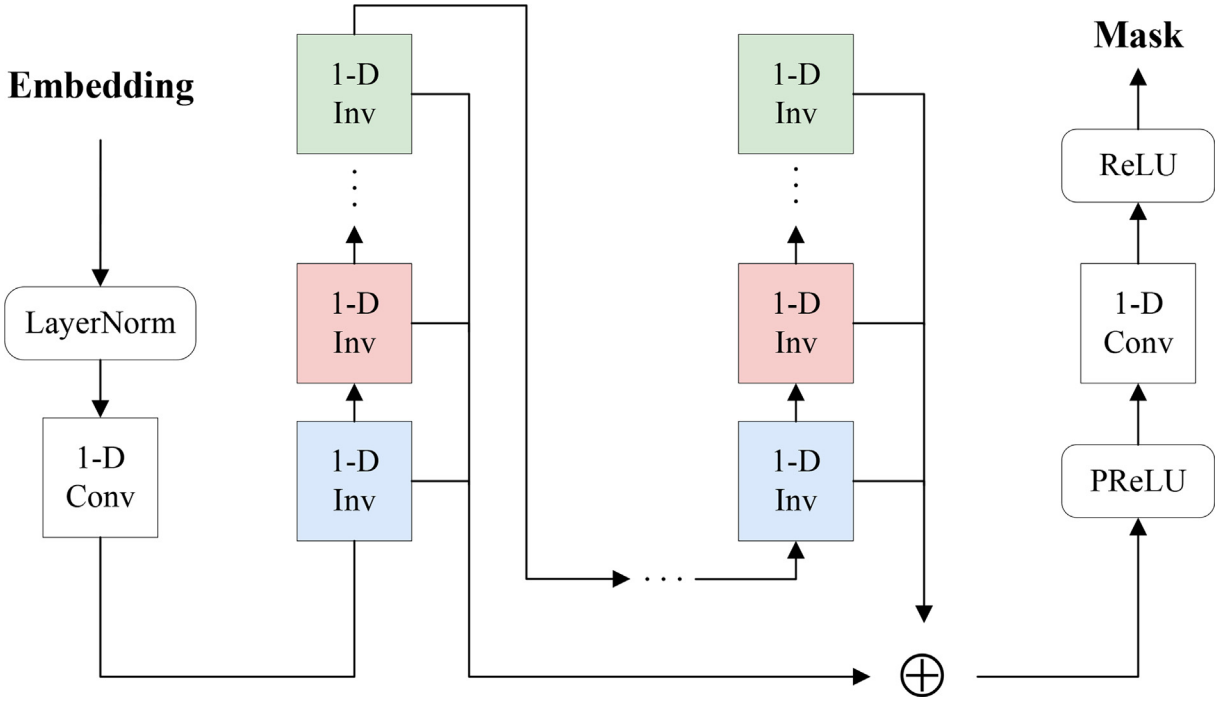

Fig. 2. The block diagram of mask estimator in MMINet. The term "1-D Conv" refers to the 1-dimensional convolution. The term "1-D Inv" refers to the 1-dimensional involution block. The term "LayerNorm" refers to layer normalization.

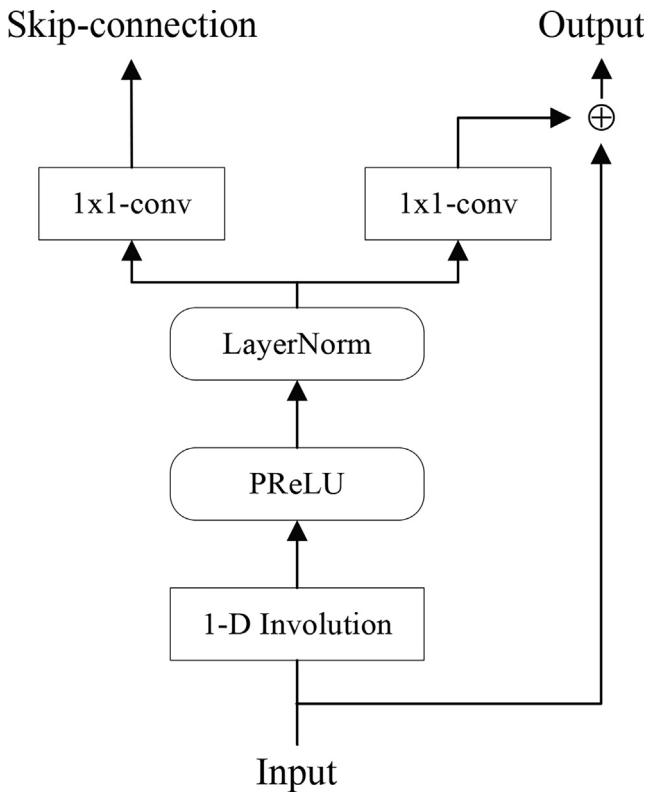

Fig. 4. 1-D Involution block in the mask estimator. The term "1 × 1-conv" refers to convolution layer with a kernel size of 1 × 1. "LayerNorm" refers to layer normalization.

2.2.3. Decoder

The decoder D linearly transforms the feature map $\mathbf{c}$ to a pack of single-channel waveforms $\mathbf{Y} \in \mathbb{R}^{L\times K}$ by a linear transform:

$$\mathbf{Y} = D(\mathbf{c}) = \mathbf{V}\mathbf{c}, \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{L\times N}$. Finally, the enhanced speech y can be obtained from $\mathbf{Y}$ with an overlap-add operation.

2.3. Training objective

Because of the mismatch of the amplitude gains between the estimated signal y and its clean reference x_a, the training objective of the proposed system maximizes the scale-invariant source-to-noise ratio (SI-SNR), which has commonly been used in end-to-end speech source separation [32]. It is defined as:

$$s = \frac{\langle y, x_a\rangle}{\|x_a\|^2}, \tag{5}$$

$$\hat{n} = y - s, \tag{6}$$

$$\text{SI}-\text{SNR} = 10\log_{10}\frac{\|s\|^2}{\|\hat{n}\|^2}, \tag{7}$$

where $\langle\cdot\rangle$ and $\|\cdot\|^2$ refer to the inner product and signal power operators respectively.

3. Experiments

3.1. Datasets

We collected AC and BC speech synchronously in an anechoic chamber. The size of anechoic chamber is $11.8 \times 4.2 \times 7.6\ \text{m}^3$. All the surfaces in the interior part of the chamber are made of sound-absorbing materials which are complied to the ISO 3745 standards [34]. The scripts were selected from 20,000 daily dialogues, and read by 100 native Chinese speakers (50 males and 50 females) wearing one headset integrating AC and BC microphones. Eventually, we collected a multi-modal speech corpus of about 42 h, with each speaker contributing about 25 min. The speech was recorded at a sampling rate of 44.1 kHz, and further downsampled to 16 kHz. The durations of utterances range from 1 to 5 s. Among the speakers, 80 speakers were used for training, 10 speakers for validating, and the remaining 10 speakers for testing.

MUSAN [35] and NOISEX-92 [36] corpora were used as noise sources of the noisy AC speech. Some noise in MUSAN was sampled from Freesound. The noise part of MUSAN was used for training. For each noise recording in NOISEX-92, half was used for validation, and the other half for testing. The noise signals were also downsampled to 16 kHz. The noisy AC utterances were constructed by corrupting the clean AC utterances with randomly selected noise segments. The SNR levels of the noisy utterances for training and validation were selected randomly from −15 dB to 5 dB. The SNR levels of the test utterances were set to −15 dB, −10 dB, −5 dB, 0 dB, and 5 dB respectively.

3.2. Experimental configurations

For all experiments, we trained the models for 30 epoches with the AdamW optimizer and a batch size of 12. The learning rate was initialized to 0.001 and was halved if the accuracy of the validation set was not improved in 4 consecutive epochs. Early stopping was applied if the performance of the model on the validation set was not improved for 10 consecutive epochs. We picked the model that has the best performance on the validation set for evaluation.

3.3. Metrics

To evaluate the quality of the enhanced speech objectively, several objective metrics were used, including narrow-band perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI), and extended STOI (ESTOI). A PESQ score ranges from −0.5 to 4.5. It measures the overall speech quality. It has a high correlation with a subjective evaluation score. STOI and ESTOI scores range from 0 to 1. They are highly relevant to the human speech intelligibility. For each of the three metrics, the higher the score is, the better the speech quality or intelligibility will be. The clean AC speech is used as a reference to calculate the scores.

To further evaluate the performance of MMINet in applications, we applied the enhanced speech to automatic speech recognition (ASR). The speech recognition system for evaluation was built on the conformer [37]. The system was first pretrained using the ESPNET TOOLKIT on a public Mandarin speech corpus AISHELL-2 [38], then fine-tuned by the clean AC speech of our datasets. The training set of AISHELL-2 contains about 1000 h of clean AC speech data recorded from more than 1900 speakers. The acoustic feature is 80-channel log-mel filterbank coefficients computed with a frame length of 25 ms and a frame shift of 10 ms. Because AISHELL-2 is a Mandarin speech corpus, character error rate (CER) was used as the evaluation metric. The lower the CER is, the better ASR performance is. The conformer achieves CERs of 9.23% and 12.9% on the clean test set of AISHELL-2 and our dataset respectively.

3.4. Comparison methods

We compared the proposed MMINet with its AC and BC component, where the component networks are labeled as the AC involution network (AC-INet) and BC involution network (BC-INet). We

also compared with multi-modal fully convolutional network (MFCN) with late fusion strategy [16], which is state-of-the-art air- and bone-conducted speech enhancement method. It first pre-trains two FCN models with the AC and BC speech respectively. Then, the outputs of the two pre-trained FCN are concatenated. Another compact FCN model with 1D convolutional layers is applied to the concatenated feature. The setting of MFCN in the comparison was the same as [16].

We also compared with some representative single-modal speech enhancement methods, including AC-based and BC based. For AC-based methods, the first one is Improved Minima Controlled Recursive Averaging (IMCRA) [30], which is a dominant traditional signal precessing method for speech enhancement in industry. Another one is the deep complex convolution recurrent network (DCCRN) [11], which is one of the state-of-the-art deep models for speech enhancement. For BC-based methods, we compared two blind enhancement methods of BC speech, i.e, transformation filter based equalization [21], and FCN-based method [16]. Finally, we also evaluate the speech quality of the AC speech and BC speech directly as a reference.

3.5. Results

Table 1 lists the results of the competing methods. From the table, we find that the quality of the BC speech drops due to the self-generated noise and the loss of the high-frequency information. Although the BC speech sounds muffled, its speech intelligibility is still acceptable. Therefore, the PESQ of BC speech is not low relatively. However, when applying the BC speech to speech recognition, the CER increases to as high as 0.88 indicating that the high frequency component of speech is important for ASR and the ASR system trained with the AC speech cannot be generalized to the BC speech at all.

As for the blind enhancement of BC speech, it is observed that transformation filter based equalization can slightly improve the performance because the low-frequency components are equalized. However, the effective bandwidth of BC speech in our corpus is quite narrow, which makes the high-frequency components cannot be reconstructed. In addition, the performance of BC-FCN is lower than original BC speech, which is coincident with [16]. The reason is the FCN architecture in [16] cannot restore the high-quality speech using BC speech. Likewise, BC-INet does not have better performance than original BC speech, because the proposed involution neural network produces a mask, which can mask some information on the feature map output from encoder, and cannot compensate the high-frequency components of BC speech.

As for the AC speech enhancement, we find that IMCRA can improve the speech quality in terms of PESQ and CER, however, it makes the speech intelligibility drop in terms of STOI and ESTOI, which indicates that the performance is limited in low SNR conditions. DCCRN outperforms the aforementioned baselines significantly in all metrics, which shows the advantage of deep learning on speech enhancement. However, the proposed AC-INet has better performance than DCCRN, which proves that involution is more suitable than convolution in modeling speech context. Finally, the proposed MMINet further outperforms AC-INet by a large margin.

From Table 1, we can also observe that the proposed MMINet significantly outperforms FCN methods in all metrics. Both FCN and MMINet are multi-modal methods in time-domain. However, their architectures and training methods are different, which accounts for the advantage of MMINet over FCN. Specifically, FCN uses a stack of convolution layers to predict the clean AC speech from the noisy AC speech and BC speech. MMINet uses an encoder-decoder architecture and estimates a mask for the feature map of the clean AC speech. Besides, FCN method needs to pretrain the two FCN branches with the AC and BC speeches respectively, which decreases the interaction between the AC and BC speeches information. On the contrary, MMINet extracts the mask from the joint embedding space of the AC and BC speeches.

Table 2 lists the detailed performance of the MMINet, its AC-INet component, and noisy speech in all five SNR levels. From the table, we find that both AC-INet and MMINet significantly improve the speech quality and ASR performance over the noisy AC speech at all levels of SNR. Comparing to AC-INet, the proposed MMINet achieves an average improvement of 0.8 in terms of PESQ, 0.11 in STOI, 0.15 in ESTOI, and about 25% in CER, highlighting the importance of leveraging the BC speech in low SNR environments. In addition, we can find that the advantage of multi-modal method becomes more salient as the SNR decreases, given the BC speech is more noise-resistant.

To demonstrate the complementary property between the BC speech and AC speech, as well as the advantage of the proposed method, we further visualize the magnitude spectrograms of the clean AC speech, BC speech, noisy AC speech and enhanced speech produced by MMINet in Fig. 5. From the figure, we first find that the BC speech only captures low-frequency bands, while the information at the frequency bands higher than 1 kHz decays severely and even completely lost. The BC speech has sharp self-generated noise, which is mainly at low-frequency bands and at a narrow-band of about 2 kHz. The noisy AC speech, which is corrupted by

Table 1
Average performance of the proposed MMINet and competing methods over the SNRs of $[-15, -10, -5, 0, 5]$ dB.

	PESQ	STOI	ESTOI	CER
Clean AC speech	-	-	-	0.134
Noisy AC speech	1.36	0.64	0.46	0.793
IMCRA [30]	1.60	0.50	0.36	0.779
DCCRN [11]	2.37	0.76	0.65	0.619
AC-INet (proposed)	2.49	0.80	0.69	0.521
BC speech	2.37	0.68	0.57	0.880
Filter method [21]	2.41	0.70	0.58	0.823
BC-FCN [16]	2.00	0.66	0.47	0.932
BC-INet (proposed)	2.39	0.68	0.56	0.845
MFCN [16]	1.77	0.68	0.54	0.761
MMINet (proposed)	**3.29**	**0.91**	**0.84**	**0.273**

Table 2
Performance comparison between MMINet and its AC-only component at five SNR levels in terms of four evaluation metrics.

	Noisy AC speech				AC-INet				MMINet			
	PESQ	STOI	ESTOI	CER	PESQ	STOI	ESTOI	CER	PESQ	STOI	ESTOI	CER
SNR5	1.99	0.82	0.70	0.604	3.29	0.93	0.87	0.243	3.63	0.94	0.90	0.163
SNR0	1.62	0.74	0.58	0.739	2.93	0.89	0.82	0.355	3.46	0.93	0.87	0.211
SNR-5	1.30	0.64	0.45	0.843	2.49	0.83	0.73	0.531	3.28	0.90	0.84	0.285
SNR-10	1.03	0.54	0.33	0.873	2.06	0.74	0.60	0.671	3.13	0.89	0.82	0.319
SNR-15	0.88	0.45	0.22	0.904	1.66	0.60	0.41	0.805	2.95	0.87	0.79	0.389
average	1.36	0.64	0.46	0.793	2.49	0.80	0.69	0.521	3.29	0.91	0.84	0.273

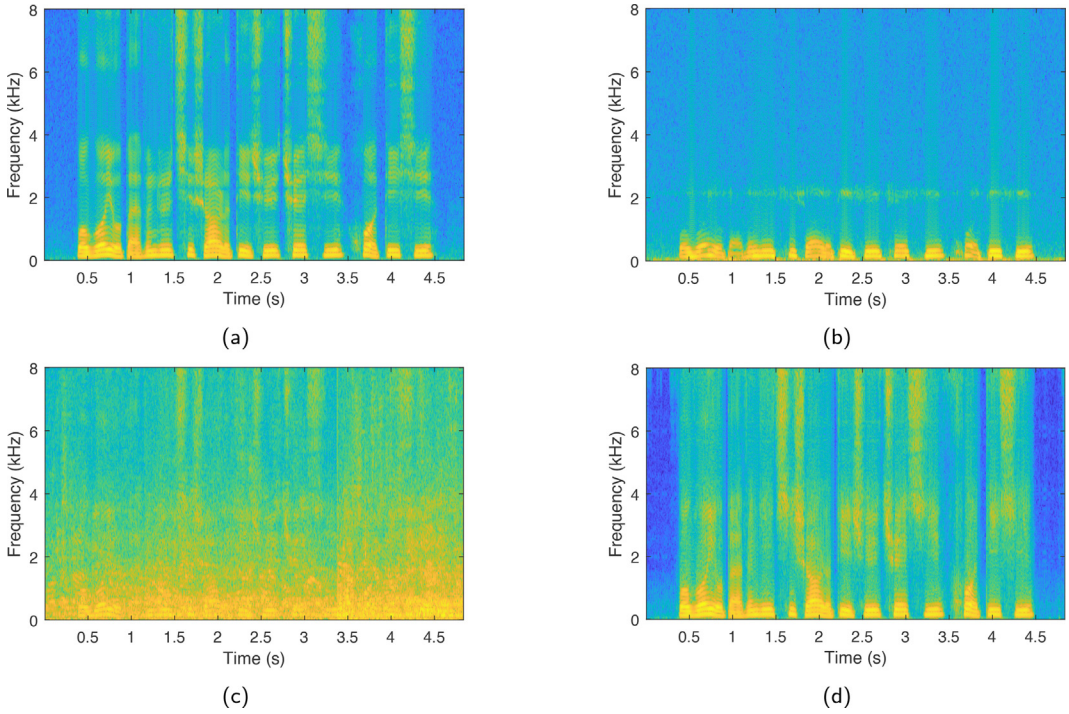

Fig. 5. (a) and (b) are spectrograms of the clean AC and synchronized BC speech respectively. (c) is spectrogram of the corresponding noisy AC speech interfered by babble noise, where the SNR is −5 dB. (d) are the spectrogram of enhanced speech with MMINet.

the babble noise at the SNR of −5 dB, does not have a clear spectrogram structure compared to the clean AC speech. Finally, the enhanced speech produced by the proposed MMINet not only suppresses the noise in both the noisy AC speech and BC speech at low SNR condition, but also recovers the information at the high-frequency bands to some extent.

4. Conclusions

In this paper, an end-to-end multi-modal speech enhancement method named MMINet is proposed. The proposed model can utilize both the noisy AC speech and its synchronized BC speech to obtain the enhanced speech at low SNR environment. Specifically, the mask estimator of our model is built with the 1-D involution network. To the best of our knowledge, it is the first time that involution operation is applied on speech to explore the context. We evaluated our model on an air- and bone-conducted multi-modal speech corpus. Experimental results demonstrate BC speech can be used as an auxiliary source to improve the speech quality and intelligibility as well as the application to ASR. Moreover, the proposed MMINet can effectively utilize the information of AC and BC speech, and outperforms the state-of-the-art methods.

CRediT authorship contribution statement

Mou Wang: Conceptualization, Methodology, Software, Writing - original draft. **Junqi Chen:** Software, Validation. **Xiaolei Zhang:** Supervision, Writing - review & editing. **Zhiyong Huang:** Supervision, Writing - review & editing. **Susanto Rahardja:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work of M. Wang was supported in part by China Scholarship Council and the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University. The work of S. Rahardja was supported in part by the Overseas Expertise Introduction Project for Discipline Innovation (111 project: B18041).

References

[1] Wang D, Chen J. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans Audio Speech Language Process 2018;26 (10):1702–26. https://doi.org/10.1109/TASLP.2018.2842159.
[2] Li A, Zheng C, Zhang L, Li X. Glance and gaze: A collaborative learning framework for single-channel speech enhancement. Appl Acoust 2022;187:. https://doi.org/10.1016/j.apacoust.2021.108499108499.
[3] Reddy H, Kar A, Østergaard J. Performance analysis of low complexity fully connected neural networks for monaural speech enhancement. Appl Acoust 2022;190:. https://doi.org/10.1016/j.apacoust.2022.108627108627.
[4] Wang Y, Narayanan A, Wang D. On training targets for supervised speech separation. IEEE/ACM Trans Audio Speech Lang Process 2014;22(12):1849–58. https://doi.org/10.1109/TASLP.2014.2352935.
[5] Xu Y, Du J, Dai L, Lee C. A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Trans Audio Speech Lang Process 2015;23 (1):7–19. https://doi.org/10.1109/TASLP.2014.2364452.
[6] Hou J, Wang S, Lai Y, Tsao Y, Chang H, Wang H. Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Trans Emerg Top Comput Intell 2018;2(2):117–28. https://doi.org/10.1109/TETCI.2017.2784878.
[7] Cui X, Chen Z, Yin F. Multi-objective based multi-channel speech enhancement with bilstm network. Appl Acoust 2021;177:. https://doi.org/10.1016/j.apacoust.2021.107927107927.
[8] Pandey A, Wang D. A new framework for cnn-based speech enhancement in the time domain. IEEE/ACM Trans Audio Speech Lang Process 2019;27 (7):1179–88. https://doi.org/10.1109/TASLP.2019.2913512.
[9] Pandey A, Wang D. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom; 2019. p. 6875–9.
[10] Pandey A, Wang D. Dense cnn with self-attention for time-domain speech enhancement. IEEE/ACM Trans Audio Speech Lang Process 2021;29:1270–9. https://doi.org/10.1109/TASLP.2021.3064421.
[11] Hu Y, Liu Y, Lv S, Xing M, Zhang S, Fu Y, Wu J, Zhang B, Xie L. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech

Enhancement. In Proc. Interspeech 2020; 2020. pp. 2472–2476. doi:10.21437/Interspeech.2020-2537.
[12] Tan X, Zhang X-L. 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). p. 6823–7. https://doi.org/10.1109/ICASSP39728.2021.9414445.
[13] Tagliasacchi M, Li Y, Misiunas K, Roblek D. SEANet: A Multi-Modal Speech Enhancement Network. In Proc. Interspeech 2020; 2020. pp. 1126–1130. doi:10.21437/Interspeech.2020-1563.
[14] Zhang Z, Liu Z, Sinclair M, Acero A, Deng L, Droppo J, Huang X, Zheng Y. Multi-sensory microphones for robust speech detection, enhancement and recognition. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3; 2004. pp. iii–781.
[15] Hershey J, Kristjansson T, Zhang Z. Model-based fusion of bone and air sensors for speech enhancement and robust speech recognition. Workshop on Statistcal and Perceptual Audio Processing 2004:139.
[16] Yu C, Hung K, Wang S, Tsao Y, Hung J. Time-domain multi-modal bone/air conducted speech enhancement. IEEE Signal Process Lett 2020;27:1035–9. https://doi.org/10.1109/LSP.2020.3000968.
[17] Chen J, Wang M, Zhang X-L, Huang Z, Rahardja S. End-to-end multi-modal speech recognition with air and bone conducted speech. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). p. 6052–6. https://doi.org/10.1109/ICASSP43922.2022.9747306.
[18] Zheng C, Yang J, Zhang X, Sun M, Yao K. Improving the spectra recovering of bone-conducted speech via structural similarity loss function. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China 2019:1485–90.
[19] Shin HS, Kang H-G, Fingscheidt T. Survey of speech enhancement supported by a bone conduction microphone. In Speech Communication; 10. ITG Symposium; 2012. pp. 1–4.
[20] Shimamura T, Tamiya T. A reconstruction filter for bone-conducted speech. In 48th Midwest Symposium on Circuits and Systems, 2005, vol. 2; 2005. pp. 1847–1850. doi:10.1109/MWSCAS.2005.1594483.
[21] Kondo K, Fujita T, Nakagawa K. On equalization of bone conducted speech for improved speech quality. 2006 IEEE International Symposium on Signal Processing and Information Technology 2006:426–31. https://doi.org/10.1109/ISSPIT.2006.270839.
[22] Vu TT, Unoki M, Akagi M. An lp-based blind model for restoring bone-conducted speech. 2008 Second International Conference on Communications and Electronics 2008:212–7. https://doi.org/10.1109/CCE.2008.4578960.
[23] Vu TT, Kimura K, Unoki M, Akagi M. A study on restoration of bone-conduted speech with mtf-based and lp-based models. J Signal Process 2006;10(6):407–17.
[24] Liu Z, Zhang Z, Acero A, Droppo J, Huang X. Direct filtering for air- and bone-conductive microphones. In IEEE 6th Workshop on Multimedia Signal Processing, 2004; 2004. pp. 363–366. doi:10.1109/MMSP.2004.1436568.
[25] Subramanya A, Droppo J, Acero A, Zhang Z, Liu Z. A graphical model for multi-sensory speech processing in air-and-bone conductive microphones. In: Proc. of the Interspeech Conference, proc. of the interspeech conference Edition. International Speech Communication Association; 2005.
[26] Tugtekin Turan MA, Erzin E. Enhancement of throat microphone recordings by learning phone-dependent mappings of speech spectra. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. p. 7049–53.
[27] Shan D, Zhang X, Zhang C, Li L. A novel encoder-decoder model via ns-lstm used for bone-conducted speech enhancement. IEEE Access 2018;6:62638–44. https://doi.org/10.1109/ACCESS.2018.2873728.
[28] Liu H, Tsao Y, Fuh C. Boneconducted speech enhancement using deep denoising autoencoder. Speech Commun 2018;104:106–12.
[29] Li D, Hu J, Wang C, Li X, She Q, Zhu L, Zhang T, Chen Q. Involution: Inverting the inherence of convolution for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 12321–30.
[30] Cohen I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans Speech Audio Process 2003;11(5):466–75. https://doi.org/10.1109/TSA.2003.811544.
[31] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861.
[32] Luo Y, Mesgarani N. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation. IEEE/ACM Trans Audio Speech Lang Process 2019;27(8):1256–66. https://doi.org/10.1109/TASLP.2019.2915167.
[33] Pan N, Wang Y, Chen J, Benesty J. A single-input/binaural-output antiphasic speech enhancement method for speech intelligibility improvement. IEEE Signal Process Lett 2021;28:1445–9. https://doi.org/10.1109/LSP.2021.3095016.
[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; 2012.
[35] Snyder D, Chen G, Povey D. MUSAN: A Music, Speech, and Noise Corpus, arXiv preprint arXiv:1510.08484v1.
[36] Varga A, Steeneken HJ. 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 1993;12(3):247–51. https://doi.org/10.1016/0167-6393(93)90095-3.
[37] Gulati A, Qin J, Chiu C, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y, et al. Conformer: Convolution-augmented transformer for speech recognition, arXiv preprint arXiv:2005.08100.
[38] Bu H, Du J, Na X, Wu B, Zheng H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In: Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), in: 2017 20th. p. 1–5.