Time-Domain Multi-modal Bone/air Conducted Speech Enhancement

Cheng Yu, Kuo-Hsuan Hung, Syu-Siang Wang, Yu Tsao, *Member, IEEE*, and
Jeih-weih Hung, *Member, IEEE*

Abstract—Previous studies have proven that integrating video signals, as a complementary modality, can facilitate improved performance for speech enhancement (SE). However, video clips usually contain large amounts of data and pose a high cost in terms of computational resources and thus may complicate the SE system. As an alternative source, a bone-conducted speech signal has a moderate data size while manifesting speech-phoneme structures, and thus complements its air-conducted counterpart. In this study, we propose a novel multi-modal SE structure in the time domain that leverages bone- and air-conducted signals. In addition, we examine two ensemble-learning-based strategies, early fusion (EF) and late fusion (LF), to integrate the two types of speech signals, and adopt a deep learning-based fully convolutional network to conduct the enhancement. The experiment results on the Mandarin corpus indicate that this newly presented multi-modal (integrating bone- and air-conducted signals) SE structure significantly outperforms the single-source SE counterparts (with a bone- or air-conducted signal only) in various speech evaluation metrics. In addition, the adoption of an LF strategy other than an EF in this novel SE multi-modal structure achieves better results.

Index Terms—multi-modal, bone/air-conducted signals, speech enhancement, fully convolutional network, fusion strategy

I. INTRODUCTION

Speech enhancement (SE) aims to improve the speech quality and the intelligibility in noisy environments and has been widely applied in numerous speech-related applications, such as automatic speech recognition [1], [2], [3], speaker recognition [4], [5], [6], emotion recognition [6], and assistive hearing devices [7], [8], to improve the system robustness against environmental noise. Recently, deep learning-based models have been popularly used as a fundamental tool in SE systems [9], [10], [11], [12], [13], [14], [15]. For most of these systems, the noise-corrupted speech is enhanced in the frequency domain using a deep-learning-based model with supervised learning. More specifically, the SE systems estimate the clean magnitude spectra or the corresponding ratio mask from the input noisy speech [16], [17], [18], [19], [20]. In addition to applying SE in the frequency domain, a fully convolutional network (FCN) based SE approach [21], [22], [23] has been proposed to directly estimate a time-domain waveform-mapping function to restore clean speech. A major advantage of this approach is the potential to circumvent the imperfect phase estimation issue. Additionally, the FCN-based SE system employs a smaller number of parameters than fully-connected neural-network-based models, making it a suitable model to operate in mobile devices.

Recent studies have shown that SE systems can leverage complementary information, such as visual cues, as an aux-iliary input to achieve better enhancement performance [24], [25], [26], [27]. As an alternative complementary information to visual cues, signal captured from a bone-conducted microphone (BCM) has the inherent capability to suppress air background noise to reduce the noise commonly recorded by an air-conducted microphone (ACM) [28], [29], [30], [31]. However, unlike an ACM-recorded speech signal, a BCM-captured waveform, in which the pronounced utterance is recorded through the vibrations from the speakers skull, may lose some high frequency components from the original spoken speech [28]. Several filtering-based and probabilistic solutions have been proposed to convert the BCM-recorded sound to its ACM version [32], [33], [34]. The authors in [35] proposed a reconstruction filter, which uses the long-term spectra of the speech, to perform the conversion. Meanwhile, some approaches have been proposed to combine ACM- and BCM-recorded sounds in the frequency domain with a linear transformation for SE tasks [36], [37].

In this study, we propose a novel FCN-based SE method that leverages the acoustic characteristics of the signals recorded by a BCM in terms of the ensemble learning approach. In the proposed algorithm, the BCM-recorded waveform is used as an auxiliary input (with noise robustness property while low precision in the high frequency region), and combined with ACM to carry out SE in the time domain. Experiments were conducted on a personalized SE scenario and results show that the newly presented method yields significant improvements in terms of standardized objective metrics over the noisy baseline. These results clearly indicate that adequately integrating BCM- and ACM-recorded signals can help FCN models learn detailed harmonic speech structures, resulting in enhanced signals of high quality and intelligibility.

II. RELATED WORKS

We briefly review some novel studies that benefit a waveform-based SE task and/or exploit various signal sources.

A. Deep learning-based model

Employing a deep learning-based model structure is usually a main procedure of a state-of-the-art SE technique. In [21], [22], an FCN model was used to directly process the input time-domain waveform. By contrast, in the studies presented in [38], [39], waveform-wise enhancement was conducted using a convolutional neural network (CNN) structure [40]. In comparison with a CNN, FCN only consists of convolutional layers, which can efficiently store information from the receptive

fields of filters in each layer while possessing much fewer parameters. In addition, an FCN has been shown to outperform the conventional deep neural network (DNN), which consists of densely connected layers, for use in SE.

B. BCM/ACM conversion

A straightforward method that can collect less distorted speech signals is to apply noise-resistant recording devices. As mentioned before, a BCM records signals through bone vibrations and is thus less sensitive to air background noise in comparison with an ACM. However, the BCM-recorded speech signals often suffer from a loss of high acoustic-frequency components, and this issue was addressed and partially alleviated through the BCM-to-ACM conversion technique applied in SE tasks [35], [41], [42].

C. Multi-modalality

Another promising direction for waveform-based SE is to adopt a multi-modal system that extracts clean-speech information from various signal sources. In [43], the authors proposed the use of audio-visual multi-modality in various speech-processing fields, and showed that integrating video modality with speech benefits various speech processing behaviors. The audio-visual system presented in [44] combines audio with lip-motion clips to access more bio-information and thereby promotes the SE performance. Despite the success of using audio-visual multi-modality for SE tasks, the corresponding high computational cost incurred and large amount of data storage required are obstacles for devices with limited computational resources.

III. PROPOSED METHOD

In this section, we present a novel time-domain SE scenario that adopts multiple FCN models to fulfill the SE task. In particular, this novel scenario possesses multi-modal characteristics because it uses both BCM- and ACM-recorded signals. As is well known, the ACM-recorded signals contain complete (full acoustic-band) clean-speech information but are vulnerable to background noise, whereas the BCM-recorded signals possess a higher SNR but lack high acoustic-frequency components. Please note in this study, the ACM-recorded signal is the major source for the SE task while the BCM-recorded signal serves as the auxiliary source. Hence, we believe that, if arranged appropriately, the two types of signals can complement each other when applied to SE.

A. The overall SE structure

A flowchart of the newly presented SE scenario is depicted in Fig. 1, which indicates two different arrangements for the input BCM- and ACM-recorded signals. These two arrangements are created by either an early-fusion (EF) strategy or a late-fusion (LF) strategy. The difference between the EF and LF is in the stage during which the BCM- and ACM-wise representations are merged. In other words, the EF strategy suggests integrating BCM- and ACM-recorded raw waveforms at the very beginning of the SE framework to serve as the initial input, whereas in the LF strategy, the two signal sources
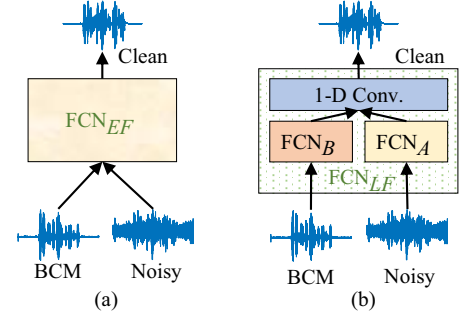

Fig. 1. Detailed structures of (a) EF strategy, FCN_{EF}, and (b) LF strategy, FCN_{LF}.

are first individually processed, and the respective outputs are then brought together for a subsequent enhancement. To the best of our knowledge, determining which strategy is better for a multi-modal analysis mostly depends on the data types and tasks associated with the given multimedia dataset. In the following sections, we provide descriptions regarding the EF and LF arrangements shown in Fig. 1 in more detail.

B. Early-fusion-strategy structure

Following the EF strategy, the waveform-level BCM- and ACM-recorded noisy signals for each utterance in the training set are directly concatenated to form an input vector, which is used to train an FCN to approximate its noise-free ACM-recorded counterpart. The corresponding input-output relationship is therefore described as follows:

$$s_{EF}[n] = \text{FCN}_{EF}\{x^{(A)}[n], x^{(B)}[n]\}, \qquad (1)$$

where $x^{(A)}[n]$ and $x^{(B)}[n]$ with respect to the time index, n, represent the ACM- and BCM-recorded signals corresponding to an arbitrary noisy utterance; $\text{FCN}_{EF}\{.\}$ denotes the FCN model operation used, and $s_{EF}[n]$ is the enhanced signal expected to approximate the noise-free version of $x^{(A)}[n]$.

In this study, FCN_{EF} is composed of seven hidden convolutional layers, each layer contains 30 kernels of size 55. In addition, to examine the impact of BCM, we construct another FCN model that is close to the FCN_{EF} while it only adopts the ACM channel. Evaluations for these models will be described in Sec. IV.

C. Late-fusion-strategy structure

In contrast to EF, the LF strategy first processes ACM- and BCM-recorded signals to obtain enhanced speech signals separately and then fuses the outputs from both sides. According to Fig. 1(b), in the presented LF structure, we first create two FCN models to conduct a BCM-to-ACM conversion and an ACM-to-ACM enhancement, respectively, for noisy BCM- and ACM-recorded signals. The resulting output feature maps from both FCNs are then concatenated to serve as the input of another FCN model with simple 1-D convolutional layers, which are expected to produce nearly clean ACM-wise signals. The input-output relationship regarding the three FCNs in this LF multi-modal process can be expressed as follows:

$$s_A[n] = \text{FCN}_A\{x^{(A)}[n]\}, \qquad (2)$$
$$s_B[n] = \text{FCN}_B\{x^{(B)}[n]\}, \qquad (3)$$

and

$$s_{LF}[n] = \text{FCN}_{LF}\{s^{(A)}[n], s^{(B)}[n]\}, \qquad (4)$$

where $\text{FCN}_A\{.\}$, $\text{FCN}_B\{.\}$ and $\text{FCN}_{LF}\{.\}$ denote the FCN model operations for the ACM-to-ACM, BCM-to-ACM, and LF, respectively. In addition, $s_A[n]$, $s_B[n]$ and $s_{LF}[n]$ represent the output signals of the above three FCNs that share a common desired target, namely, a clean version of the ACM-recorded signal $x^{(A)}[n]$. The characteristics of each FCN model used here are further described as follows:

- The ACM-to-ACM enhancement FCN model, FCN_A, which aims to reduce noise distortions in the original ACM-recorded signals, is created following our recent study [22]. That is, the FCN_A model has seven hidden layers, 33 filters for each layer and length 55 for each filter. According to [22], this model can enhance ACM-recorded signals significantly.
- Unlike FCN_A, the FCN_B model conducting BCM-to-ACM conversion is designed in a compact manner, consisting of four convolutional layers in the order of (1, 257), (3, 1), (5, 15) and (1, 513), which follow the setting representation "(number of kernels, kernel size)".
- The fusion function enhancing both $s_A[n]$ and $s_B[n]$ is constructed by two hidden convolutional layers with layer settings, (15, 55) and (1, 55), in the model.

IV. EXPERIMENTS

A. Experimental setup

We conducted the experiments on the Taiwan Mandarin hearing in noise test script (TMHINT) dataset [45]. TMHINT is a balanced corpus consisting of 320 sentences and 10 Chinese characters in each sentence. The utterances in TMHINT were pronounced by a native Mandarin male speaker and recorded simultaneously with an ACM and a BCM in a silent meeting room at a sampling rate of 16 kHz.

During the experiments, we split 320 utterances into three parts: 243 utterances for training, 27 utterances for validation, and 50 utterances for testing. In this study, we considered a personalized SE scenario, where only a small number of utterances pronounced by a single speaker was used for training and testing. For the training set, we added noise to the ACM-recorded utterances with several noise types (two talkers, piano music, a siren, and speech-spectrum-shaped (SSN) noise) at five SNR levels, -4, -1, 2 and 5 dB. For the test set, three noise types (car, baby-cry and helicopter), which were unseen noise types during the training, were added to ACM-recorded utterances at four SNR levels, -5, 0, 5 and 10 dB, to simulate mismatched conditions relative to the training set. Meanwhile, noise-free BCM signals were used as an assistant channel in both training and testing stages.

To evaluate the SE performance of the presented scenario, several objective metrics were used, including a perceptual evaluation of speech quality (PESQ) [46] with the wide-band configuration, short-time objective intelligibility (STOI) [47] and extended STOI (ESTOI) [48]. PESQ indicates the speech quality with a score ranging from -0.5 to 4.5, whereas the STOI and ESTOI metrics typically having score range from 0
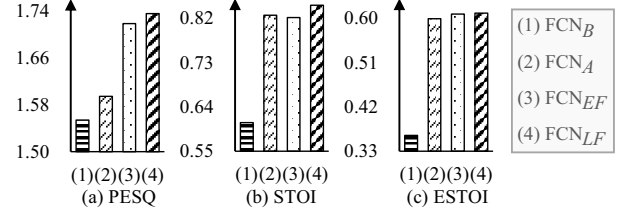

Fig. 2. Scores of different enhancement methods: FCN_B, FCN_A, FCN_{EF}, and FCN_{LF} evaluated with (a) PESQ, (b) STOI, and (c) ESTOI.

to 1 reflect the speech intelligibility (even though they might be also negative).

B. Evaluation results and discussions

Several FCN-wise SE scenarios are compared here, including FCN_B which applies a BCM-to-ACM conversion; FCN_A, which applies an ACM-to-ACM enhancement; and two novel multi-modal approaches, FCN_{EF} and FCN_{LF}.

Table I listed the metric scores for the original and the FCN_B-processed BCM-recorded utterances. From this table, we can see that the original BCM-recorded utterances exhibit a relatively low speech quality and intelligibility even though they do not encounter a noise distortion, which is primarily caused by a lack of high-frequency components. Next, the BCM-to-ACM conversion by the FCN_B model moderately improves the speech quality in terms of PESQ scores, whereas the speech intelligibility is slightly worse than BCM.

Next, the metric scores for the original noisy ACM-recorded utterances and their three enhanced versions (updated using FCN_A, FCN_{EF} or FCN_{LF}) are listed in Tables II and III. From both tables, we can observe the following:

TABLE I
EVALUATION SCORES OF BCM SIGNALS AND FCN_B.

	BCM			FCN_B		
	PESQ	STOI	ESTOI	PESQ	STOI	ESTOI
Avg.	1.247	**0.619**	**0.395**	**1.554**	0.608	0.362

TABLE II
EVALUATION ON NOISY ACMS AND FCN_A IN DIFFERENT SNR LEVELS.

	Noisy ACM			FCN_A		
	PESQ	STOI	ESTOI	PESQ	STOI	ESTOI
10dB	1.722	0.912	0.750	**1.965**	**0.915**	**0.761**
5dB	1.452	0.849	0.624	**1.682**	**0.877**	**0.673**
0dB	1.273	0.766	0.500	**1.446**	**0.809**	**0.552**
-5dB	1.175	0.671	0.386	**1.284**	**0.701**	**0.410**
Avg.	1.405	0.799	0.565	**1.594**	**0.826**	**0.599**

TABLE III
EVALUATION ON FCN_{EF} AND FCN_{LF} IN DIFFERENT SNR LEVELS.

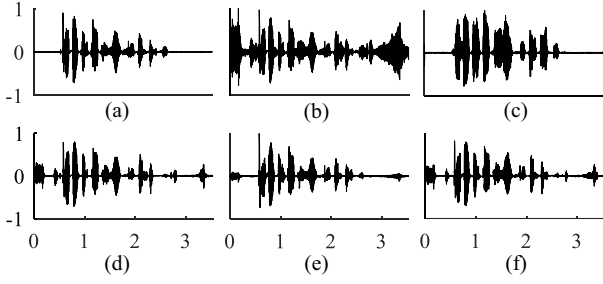|	FCN_{EF}			FCN_{LF}		
	PESQ	STOI	ESTOI	PESQ	STOI	ESTOI
10dB	2.066	0.883	0.722	**2.150**	**0.920**	**0.757**
5dB	1.791	0.853	0.660	**1.858**	**0.889**	**0.678**
0dB	**1.594**	0.804	**0.574**	1.577	**0.833**	0.570
-5dB	**1.422**	**0.744**	**0.475**	1.357	0.740	0.433
Avg.	1.718	0.821	0.608	**1.735**	**0.846**	**0.610**

Fig. 3. The waveform of (a) clean ACM, (b) noisy ACM, (c) BCM, (d) noisy enhanced by FCN$_A$, and (e) FCN$_{EF}$ enhanced speech and the (f) FCN$_{LF}$ enhanced version.
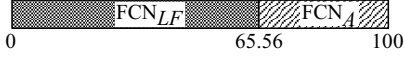

Fig. 4. The results (in percentage, %) for the AB test that compares FCN$_{LF}$ and FCN$_A$ to determine which one brings less signal distortion.

1) Comparing the results for "BCM" in Table I and "Noisy ACM" in Table II, we see that BCM behaves worse than the unprocessed ACM noisy case, revealing that BCM alone fails to enhance speech signals.

2) The FCN$_A$ model, which was purely trained with ACM-recorded signals, behaves satisfactorily in promoting both quality and intelligibility of noisy ACM-recorded utterances. For example, the improvements in the averaged PESQ, STOI, and ESTOI scores are 0.189, 0.025 and 0.034, respectively.

3) Both multi-modal FCN$_{EF}$ and FCN$_{LF}$ structures, which integrate the information from both ACM and BCM, reveal higher PESQ, STOI, and ESTOI scores than the noisy baseline in all SNR cases. These results indicate the success of the presented multi-modal SE scenarios.

4) FCN$_{LF}$ achieves higher evaluation scores at high SNRs (5 dB and 10 dB), and lower performances at low SNRs (0 dB and −5 dB) when compared with FCN$_{EF}$. One possible explanation is that FCN$_{EF}$ possesses better noise-robustness capability when it is employed in a more distorted situation.

5) FCN$_{EF}$ performs especially well and outperforms both FCN$_A$ and FCN$_{LF}$ for lower SNR cases (0 dB and −5 dB), but is less effective than FCN$_A$ in terms of STOI and ESTOI at SNRs of 5 dB and 10 dB. In comparison, FCN$_{LF}$ achieves better PESQ, STOI and ESTOI scores than FCN$_A$ under all SNR conditions.

The evaluation scores from the previous tables averaged over different SNR cases are summarized in Fig. 2 for comparisons. From this figure, we further confirmed that integrating speech sources from both BCM and ACM as in the FCN$_{EF}$ and FCN$_{LF}$ models, can achieve better SE performance in most noisy situations compared with FCN$_A$ and FCN$_B$, in which the models are created with a single speech source. Moreover, the LF strategy for multi-modal as in FCN$_{LF}$ appears to be a better choice here because it outperforms the others in all evaluation indices.

In addition to the objective evaluations, subjective listening tests were also conducted to compare FCN$_{LF}$ and FCN$_A$. We selected two utterances from each of the 9 noisy conditions with 3 noise types (car, baby-cry and helicopter) at 3 SNR

levels (-5, 0 and 5 dB), amounting to 18 ($2 \times 3 \times 3$) utterances. Each utterance was processed by either of FCN$_{LF}$ and FCN$_A$, thus generating 18 utterance pairs, which were used to perform single-blind-designed subjective AB tests in a quiet environment. 20 subjects with normal hearing were recruited to participate the tests. Each subject was asked to listen to a pair of processed utterances and select the one with lower signal distortion. The results are shown in Fig. 4, revealing that 65.56% voted for FCN$_{LF}$ generating low signal distortions and the remaining 34.44% voted for FCN$_A$ from the 360 tests (18 pairs × 20 subjects). In addition, we conducted a matched-pair T-test to confirm the improvement of FCN$_{LF}$ over FCN$_A$ as follows: First, we calculated the votes for FCN$_{LF}$ and FCN$_A$ for each of the 18 utterances from the 20 participants. Next, these 18 paired votes were used to determine the p-value for the T-test, which equals 0.00088. Such a small p-value confirms that the improvement of FCN$_{LF}$ over FCN$_A$ is significant in the subjective evaluation tests.

Finally, Figs. 3(a)-(f) illustrate the waveforms of an utterance under six conditions: (a) clean ACM, (b) noisy ACM, (c) BCM-recorded clean, the noisy ACM enhanced by (d) FCN$_A$, and the concatenated BCM and noisy ACM signal enhanced by (e) FCN$_{EF}$ and (f) FCN$_{LF}$. When comparing the waveform of (c) with that of (a) in the figure, we can observe and confirm again that the BCM-captured speech is similar to the clean ACM counterpart while it has a smoother trajectory. Meanwhile, comparing Fig. 3(b) with Fig. 3(d), we see that FCN$_A$ can reduce noise distortion significantly. However, both FCN$_{EF}$ and FCN$_{LF}$ are shown to outperform FCN$_A$ by providing even less distorted signals when comparing Figs. 3(e)(f) and Fig. 3(d). As a result, we show that integrating BCM and ACM signals appropriately benefits a lot for speech enhancement as in the presented FCN$_{EF}$ and FCN$_{LF}$.

V. CONCLUSION

In this study, we proposed a novel time-domain multi-modal bone/air conducted SE system, where BCM-signals were used as the auxiliary source to promote the conventional ACM-based SE system. We have derived and investigated two ensemble-learning-based fusion strategies, namely EF and LF, to perform multi-modal SE. In particular, for the LF multi-modal structure, two pre-trained FCN models (for BCM- and ACM-recorded signals, respectively) are concatenated and then followed by another compact FCN model with 1-D convolutional layers, along with the normalization and non-linear activation output layers. This structure provides signals with significantly improved PESQ, STOI, and ESTOI metric scores and consistently outperforms the FCN model which uses only ACM-recorded signals for training. Due to its compact model architecture as well as small input data size, the presented multi-modal scenario is quite suitable for implementations on mobile devices, such as cellphones, tablets, and even hearing aids. This study has revealed the effectiveness of the proposed multi-modal SE system on a speaker-specific scenario, such as mobile and personalized device applications. In the future, we will work to improve this system in its SE performance and extend its application in more severe environments.

References

[1] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust automatic speech recognition: a bridge to practical applications*. Academic Press, 2015.

[2] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, *et al.*, "Recent advances in deep learning for speech research at microsoft," in *Proc. ICASSP*, 2013.

[3] X. Chang, W. Zhang, Y. Qian, J. L. Roux, and S. Watanabe, "Mimo-speech: End-to-end multi-channel multi-speaker speech recognition," *arXiv preprint arXiv:1910.06522*, 2019.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018.

[5] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, "Deep feature for text-dependent speaker verification," *Speech Communication*, vol. 73, pp. 1–13, 2015.

[6] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017.

[7] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *The Journal of the Acoustical Society of America*, vol. 139, no. 5, p. 2604, 2016. PMID: 27250154 PMCID: PMC5392064.

[8] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, pp. 1568–1578, 7 2017.

[9] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. INTERSPEECH.*, 2015.

[10] M. Kolbk, Z.-H. Tan, J. Jensen, M. Kolbk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 153–167, 2017.

[11] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 7–19, 2015.

[12] J. Qi, H. Hu, Y. Wang, C.-H. H. Yang, S. M. Siniscalchi, and C.-H. Lee, "Tensor-to-vector regression for multi-channel speech enhancement based on tensor-train network," *arXiv preprint arXiv:2002.00544*, 2020.

[13] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proc. ICASSP*, pp. 5751–5755, 2019.

[14] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," *arXiv preprint arXiv:1709.01703*, 2017.

[15] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. ICASSP*, pp. 116–120, 2015.

[16] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. INTERSPEECH*, 2013.

[17] G. Hu and D. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proc. WASPAA*, 2001.

[18] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 14, pp. 2236–2252, 2003.

[19] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[20] E. M. Grais, D. Ward, and M. D. Plumbley, "Raw multi-channel audio source separation using multi-resolution convolutional auto-encoders," in *Proc. EUSIPCO*, pp. 1577–1581, 2018.

[21] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. APSIPA*, 2017.

[22] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1570–1584, 2018.

[23] S. Gong, Z. Wang, T. Sun, Y. Zhang, C. D. Smith, L. Xu, and J. Liu, "Dilated fcn: Listening longer to hear better," in *Proc. WASPAA*, pp. 254–258, 2019.

[24] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. INTERSPEECH*, pp. 3364–3368, 2017.

[25] Z. Wu, S. Sivadas, Y. K. Tan, M. Bin, and R. S. M. Goh, "Multi-modal hybrid deep neural network for speech enhancement," *arXiv preprint arXiv:1606.04750*, 2016.

[26] F.-K. Chuang, S.-S. Wang, J.-w. Hung, Y. Tsao, and S.-H. Fang, "Speaker-aware deep denoising autoencoder with embedded speaker identity for speech enhancement," in *Proc. INTERSPEECH*, pp. 3173–3177, 2019.

[27] N. Tawara, T. Kobayashi, and T. Ogawa, "Multi-channel speech enhancement using time-domain convolutional denoising autoencoder," *Proc. INTERSPEECH*, pp. 86–90, 2019.

[28] H. S. Shin, H.-G. Kang, and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphone," in *Proc. Speech Communication; 10. ITG Symposium*, pp. 1–4, 2012.

[29] T. Shimamura, J. Mamiya, and T. Tamiya, "Improving bone-conducted speech quality via neural network," in *Proc. ISSPIT*, pp. 628–632, 2006.

[30] C. H. Lee, B. D. Rao, and H. Garudadri, "Bone-conduction sensor assisted noise estimation for improved speech enhancement.," in *Proc. INTERSPEECH*, pp. 1180–1184, 2018.

[31] Y. Xing, C. Zheng, X. Zhang, and Z. Fang, "A speech enhancement scheme based on bone-conducted speech," in *in Proc. ICMSSP*, pp. 75–79, 2018.

[32] H.-P. Liu, Y. Tsao, and C.-S. Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," *Speech Communication*, vol. 104, pp. 106–112, 2018.

[33] Y. Tajiri, H. Kameoka, and T. Toda, "A noise suppression method for body-conducted soft speech based on non-negative tensor factorization of air-and body-conducted signals," in *Proc. ICASSP*, pp. 4960–4964, 2017.

[34] T. V. Thang, K. Kimura, M. Unoki, and M. Akagi, "A study on restoration of bone-conducted speech with MTF-based and LP-based models," *Journal of signal processing*, vol. 10, pp. 407–417, 2006.

[35] T. Shinamura and T. Tomikura, "Quality improvement of bone-conducted speech," in *Proc. ECCTD*, 2005.

[36] M. S. Rahman, A. Saha, and T. Shimamura, "Multisensory speech enhancement using lower-frequency components from bone-conducted speech," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 14, no. 11, pp. 1661–1666, 2019.

[37] M. Li, I. Cohen, and S. Mousazadeh, "Multisensory speech enhancement in noisy environments using bone-conducted and air-conducted microphones," in *Proc. ChinaSIP*, pp. 1–5, 2014.

[38] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. INTERSPEECH*, 2019.

[39] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain," in *Proc. INTERSPEECH*, 2018.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[41] Y. Zheng, Z. Liu, Z. Zhang, M. Sinclair, J. Droppo, L. Deng, A. Acero, and X. Huang, "Air- and bone-conductive integrated microphones for robust speech detection and enhancement," in *Proc. ASRU*, 2003.

[42] B. Huang, Y. Gong, J. Sun, and Y. Shen, "A wearable bone-conducted speech enhancement system for strong background noises," in *Proc. ICEPT*, pp. 1682–1684, 2017.

[43] D. G. Stork and M. E. Hennecke, *Speechreading by humans and machines: models, systems, and applications*. Springer Science & Business Media, 2013.

[44] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.

[45] M.-W. Huang, "Development of taiwan mandarin hearing in noise test," *Master thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health Sciences*, 2005.

[46] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, pp. 749–752, 2001.

[47] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[48] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 2009–2022, 2016.