TIME-DOMAIN MULTI-MODAL BONE/AIR CONDUCTED SPEECH ENHANCEMENT

Cheng Yu¹, Yan-Ting Lin¹, Kuo-Hsuan Hung¹, Syu-Siang Wang¹, Szu-Wei Fu¹, Yu Tsao¹, and Jeih-weih Hung²

¹ Research Center for Information Technology Innovation, Academia Sinica, Taiwan
² Department of Electrical Engineering, National Chi Nan University, Taiwan

ABSTRACT

Integrating modalities, such as video signals with speech, has been shown to provide a standard quality and intelligibility for speech enhancement (SE). However, video clips usually contain large amounts of data and pose a high cost in terms of computational resources, which may complicate the respective SE. By contrast, a bone-conducted speech signal has a moderate data size while it manifests speech-phoneme structures, and thus complements its air-conducted counterpart, benefiting the enhancement. In this study, we propose a novel multi-modal SE structure that leverages bone- and air-conducted signals. In addition, we examine two strategies, early fusion and late fusion (LF), to process the two types of speech signals, and adopt a deep learning-based fully convolutional network to conduct the enhancement. The experiment results indicate that this newly presented multi-modal 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

1. INTRODUCTION

Speech enhancement (SE) aims to improve the speech quality and the intelligibility in a noisy environments, and has been widely applied in many tasks, such as speaker, speech, and emotion recognition [1-3, 4, 5], to improve the system robustness against environmental noise. Conventional SE approaches can be divided into filtering, spectral restoration, and speech model techniques [6]. The basic idea of these approaches on reducing the noise components is to apply a filter functions on the noisy input. The function is normally designed based on the distinct statistic properties between the clean speech and background interference. Some famous methods include Wiener filter [7] minimum mean square error spectral estimator [7, 8], harmonic model [6] and the hidden Markov model [9].

Recently, deep-learning-based methods have shown compatibility to learn non-linear mapping functions for SE [10, 11, 12]. For these approaches, the noise-corrupted speech is mostly enhanced in the spectrum with a deep-learning-based model. In the frequency domain, supervised learning aims to estimate the clean magnitude spectra or the corresponding signal-to-noise ratio (SNR) from the input noisy speech [13, 14, 15]. In addition to applying SE in the frequency domain, a fully convolutional network (FCN) [16, 17] is used to directly estimate a temporal mapping function that circumvents the interference caused by the noisy phase when recovering speech from its processed spectrogram. The results show that FCN not only decreases the number of parameters of a deep-learning model, but also restores better precision in high acoustic-frequency components of the speech waveform.

In addition, the 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). 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. Several filtering-based and probabilistic solutions have been proposed to convert the BCM-recorded sound to its ACM version. Shimamura et. al. [18] used a reconstruction filter, which is designed using the long-term spectra of the speech, to perform the conversion. Meanwhile, numerous approaches have been proposed to combine ACM- and BCM-recorded sound in hardware devices with a linear transformation for SE and speech recognition tasks [19, 20].

In this study, we propose a novel deep-learning-based SE method that leverages the acoustic characteristics between signals recorded using a BCM and a normal ACM. This method primarily takes advantage of the noise robustness of BCM-recorded signals and the capability of an FCN model in restoring the high acoustic-frequency components in the signals. Experimental results show that the newly presented method is a significant improvement in terms of various 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.

2. RELATED WORKS

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

Employing a deep learning-based model structure is a main element of an SE technique. In [16, 17], an FCN model was used to directly process the input time-domain waveform. By contrast, in the studies presented in [21, 22], waveform-wise enhancement was conducted using a convolutional neural network (CNN) structure. 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.

2.2. BCM/ACM conversion

A straightforward method used to collect less distorted speech signals applies 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 [18, 19].

2.3. 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 [23], 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 [24] combines audio with lip-motion clips to access more bio-information and thereby promote the performance of using an SE system. 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.

3. 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. Hence, we believe that, if arranged appropriately, the two types of signals can complement each other when applied to SE.

3.1. 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, both kinds of signals 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.

3.1.1. 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:

\[ s_{EF}[n] = FCN_{EF}\{x^{(A)}[n], x^{(B)}[n]\}, \]  

(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; \( FCN_{EF}\{.\} \) denotes the FCN model operation used; and \( s_{EF}[n] \) is an enhanced signal expected to approximate the cleanliness of \( x^{(A)}[n] \).

In addition, to examine the impact of the BCM, we constructed another FCN model that is close to the \( FCN_{EF} \), which only adopts the ACM channel. Evaluations between these models are described in Sec. 4.
### 3.1.2. Late-fusion-strategy structure

In contrast to EF, the LF strategy suggests enhancing ACM- and BCM-recorded signals separately, and then integrates the outputs from both sides. However, because the two separate outputs might lose mutual correlations, it is often crucial to apply another model to appropriately integrate them to obtain the ultimately enhanced signal. 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 a simple 1-D convolutional layer, which is expected to produce mostly clean ACM-wise signals. The input-output relationship regarding the three FCNs in this LF multi-modal process can be expressed as follows:

\[ s_A[n] = \text{FCN}_A\{x^{(A)}[n]\}, \]
\[ s_B[n] = \text{FCN}_B\{x^{(B)}[n]\}, \]
\[ s_{LF}[n] = \text{FCN}_{LF}\{s^{(A)}[n], s^{(B)}[n]\}, \]

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, \(\text{FCN}_A\), which aims to reduce noise distortions in the original ACM-recorded signals, is created following our recent study [17]. According to [17], this FCN model enhances the ACM-recorded signal significantly.
- Unlike \(\text{FCN}_A\), the \(\text{FCN}_B\) model conducting BCM-to-ACM conversion is designed in a compact manner, consisting of only convolutional layers, normalization layers and one hyperbolic tangent output layer.

### 4. EXPERIMENTS

#### 4.1. Experimental setup

We conducted the experiments on the Taiwan Mandarin hearing in noise test script (TMHINT) dataset [25]. 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. For the training and validation sets, 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 of -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 of -5, 0, 5 and 10 dB, to simulate mismatched conditions relative to the training set.

To evaluate the SE performance of the presented scenario, several objective metrics were used, which comprised a perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI) and extended STOI (ESTOI). PESQ indicates the speech quality with a score ranging from -0.5 to 4.5, whereas the STOI and ESTOI metrics reflect the speech intelligibility with a score ranging from 0 to 1.

#### 4.2. Evaluation results and discussions

Several FCN-wise SE scenarios are compared here, including \(\text{FCN}_B\) which applies a BCM-to-ACM conversion; \(\text{FCN}_A\), which applies an ACM-to-ACM enhancement; and two novel multi-modal approaches, \(\text{FCN}_{EF}\) and \(\text{FCN}_{LF}\).

Table 1 listed the metric scores for the original and the \(\text{FCN}_B\)-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 brought about by the \(\text{FCN}_B\) model moderately improves the speech quality from 1.247 to 1.554 in terms of PESQ scores, whereas the speech intelligibility does not benefit from \(\text{FCN}_B\).

Next, the metric scores for the original noisy ACM-recorded utterances and their three enhanced versions (updated using \(\text{FCN}_A\), \(\text{FCN}_{EF}\) or \(\text{FCN}_{LF}\)) are listed in Tables 2 and 3. From these two tables, we can observe the following:

**Table 1.** Evaluation Scores of BCM signals and FCN\(_B\) in different SNR levels.

| SNR  | BCM PESQ | BCM STOI | BCM ESTOI | FCN\(_B\) PESQ | FCN\(_B\) STOI | FCN\(_B\) ESTOI |
|------|----------|----------|----------|----------------|----------------|----------------|
| Avg. | 1.247    | 0.619    | 0.395    | 1.554          | 0.608          | 0.362          |
1. 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.

2. The two multi-modal FCN structures, FCN$_{EF}$ and FCN$_{LF}$, 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.

3. 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 for this is the better noise-robustness capability when applying the FCN$_{EF}$ SE approach to noisy speech in a severely noisy environment.

4. 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 each of comparison. 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, in comparisons with FCN$_A$.

Table 2. Evaluation Scores of noisy ACM signals and FCN$_A$ in different SNR levels.

| SNR  | 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 3. Evaluation Scores of FCN$_{EF}$ and proposed FCN$_{LF}$ in different SNR levels.

| SNR  | 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.

In this study, we proposed a novel multi-modal SE scenario using two different fusion strategies, namely early fusion and late fusion. In particular, for the late-fusion multi-modal structure, two pre-trained FCN models (for BCM- and ACM-recorded signals, respectively) are concatenated, followed by another compact FCN model with a 1-D convolutional layer, along with the normalization and non-linear activation output layers. This structure achieves 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.

5. CONCLUSION

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 signal on some levels at a smooth trajectory. Meanwhile, FCN$_A$ in Fig. 3(d) shows small noise components in the enhanced speech, and suggests the effectiveness of the applied model on enhancing the noisy waveform, which is depicted in Fig. 3(b). However, both FCN$_{EF}$ and FCN$_{LF}$ can provide more noise-free speech when comparing the waveform in Figs. 3(e) and (f) with that in (d). Clear utterances in both FCN$_{EF}$ and FCN$_{LF}$ enhanced speech imply that integrating the BCM signal can promote the performance of an SE system.
6. REFERENCES

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

[2] 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.

[3] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. ICASSP*, 2013.

[4] 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.

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

[6] J. Chen, J. Benesty, Y. Huang, and E. Diethorn, “Fundamentals of noise reduction in spring handbook of speech processing-chapter 43,” 2008.

[7] P. Scalart and J. V. Filho, “Speech enhancement based on a priori signal to noise estimation,” in *Proc. ICASSP*, 1996.

[8] J. H. Hansen, V. Radhakrishnan, and K. H. Arehart, “Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system,” *IEEE/ACM TASLP*, vol. 14, pp. 2049–2063, 2006.

[9] L. R. Rabiner and B.-H. Juang, “An introduction to hidden markov models,” *IEEE ASSP Magazine*, vol. 3, pp. 4–16, 1986.

[10] 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. Annu. Conf. Int. Speech Commun. Assoc.*, 2015.

[11] 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 TASLP*, vol. 25, pp. 153–167, 2017.

[12] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks.” *IEEE/ACM TASLP*, vol. 23, pp. 7–19, 2015.

[13] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in *Proc. Interspeech*, 2013.

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

[15] 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.

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

[17] 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 TASLP*, vol. 26, pp. 1570–1584, 2018.

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

[19] 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.

[20] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, “Multi-sensory microphones for robust speech detection, enhancement, and recognition,” in *Proc. ICASSP*, 2004.

[21] A. Pandey and D. Wang, “TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in *Proc. Interspeech*, 2019.

[22] A. Pandey and D. Wang, “A new framework for supervised speech enhancement in the time domain,” in *Proc. Interspeech*, 2018.

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

[24] 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 TETCI*, vol. 2, pp. 117–128, 2018.

[25] 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.