DEEP COMPLEX U-NET WITH CONFORMER FOR AUDIO-VISUAL SPEECH ENHANCEMENT

Shafique Ahmed[1,2], Chia-Wei Chen[3], Wenze Ren[3], Chin-Jou Li[3], Ernie Chu[3], Jun-Cheng Chen[1], Amir Hussain[4], Hsin-Min Wang[1], Yu Tsao[1], and Jen-Cheng Hou[1]

[1]Academia Sinica [2]National Tsing Hua University, [3]National Taiwan University [4]Edinburgh Napier University

shafique.khattak13@gmail.com, r11943130@ntu.edu.tw, r11942166@ntu.edu.tw, b09902035@csie.ntu.edu.tw
shchu@citi.sinica.edu.tw, pullpull@citi.sinica.edu.tw, a.hussain@napier.ac.uk, whm@iis.sinica.edu.tw
yu.tsao@citi.sinica.edu.tw, jchou@citi.sinica.edu.tw

ABSTRACT

Recent studies have acknowledged the advantages of incorporating visual data into SE systems. We introduce a novel audio-visual SE approach, termed DCUC-Net (deep complex U-Net with conformer network). DCUC-Net leverages complex domain features and conformer blocks. The encoder and decoder of DCUC-Net use a complex U-Net-based framework. Audio and visual signals are processed using a complex encoder and ResNet-18 model. These signals are fused using conformer blocks and transformed into enhanced speech waveforms via a complex decoder. The conformer blocks consist of self-attention mechanisms and convolutional operations, enabling DCUC-Net to capture global and local audio-visual dependencies. Our experimental results show DCUC-Net outperforms the baseline model from the COG-MHEAR AVSE Challenge 2023 by 0.14 in terms of PESQ. Additionally, DCUC-Net performs comparably to a state-of-the-art model and outperforms other models on the Taiwan Mandarin speech with video (TMSV) dataset.

Index Terms— audio-visual speech enhancement, speech enhancement, complex ratio masking, conformer

1. INTRODUCTION

Effective communication through spoken language is the cornerstone of human interaction. In various fields such as telecommunications, voice assistants, hearing aids, and video conferencing, ensuring the quality and intelligibility of speech is crucial. However, the challenge lies in the ability to consistently achieve good quality and intelligibility, especially in adverse acoustic environments characterized by background noise, reverberation, or limited audio quality. In the pursuit of addressing these challenges and enhancing speech quality and intelligibility, speech enhancement (SE) has emerged as a critical area of research and development [1]. The field of SE has witnessed significant advancements through the integration of deep learning methods. While deep learning-based SE techniques [2, 3] have demonstrated remarkable success by primarily focusing on audio signals alone, it is notable that incorporating the visual modality can bring substantial benefits to improving the performance of SE systems in challenging acoustic environments [4, 5, 6].

Depending on the input type, audio-only SE methods can be roughly divided into two categories: time-domain methods and time-frequency (TF) domain methods. Conventional TF domain methods typically rely on amplitude spectral features; yet, studies have shown that SE performance can be limited because phase information is not adequately considered [3]. To address this issue, approaches that employ complex-valued features, such as complex spectral mapping (CSM) [7] and complex ratio masking (CRM) [8], as SE input have recently been proposed. Many CSM and CRM techniques are built upon real-valued neural networks, while others employ complex-valued neural networks to process the complex input. Deep Complex U-NET (DCUNET) [9] and Deep Complex Convolution Recurrent Network (DCCRN) [10] are notable complex-valued neural networks for the SE task. In this study, we also use a complex-valued neural network to process audio data.

The core idea behind audio-visual speech enhancement (AVSE) is to integrate visual input as supplementary data into an audio-only SE system, aiming to improve SE performance with the help of supplementary information. Several earlier studies have demonstrated the efficacy of incorporating visual input to enhance the performance of SE systems [4, 11, 12]. Most previous AVSE systems focused primarily on processing audio in the TF domain [13, 6]. Some previous studies performed audio-visual speech separation tasks in the time domain [14]. Recently, self-supervised learning (SSL) embeddings have been used to improve the performance of AVSE. For example, Richard et al. [15] introduced the SSL-AVSE method, which integrates visual cues with audio signals. These combined audio-visual data are fed into a Transformer-based SSL AV-HuBERT model to extract features, which are subsequently processed using a BLSTM-based SE model.

In this paper, we propose a novel AVSE approach, termed DCUC-Net (deep complex U-Net with conformer network). DCUC-Net employs a deep complex U-Net architecture to incorporate phase information into AVSE. The audio stream is processed by a complex encoder to create a complex representation of the audio data and then combined with visual features. To further enhance the combined audio-visual features, DCUC-Net integrates conformer blocks into the deep complex U-Net architecture. These conformer blocks enable DCUC-Net to effectively capture long-range dependencies, both local and global, and fine-grained contextual information between audio and visual modalities. The resulting features are then decoded by a complex decoder to estimate a complex mask. By multiplying the complex spectra from the noisy input with the estimated complex mask, we can obtain enhanced complex spectra, which are then converted to enhanced waveforms as the final output. We evaluate the performance of DCUC-Net through experiments on the dataset used in the COG-MHEAR AVSE Challenge 2023, which is built on the LRS3 dataset [16], and the Taiwan Mandarin speech with video (TMSV) dataset. The results demonstrate the efficacy of DCUC-Net, as it surpasses existing baselines and demonstrates robust denoising capabilities.

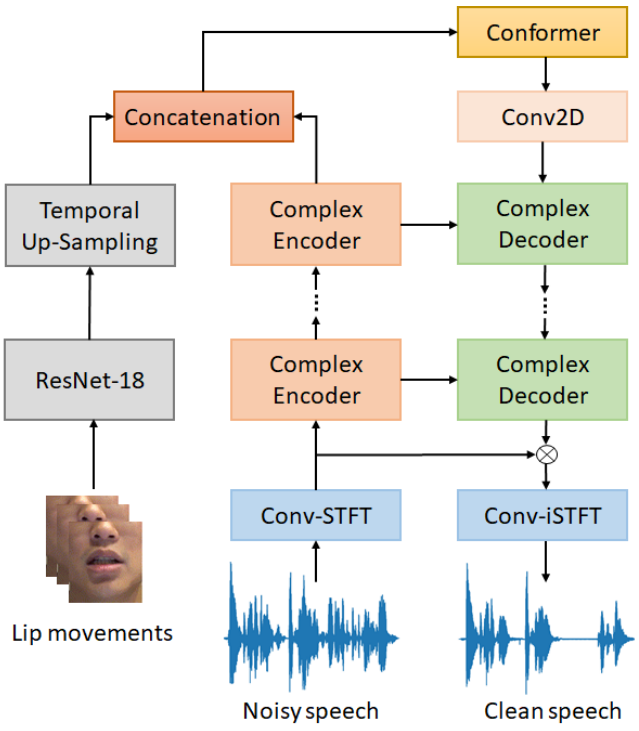


Fig. 1. Architecture of the proposed DCUC-Net.

2. THE PROPOSED DCUC-NET AVSE SYSTEM

The proposed DCUC-Net aims to obtain enhanced speech signals with improved intelligibility and quality via the integration of audio and visual information. Figure 1 shows the overall architecture of DCUC-Net. In DCUC-Net, we utilize ResNet-18 to process lip movement cues, thereby preparing the visual information. Additionally, we employ a complex encoder to transform the audio input into complex-valued features. The complex encoder of DCUC-Net comprises five Conv2d blocks, which are designed to extract encoded features from the audio. These extracted audio features are then combined with visual information, and we leverage conformer blocks [17] to effectively process this combined audio-visual information. During decoding, the symmetric complex-valued encoder-decoder structure is considered and used to reconstruct the lower-dimensional representation to the original size of the input, facilitating the restoration of the enhanced speech waveforms. The complex encoder and decoder block in DCUC-Net is constructed based on the implementation details described in [18]. It comprises three main components: complex Conv2d [10], complex batch normalization [18], and real-valued PReLU [19]. The complex Conv2d operation enables the manipulation of complex-valued features. In this operation, the complex-valued convolutional filter, denoted as W, is defined as $W = W_r + jW_i$, where W_r represents the real part, and W_i represents the imaginary part. Similarly, the input complex spectra, denoted as X, is defined as $X = X_r + jX_i$, where X_r and X_i denote the real and imaginary parts, respectively. The complex convolution operation, denoted as $X \bigotimes W$, is performed to obtain the complex output. The output feature of each complex layer, represented as F_{out}, is calculated as:

$$F_{out} = (X_r \times W_r - X_i \times W_i) + j(X_r \times W_i + X_i \times W_r). \quad (1)$$

Equation (1) within DCUC-Net enables the manipulation of both the real and imaginary parts of the complex-valued features, facilitating effective processing of the audio part.

Expanding on this capability, we integrate the extracted visual features into DCUC-Net by concatenating them with the real values from the complex audio feature. However, this concatenation process introduces a sequential misalignment between the visual and audio features. To address this discrepancy, we employ a temporal upsampling technique to align the temporal sequences of the visual features with the temporal resolution of the audio features. This alignment process ensures proper fusion of the two modalities. Subsequently, the aligned audio and visual features are input into the Conformer blocks.

The Conformer block comprises two Feed-Forward (FFN) modules positioned around the Multi-Headed Self-Attention (MHSA) module and the Convolution (Conv) module. In contrast to the original Transformer block, which incorporates a

single feed-forward layer, the Conformer block features two half-step feed-forward layers: one preceding the attention layer and one succeeding it. To put it simply, when the i Conformer block receives an input, denoted as x_i the resulting output y_i is determined through the following mathematical expression:

$$\begin{aligned} \hat{x}_i &= x_i + \frac{1}{2}\text{FFN}(x_i) \\ \check{x}_i &= \hat{x}_i + \text{MHSA}(\hat{x}_i) \\ \bar{x}_i &= \check{x}_i + \text{Conv}(\check{x}_i) \\ y_i &= \text{Layernorm}(\bar{x}_i + \frac{1}{2}\text{FFN}(\bar{x}_i)) \end{aligned} \tag{2}$$

The conformer blocks play a critical role in capturing both global and local audio-visual dependencies within our DCUC-Net. These blocks facilitate the effective combination of the audio and visual modalities by leveraging self-attention mechanisms and convolutional operations. The self-attention mechanisms establish relationships between different parts of the audio and visual features, enabling the model to capture long-range dependencies and effectively leverage global audio-visual cues. Simultaneously, the convolutional operations extract local contextual information, ensuring the model's robustness in capturing fine-grained audio-visual details.

With the incorporation of the audio and visual information, the output of the conformer blocks is passed to the complex decoder component in DCUC-Net. The role of this decoder is to estimate the complex ratio mask and obtain enhanced complex spectra by multiplying them with the noisy complex spectra. Subsequently, the complex spectra are transformed into an enhanced speech waveform by applying the Convolutional inverse Short-Time Fourier Transform (ConviSTFT). It's worth noting that the decoder benefits from a skip-connection mechanism, which provides access to encode features via the U-Net architecture.

3. EXPERIMENTS

3.1. Datasets

We evaluated the proposed DCUC-Net on two datasets: the COG-MHEAR Audio-Visual Speech Enhancement Challenge 2023 [20] dataset and the Taiwan Mandarin speech with video (TMSV) dataset. The COG-MHEAR dataset is built on the LRS3 dataset and contains two types of interference sources: speech and noise. The speech interference source comes from the LRS3 dataset, while the noise interference source comes from three different sources: the 1st Clarity Enhancement Challenge, DEMAND [21], and Deep Noise Suppression Challenge (DNS) [22] 2nd version. In the training set, interference sources come from 405 competing speakers and 7,346 noise files covering 15 noise categories. There are a total of 34,524 scenes with 605 target speakers. On the other hand, For the test set, we utilized the development set provided by the challenge, in which interference sources are selected from 30 competing speakers and 1,825 noise files in the same 15 noise categories mentioned above. The test set contains 3,306 scenes with 85 target speakers. The TMSV dataset comprises video recordings from 18 native Mandarin speakers (13 males and 5 females). Each speaker recorded 320 Mandarin sentences, with each sentence consisting of 10 Chinese characters. The duration of each utterance ranges from approximately 2 to 4 seconds. To ensure methodological congruence and experiment reproducibility, we followed the procedures detailed in [12] for the introduction of noise interference into the dataset and train-test split.

3.2. Experimental Setup

DCUC-Net preprocesses the input audio using a Convolutional Short-Time Fourier Transform (ConvSTFT) with a window length of 400 samples, a window increment of 160 samples, and an FFT length of 512 samples. The model encoder consists of 5 complex Conv2D blocks for audio processing, ResNet-18 for visual feature extraction (followed by temporal upsampling), 2 conformer blocks for capturing audio-visual dependencies, and 5 decoder blocks for estimating the complex mask. The enhanced speech signal is obtained by multiplying the estimated mask with the complex spectrogram and applying the Convolutional inverse Short-Time Fourier Transform (Conv-iSTFT).

To train DCUC-Net, we employed the Scale-Invariant Signal-to-Noise Ratio (SI-SNR) [23] as the loss function, which is defined as:

$$\text{SI-SNR} = 10\log_{10}\left(\frac{\|s_{\text{target}}\|^2}{\|e_{\text{noise}}\|^2}\right), \tag{3}$$

where $\|s_{\text{target}}\|^2$ is the power of the target speech signal, and $\|e_{\text{noise}}\|^2$ represents the power of the estimated noise. The estimated noise is obtained by subtracting the target speech signal from the estimated speech signal. SI-SNR is a commonly used evaluation metric that provides a more robust measure of performance than mean square error (MSE).

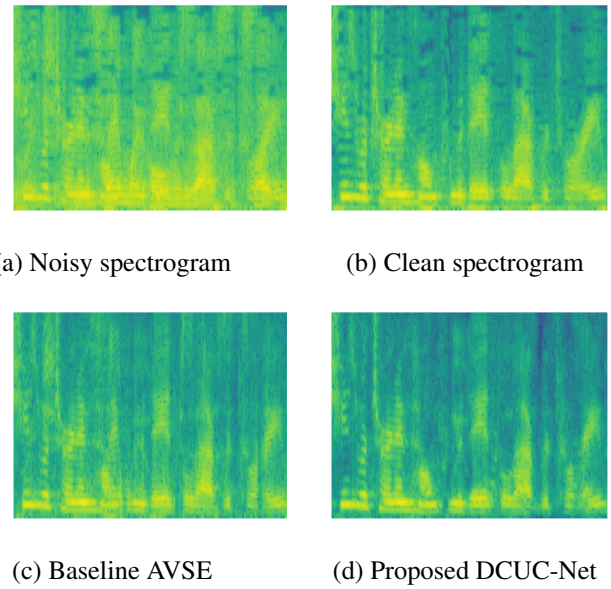
3.3. Evaluation Results on the COG-MHEAR Dataset

We evaluated three types of speech, namely noisy speech, enhanced speech generated by baseline models from the COG-MHEAR AVSE Challenge in 2022 and 2023, and enhanced speech generated by DCUC-Net, with two standard metrics: Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI). Several noteworthy observations can be drawn from the results in Table 1. First, compared with noisy speech, the enhanced speech output of all AVSE models exhibits superior quality (as indicated by PESQ) and higher intelligibility (as shown by STOI). Second, DCUC-Net outperforms current and previous years' baselines. Finally, DCUC-Net with conformers

for combining audio and visual information outperforms all other models (especially the one that used Bidirectional Long Short-Term Memory (BLSTM) and transformer for combining audio and visual information) in both evaluation metrics, providing strong evidence for the efficacy of integrating the conformer blocks into DCUC-Net model architecture.

Table 1. Objective assessment scores of noisy speech and enhanced speech generated by the baseline and the proposed DCUC-net on the COG-MHEAR dataset.

	PESQ	STOI
Noisy	1.15	0.64
Baseline (2022)[20]	1.30	0.67
Baseline (2023)[20]	1.70	0.83
Ours (BLSTM)	1.62	0.81
Ours (Transformer)	1.79	0.83
Ours (Conformer)	**1.84**	**0.84**



(a) Noisy spectrogram (b) Clean spectrogram

(c) Baseline AVSE (d) Proposed DCUC-Net

Fig. 2. Spectrograms of (a) noisy speech, (b) clean speech, (c) enhanced speech by the COG-MHEAR AVSE Challenge 2023 baseline, and (d) enhanced speech by DCUC-Net for an example in the COG-MHEAR development dataset. The vertical axis represents frequency, and the horizontal axis represents time.

We performed spectrogram analysis on noisy speech, clean speech, and enhanced speech generated by the COG-MHEAR AVSE Challenge 2023 baseline and the proposed DCUC-Net. Figure 4 shows the corresponding spectrograms associated with an example in the development set. It is obvious from the figure that our model effectively suppresses the noise components present in the noisy speech spectrogram. Comparing the enhanced speech spectrograms of DCUC-Net and the baseline model, DCUC-Net provides significantly better noise reduction capabilities than the baseline model.

3.4. Evaluation Results on the TMSV Dataset

Table 2. Objective assessment scores of noisy speech and enhanced speech generated by various AVSE models on the TMSV dataset.

	PESQ	STOI
Noisy	1.18	0.60
LogMMSE(Audio-only)[24]	1.21	0.61
AVCVAE [25]	1.34	0.63
LAVSE [12]	1.31	0.61
SSL-AVSE [13]	1.40	**0.68**
Ours (Conformer)	**1.41**	0.66

We also evaluated noisy speech and enhanced speech generated by different SE and AVSE models on the TMSV dataset, including one audio-only traditional SE approach (MMSE), and three deep learning-based AVSE systems (AVCVAE [25], LAVSE [12], and SSL-AVSE [13]). From Table 2, we can see that DCUC-Net outperforms all compared models in PESQ and STOI metrics, except SSL-AVSE [13] in STOI metric. It is worth noting that AVSE adopts the pre-trained self-supervised learning model AV-Hubert for feature extraction, but the proposed DCUC-Net is trained from scratch. The model size of DCUC-Net is also much smaller than SSL-AVSE. Therefore, DCUC-Net has its advantages in terms of practical implementation.

4. CONCLUSIONS

This paper introduces a novel AVSE framework in which we incorporate lip movement cues as visual features into the audio stream, and we emphasize the importance of phase information by leveraging complex features in a deep complex U-Net architecture. The visual features are combined with the output of the complex encoder for the audio stream, and the resulting concatenated features are processed by conformer blocks. The proposed framework effectively captures global and local audio-visual dependencies. Experimental results on the COG-MHEAR dataset demonstrate the superior performance of the proposed DCUC-Net AVSE framework over two baselines. Furthermore, on the TMSV dataset, DCUC-Net performs comparably to a state-of-the-art model that uses the pre-trained self-supervised learning AV-HuBERT model for feature extraction, and outperforms all other compared models.

5. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder.," in *Proc. Interspeech*, 2013, vol. 2013, pp. 436–440.

[3] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. ICASSP*, 2014, pp. 1562–1566.

[4] 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.

[5] B. Xu, C. Lu, Y. Guo, and J. Wang, "Discriminative multi-modality speech recognition," in *CVPR*, 2020.

[6] D. Michelsanti *et al.*, ," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1368–1396, 2021.

[7] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *Proc. ICASSP*, 2019, pp. 6865–6869.

[8] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[9] H.-S. Choi, J. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," in *Proc. ICLR*, 2019.

[10] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.

[11] J. S. Chung T. Afouras and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Interspeech*, 2018.

[12] S.-Y. Chuang, H.-M. Wang, and Y. Tsao, "Improved lite audio-visual speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1345–1359, 2022.

[13] I-C. Chern, K.-H. Hung, Y.-T. Chen, T. Hussain, M. Gogate, A. Hussain, Y. Tsao, and J.-C. Hou, "Audio-visual speech enhancement and separation by utilizing multi-modal self-supervised embeddings," in *Proc. ICASSPW*, 2023, pp. 1–5.

[14] Y. Wu, C. Li, J. Bai, Z. Wu, and Y. Qian, "Time-domain audio-visual speech separation on low quality videos," in *Proc. ICASSP*, 2022, pp. 256–260.

[15] R. L. Lai, J.-C. Hou, M. G., K. Dashtipour, A. Hussain, and Y. Tsao, "Audio-visual speech enhancement using self-supervised learning to improve speech intelligibility in cochlear implant simulations," 2023.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," 2018.

[17] A. Gulati, C.-C. Chiu, J. Qin, J. Yu, N. Parmar, R. Pang, S. Wang, W. Han, Y. Wu, Y. Zhang, and Z. Zhang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020, pp. 5036–5040.

[18] C. Trabelsi *et al.*, "Deep complex networks," in *Proc. ICLR*, 2018.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. ICCV*, 2015, pp. 1026–1034.

[20] "2nd COG-MHEAR Audio-Visual Speech Enhancement Challenge, 2023," `https://challenge.cogmhear.org`.

[21] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," 2017.

[22] C. K. A. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. Interspeech*, 2021, pp. 2796–2800.

[23] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[24] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.

[25] M. Sadeghi *et al.*, "Audio-visual speech enhancement using conditional variational auto-encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1788–1800, 2020.