Lite Audio-Visual Speech Enhancement

Shang-Yi Chuang[1], *Yu Tsao*[1], *Chen-Chou Lo*[2], *Hsin-Min Wang*[3]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]EAVISE, Dept. of Electrical Engineering, KU Leuven, Belgium
[3]Institute of Information Science, Academia Sinica, Taiwan

{sychuang,yu.tsao}@citi.sinica.edu.tw, chenchou.lo@kuleuven.be, whm@iis.sinica.edu.tw

Abstract

Previous studies have confirmed the effectiveness of incorporating visual information into speech enhancement (SE) systems. Despite improved denoising performance, two problems may be encountered when implementing an audio-visual SE (AVSE) system: (1) additional processing costs are incurred to incorporate visual input and (2) the use of face or lip images may cause privacy problems. In this study, we propose a Lite AVSE (LAVSE) system to address these problems. The system includes two visual data compression techniques and removes the visual feature extraction network from the training model, yielding better online computation efficiency. Our experimental results indicate that the proposed LAVSE system can provide notably better performance than an audio-only SE system with a similar number of model parameters. In addition, the experimental results confirm the effectiveness of the two techniques for visual data compression.

Index Terms: speech enhancement, audio-visual, data compression, quantization

1. Introduction

Speech enhancement (SE) has been widely used in the front-end processing of numerous speech-related applications such as automatic speech recognition (ASR), assistive hearing technologies, and speaker recognition. The goal of SE is to improve speech quality and intelligibility by converting low-quality speech into high-quality speech. Traditional SE approaches are derived based on the assumed audio characteristics of clean speech signals and distortion sources and can be broadly divided into three categories. The first category aims to estimate the gain function to suppress noise components in noisy speech [1, 2, 3]. The second category aims to divide the noise and clean speech components into distinct subspaces, and the elements in the clean speech subspace are then used to generate enhanced speech [4, 5, 6]. The third category utilizes a nonlinear model to directly map noisy speech signals to clean ones [7, 8, 9].

Recently, deep-learning (DL) models have become popular in several fields, such as image recognition [10], ASR [11], and natural language processing [12]. In the SE field, DL models have been extensively used as fundamental units as well. Various network architectures, including the deep denoising autoencoder [13, 14], fully connected network [15, 16, 17], convolutional neural network (CNN) [18, 19], recurrent neural network (RNN), and long short-term memory (LSTM) [20, 21, 22], have been confirmed to notably enhance SE capabilities over those of traditional SE approaches. The success of these DL models can be attributed to their outstanding nonlinear mapping properties, which can accurately characterize complex transformations from noisy to clean speech signals.

Another advantage of DL models is their ability to fuse multimodal data. In research on SE, visual information has been adopted as auxiliary information to facilitate better SE performance [23, 24, 25, 26]. Compared to audio-only SE systems, audio-visual SE (AVSE) systems have been proven to provide improved enhancement capabilities. In exchange for the improved performance, however, two problems may be encountered when implementing an AVSE system. The first concerns the increase in input data. Compared to audio-only systems, the implementation of AVSE systems requires additional image inputs, which can be immense. Accordingly, additional hardware and computation costs are incurred during the training process. Second, the incorporation of face or lip images, which are particularly sensitive and personal, can generate privacy problems. Therefore, when implementing AVSE systems, particularly on embedded systems, it is essential to effectively reduce the size of visual input and user identifiability.

In this study, we propose an autoencoder (AE)-based compression network and a data compression scheme [27, 28] to address these two problems in AVSE. The AE-based compression network can significantly reduce the size of the visual input while extracting highly informative visual information. A data compression scheme is applied to further reduce the bits of the extracted representation. Because we focus on reducing both the data size and training network, we refer to the proposed method as the Lite AVSE (LAVSE) system[1]. We tested the proposed LAVSE system on the dataset of Taiwan Mandarin speech with video (TMSV)[2], an audio-visual version of Taiwan Mandarin hearing in noise test [29]. The dataset was recorded by 18 speakers (13 males and 5 females), each providing 320 video clips. The experimental results indicate that the proposed LAVSE system can yield better performance than an audio-only SE baseline. Moreover, it is confirmed that the user identity can be removed from the compressed visual data, thereby addressing the privacy problem.

The remainder of this paper is organized as follows. Related works are described in Section 2. The overall architecture of the proposed LAVSE system is introduced in Section 3. The experimental setup and results are presented in Section 4. Finally, the concluding remarks are provided in Section 5.

2. Related Works

In this section, previous research works related to the proposed LAVSE system are reviewed.

[1]https://github.com/kagaminccino/LAVSE
[2]https://bio-asplab.citi.sinica.edu.tw/Opensource.html#TMSV

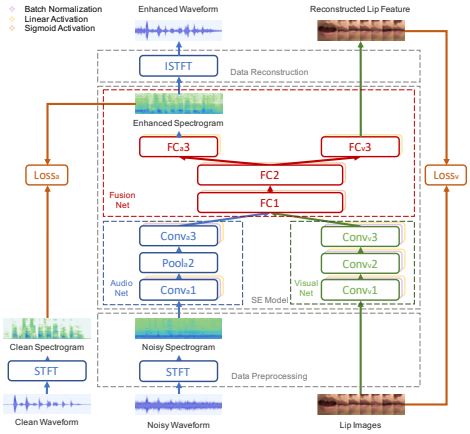


Figure 1: *The AVDCNN architecture[23].*

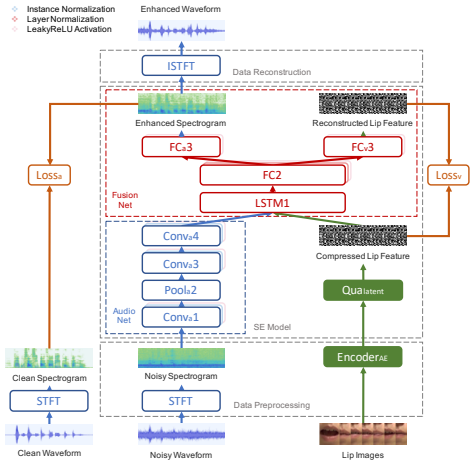


Figure 2: *The LAVSE architecture with two visual data compression units (EncoderAE and Qualatent).*

2.1. Audio-only SE

DL-based SE systems can be broadly divided into two categories: masking- and mapping-based. For both categories [30], paired audio data of clean and noisy speech utterances are prepared in the training phase. The noisy speech utterances are used as the input, and the goal is to obtain enhanced speech signals as the output. During training, a loss function is used, such as L1 and L2 norms, to measure the difference between the enhanced and clean signals. The DL-based model is trained with the aim of minimizing the loss function. Although the goals are the same, the implementations of the masking- and mapping-based SE approaches are different. The masking-based approaches estimate a mask and then multiply it with the noisy audio signal to generate an enhanced signal [31]. In contrast, mapping-based approaches directly estimate enhanced speech. According to the type of input, the methods are based on spectral mapping [15, 16, 17, 13], complex spectral mapping [32, 33], and waveform mapping [19, 34].

2.2. AVSE

The concept of AVSE is to incorporate visual input as auxiliary data into audio-only SE systems to provide complementary information. Numerous prior studies have verified the effectiveness of using visual input in improving SE performance [23, 24, 25, 26]. Among them, as depicted in Fig. 1, audio-visual speech enhancement using multimodal deep convolutional neural networks (AVDCNN) [23] receives noisy audio and lip images as the input and generates enhanced audio and lip images as the output. The combination of audio loss, Lossa and visual loss, Lossv (with specific weights) is used as the final loss to optimize the AVSE model. In this study, we used AVDCNN (with the audio and fusion net of the LAVSE system) as the basic AVSE system. We derived several algorithms to reduce the system's computation cost and storage requirements and address privacy problems.

2.3. Bit-wise Data Compression

The single-precision floating-point in IEEE 754 [35] is a data format widely used to represent 32-bit floating-point variables. The 32 bits consist of 1 sign bit, 8 exponential bits, and 23 mantissa bits. The sign bit indicates whether the value represented is positive or negative, the exponential bits determine the representation range of the value, and the mantissa bits account for the significant figures. The exponential and mantissa terms are not explicitly assigned and should be computed using addition and multiplication.

Because quantization on the mantissa bits does not change the represented value itself but only reduces the precision, a previous work suggests an exponent-only floating-point (EOFP) [28] format for audio SE tasks. Based on this design, the model size can be notably reduced, and the online computation efficiency can be effectively enhanced [36] while maintaining the SE performance in terms of speech quality and intelligibility.

3. Proposed LAVSE System

The goal of this study is to reduce the amount of storage, enhance the online computation efficiency, and address privacy problems of the AVDCNN system, which is depicted in Fig. 1. In this section, we first introduce the overall architecture of the proposed LAVSE system and then detail two approaches for reducing the amount of visual data.

3.1. Overall Architecture

In Fig. 2, the architecture of the LAVSE system is depicted, which includes three parts: data preprocessing, SE model, and data reconstruction. The two green solid blocks outline the proposed visual data compression. EncoderAE reduces the dimension of the visual data, whereas Qualatent conducts bit quantization.

Before training, noisy waveforms are transformed into audio features using the short-time Fourier transform (STFT). Lip images are fed into the AE model illustrated in Fig. 3, in the preprocessing stage, and the latent representations are used as the visual features. With EncoderAE, the visual feature extraction net can be fully removed from the training model. Next, in the SE model stage, the audio features are initially enhanced, whereas the visual features are processed by Qualatent to reduce the precision. The processed audio and visual features are then concatenated and fed into a fusion net, which is formed using LSTM and FC layers. Finally, enhanced audio features and restored visual features are obtained. With reference to clean audio data and compressed visual data, we can estimate the combined loss (from the audio and visual parts) and then use it to update the model parameters.

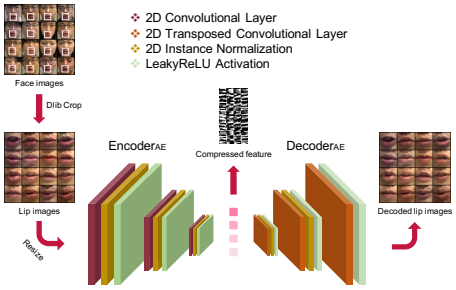


Figure 3: *The AE model for visual data compression.*

During testing, noisy waveforms and lip images first go through the preprocessing stage, and then they are fed into the trained SE model to generate enhanced spectrograms, which are then converted into the waveform domain through inverse STFT with the phases of the noisy audio in the reconstruction stage.

3.2. Visual Data Compression

The visual data compression module consists of two parts. The original lip images were first processed by EncoderAE. The AE model is trained in a self-learning manner in advance, which consists of three 2D convolutional layers as the encoder and three 2D transposed convolutional layers as the decoder. The dimensions of the latent representation of the AE are 32 $\times$ 8 $\times$ 8 which yield 2048 dimensions after flattened. This latent representation is used as the visual feature, and its size is only 16.67% of that of the original lip image (3 $\times$ 64 $\times$ 64).

Next, the extracted visual features are processed by Qualatent (based on EOFP) to further reduce the number of bits for each feature element. The compressed visual features are then used in the SE model. In real-world applications, the EncoderAE and Qualatent modules can be installed in the visual sensor so that the online computation efficiency is enhanced. Through the two-stage visual data compression process, the transmission cost can be reduced significantly and the lip images can be blurred out so that the matter of privacy can be addressed moderately.

4. Experiments

This section first introduces the experimental setup, including data preparation and system settings, and then presents the experimental results and discussions.

4.1. Experimental Setup

The TMSV dataset contained 5,760 speech utterances from 18 speakers, each providing 320 utterances. As there were only 5 females, we used 5 speakers from each gender to create a gender-balanced situation. The training set consisted of the 1st to the 200th utterance from each of the eight speakers (four males and four females), amounting to 1600 clean utterances in all. These utterances were corrupted by 100 types of noise [37] at five signal-to-noise ratios (SNRs) from -12 dB to 12 dB with a step of 6 dB. The test set consisted of the 201st to the 320th utterance from each of the other two speakers (one male and one female) for a total of 240 utterances. We designed a car driving application scenario for which we adopted six types of noise that are more common in car driving, including the cries of a baby, engine noise, background talkers, music, pink noise, and street noise, to form the noisy testing data. The SNR levels were set to -1, -4, -7, and -10 dB, which are considered challenging. Note that the speakers, noise types, and SNR levels are all mismatched in the training and testing settings.

Two standard evaluation metrics were used to evaluate the performance: perceptual evaluation of speech quality (PESQ) [38] and the short-time objective intelligibility measure (STOI) [39]. PESQ was used to evaluate the quality of speech, with a score ranging from -0.5 to 4.5. STOI was designed to evaluate the intelligibility of speech, with a score ranging from 0 to 1. Higher PESQ and STOI scores indicate that the enhanced speech has better speech quality and intelligibility, respectively.

We adopted the log1p magnitude spectrum as the audio feature for the SE model[3]. For the visual part, the lip or face image contours were positioned using Dlib [40], which is a 68-point facial landmark detector, and then compressed using EncoderAE. The audio-visual combined loss is $loss = loss_a + \mu \times loss_v$, where μ is empirically determined as 10^{-3} and the individual losses are mean square errors. We have attempted several visual feature extraction approaches to build the baseline AVSE system, which will be detailed in Section 4.2.1. The audio-only SE system was implemented for comparison. The total number of model parameters of the audio-only SE system were designed to be comparable to that of the AVSE system. For all the SE models, the optimizer is Adam [41] with a learning rate of 5×10^{-5}.

4.2. Experimental Results

4.2.1. LAVSE versus Audio-only SE

First, we intend to compare the performance of the proposed LAVSE with EncoderAE (LAVSE(AE)) with several baseline systems, including an audio-only SE (Audio-only) system and three AVSE systems with different visual data, namely, (1) AVSE(VGGface)—face features processed by VGGface [42], (2) AVSE(face)—raw face images, and (3) AVSE(lip)—raw lip images. The performances of these five systems are presented in Table 1, where the scores of the original noisy speech are also listed for comparison. The results indicate that Audio-only, AVSE(lip), and LAVSE(AE) can notably improve the speech quality (in terms of PESQ) and intelligibility (in terms of STOI). However, AVSE(VGGface) and AVSE(face) cannot yield further improvements compared to the Audio-only system. This may be because the entire face image may contain redundant information that does not directly benefit the SE task. It is also noted that LAVSE(AE) outperforms Audio-only and AVSE(lip) in terms of both PESQ and STOI, suggesting the positive effect of incorporating compressed visual lip information into the SE system. In the following discussion, we present only the results obtained using lip images as visual data.

Table 1: *PESQ and STOI scores of the LAVSE(AE) system and the baselines.*

	PESQ	STOI
Noisy	1.001	0.587
Audio-only	1.283	0.610
AVSE(VGGface)	0.797	0.492
AVSE(face)	1.270	0.616
AVSE(lip)	1.337	0.641
LAVSE(AE)	**1.374**	**0.646**

[3]According to our preliminary experiments, the log1p magnitude spectrum can yield better SE performance than the log power spectrum.



(a) *Original lip images.* (b) *AE reconstructed images.*

Figure 4: *Original and AE reconstructed lip images.*

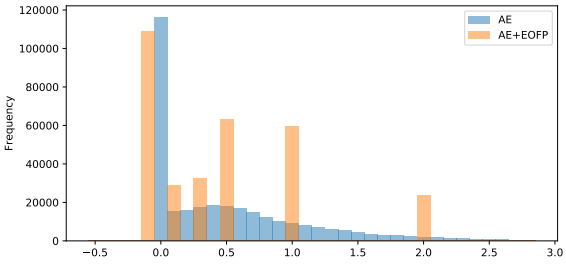


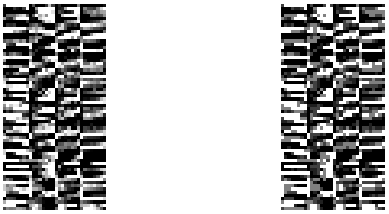
Figure 5: *The distributions of visual features before and after applying Qualatent.*

4.2.2. Visual Data Compression

Next, we investigate the effects of the two visual data compression modules: EncoderAE and Qualatent. The original lip images and the lip images that are put through the AE are presented in Fig. 4. Through visual inspection, we can affirm that the AE can encode and decode lip images well, suggesting that the latent representation can carry enough lip image information. The compression ratio of the EncoderAE module R_{AE} was 6 ($= (3 \times 64 \times 64) \div 2048$).

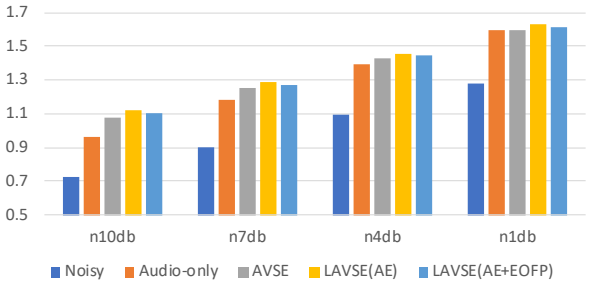
Subsequently, we investigate the effect of Qualatent by comparing the visual features before and after applying Qualatent. In this study, we applied Qualatent to reduce the original 32-bit data point to a 4-bit representation, including 1 sign bit and 3 exponential bits. The compressed representation was then used as the visual feature and put through the SE model. Note that by applying Qualatent to the latent representation, the sizes of the visual features were notably reduced——each feature element was reduced from 32 to 4 bits. The compression ratio of the Qualatent module R_{Qua} is 8 ($= (1+8+23) \div (1+3+0)$). The LAVSE system with both EncoderAE and Qualatent is termed LAVSE(AE+EOFP). By comparing the distributions presented in Fig. 5, we observe that when using Qualatent, the distribution of the compressed visual features (orange bars) covers the entire distribution of uncompressed features (blue bars) well. In Figs. 6a and 6b, the latent representations of lip features before and after applying Qualatent, respectively, are depicted. From the figures, we can easily see that the user identity has been removed almost completely, thereby moderately addressing the privacy problem.

Finally, by combining EncoderAE and Qualatent, the overall compression ratio, R_{Comp} is 48 ($= R_{\text{AE}} \times R_{\text{Qua}}$), confirming that the size of the visual input can be significantly reduced using the LAVSE(AE+EOFP) system. From Fig. 7, it is evident that even with such significant data compression, LAVSE(AE+EOFP) can still yield satisfactory enhancement performance of PESQ = 1.358 and STOI = 0.643, which are comparable to those of the best AVSE systems and notably better than those of the Audio-only system (refer to Table 1).

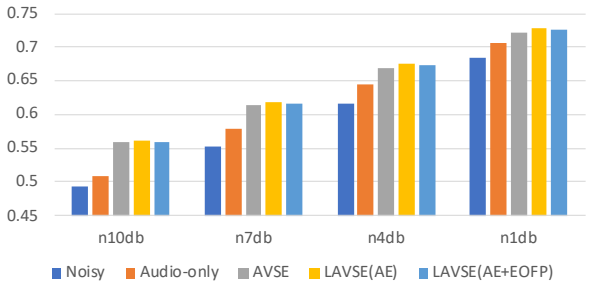


(a) *AE feature.* (b) *AE+EOFP feature.*

Figure 6: *Visual latent features of lips.*



(a) *PESQ.*



(b) *STOI.*

Figure 7: *PESQ and STOI scores at specific SNR levels.*

5. Conclusion

In this study, a LAVSE system with novel visual data compression approaches was proposed to address the problems that may be encountered when implementing an AVSE system for practical applications. We designed a lighter model structure by removing the visual feature extraction network from the training model. Compared to the data sizes of the original lip images, the data sizes of the visual features obtained using the combined EncoderAE and Qualatent modules are reduced by a significant factor of 48. Experimental results confirm that the AVSE system with the proposed visual data compression provides better performance than the Audio-only and AVSE systems without using visual data compression. The contributions of this study are threefold. First, we verified the effectiveness of incorporating visual information into SE. Second, we confirmed that even if the visual data is significantly reduced, it can still provide significant complementary information for the SE task. Third, using the EncoderAE and Qualatent modules together, privacy problems may be addressed as well. We expect that the findings in this study would be helpful for implementing the DL-based AVSE model in real-world applications. In the future, we will investigate time-domain compression to further enhance the online computation efficiency of the proposed LAVSE system.

6. References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[2] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP 1996*.

[3] J. Chen, J. Benesty, Y. A. Huang, and E. J. Diethorn, "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*. Springer, 2008, pp. 843–872.

[4] A. Rezayee and S. Gazor, "An adaptive klt approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 2, pp. 87–95, 2001.

[5] Y. Hu and P. C. Loizou, "A subspace approach for enhancing speech corrupted by colored noise," in *Proc. ICASSP 2002*.

[6] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. ICASSP 2012*.

[7] N. Yoma, F. McInnes, and M. Jack, "Lateral inhibition net and weighted matching algorithms for speech recognition in noise," in *Proc. VISP 1996*.

[8] G. Cocchi and A. Uncini, "Subband neural networks prediction for on-line audio signal recovery," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 867–876, 2002.

[9] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, "Nonlinear speech enhancement: An overview," in *Progress in nonlinear speech processing*. Springer, 2007, pp. 217–248.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR 2016*.

[11] 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*.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NIPS 2013*.

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

[14] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[15] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. Interspeech 2014*.

[16] 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 (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[17] M. Kolbæk, 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 (TASLP)*, vol. 25, no. 1, pp. 153–167, 2017.

[18] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian wavenet," in *Proc. Interspeech 2017*.

[19] 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, no. 9, pp. 1570–1584, 2018.

[20] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. ICASSP 2014*.

[21] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP 2015*.

[22] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for lstm-rnn based speech enhancement," in *Proc. HSCMA 2017*.

[23] 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.

[24] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Deep-learning-based audio-visual speech enhancement in presence of lombard effect," *Speech Communication*, vol. 115, pp. 38–50, 2019.

[25] M. L. Iuzzolino and K. Koishida, "Av (se) 2: Audio-visual squeeze-excite speech enhancement," in *Proc. ICASSP 2019*.

[26] W. Wang, C. Xing, D. Wang, X. Chen, and F. Sun, "A robust audio-visual speech enhancement model," in *Proc. ICASSP 2010*.

[27] S. Wu, G. Li, F. Chen, and L. Shi, "Training and inference with integers in deep neural networks," *arXiv preprint arXiv:1802.04680*, 2018.

[28] Y.-T. Hsu, Y.-C. Lin, S.-W. Fu, Y. Tsao, and T.-W. Kuo, "A study on speech enhancement using exponent-only floating point quantized neural network (eofp-qnn)," in *Proc. SLT 2018*.

[29] M. Huang, "Development of taiwan mandarin hearing in noise test," *Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.

[30] 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.

[31] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[32] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proc. MLSP 2016*.

[33] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *Proc. ICASSP 2020*.

[34] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[35] Institute of Electrical and Electronics Engineers, *IEEE Standard for Binary Floating-point Arithmetic*. ANSI/IEEE Std 754-1985, 1985.

[36] Y.-C. Lin, Y.-T. Hsu, S.-W. Fu, Y. Tsao, and T.-W. Kuo, "Ia-net: Acceleration and compression of speech enhancement using integer-adder deep neural network," in *Proc. Interspeech 2019*.

[37] G. Hu, "100 nonspeech environmental sounds," 2004, available: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[38] 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 2001*.

[39] 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.

[40] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[42] O. M. Parkhi, A. Vedaldi, and A. Zisserman, *Deep face recognition*. British Machine Vision Association, 2015.