Audio-Visual Speech Enhancement using Hierarchical Extreme Learning Machine

Tassadaq Hussain[‡]
Taiwan International Graduate Program in Social Network and Human-Centered Computing
Institute of Information Science
Academia Sinica
Taipei, Taiwan
tass.hussain@iis.sinica.edu.tw

Yu Tsao
Research Center for Information Technology Innovation
Academia Sinica
Taipei, Taiwan
yu.tsao@citi.sinica.edu.tw

Hsin-Min Wang
Institute of Information Science
Academia Sinica
Taipei, Taiwan
whm@iis.sinica.edu.tw

Jia-Ching Wang
Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
jcw@csie.ncu.edu.tw

Sabato Marco Siniscalchi
Department of Computer Engineering
Kore University of Enna
Enna, Italy
marco.siniscalchi@unikore.it

Wen-Hung Liao
[‡]*Department of Computer Science*
National Chengchi University
Taipei, Taiwan
whliao@cs.nccu.edu.tw

***Abstract*—Recently, the hierarchical extreme learning machine (HELM) model has been utilized for speech enhancement (SE) and demonstrated promising performance, especially when the amount of training data is limited and the system does not support heavy computations. Based on the success of audio-only-based systems, termed AHELM, we propose a novel audio-visual HELM-based SE system, termed AVHELM that integrates the audio and visual information to confrontate the unseen non-stationery noise problem at low SNR levels to attain improved SE performance. The experimental results demonstrate that AVHELM can yield satisfactory enhancement performance with a limited amount of training data and outperforms AHELM in terms of three standardized objective measures under matched and mismatched testing conditions, confirming the effectiveness of incorporating visual information into the HELM-based SE system.**

***Index Terms*—Speech Enhancement, Hierarchical Extreme Learning Machine, Audio-Visual, Multi-Modal**

I. Introduction

Background noises severely deteriorate the quality, and the intelligibility of a speech signal limiting the applicability of speech-related applications in real-world conditions. Numerous approaches have been proposed in the past to mitigate the background noise problem [1]–[4]; a notable class is speech enhancement (SE). The goal of SE is to generate an enhanced speech with better speech quality and intelligibility by suppressing the background noise components in the noisy speech. Recently, data-driven-based SE approaches have demonstrated significant success. For these approaches, a mapping function that aims to transform noisy speech to clean speech is realized by a machine-learning-based model. The model is trained using noisy-clean utterances pairs. As the relation between noisy and clean speech is typically complicated, adopting a nonlinear mapping function can yield better SE performance.

Artificial neural network (ANN) models with deep structures are suitable candidates to characterize the nonlinear mapping function. Thus far, many ANN models have been adopted for the SE task and demonstrated outstanding performance; for example, deep feedforward neural networks [5] [6], deep denoising autoencoders (DDAE) [7] [8], long short-term memory recurrent neural network (LSTM) [9] [10], and convolutional neural network (CNN) [11] [12]. Despite the notable improvements over traditional methods, ANN-based SE approaches typically require a sufficient amount of training data for an effective backpropagation process. More recently, the authors in [13]–[15] employed the ELM and the hierarchical extreme learning machine (HELM) as an alternative model for the SE task. In the training stage of the HELM, the model first converts the original speech into a high-dimensional sparse representation using multiple layers of ELM-based autoencoders. Next, a mapping function is estimated to transform the sparse representation to the target clean speech. In the test phase, the noisy utterance is processed by the same multiple layers of the ELM-based autoencoders to obtain the enhanced utterance. Since there is no backpropagation involved, the HELM-based SE system can perform well even when the amount of training data is limited. Moreover, because no fine-tuned model is involved, the HELM-based SE system can be installed in devices that cannot support heavy computations.

Studies have shown that visual modality carries important information, such as lip motions and mouth articulations that can help discriminate similar speech sounds in noisy conditions [16]–[18]. Recently, several SE methods that in-

tegrate audio and visual information have been proposed. For example, in [19], [20], feedforward and convolutional neural network models were used to build an audio-visual SE system and have improved the noise reduction performance successfully compared to audio-only frameworks. In [21], the authors proposed a deep learning-based framework to investigate the impact of Lombard effect on the performance of the audio-visual speech enhancement system. In [22], a speech separation system was proposed that incorporated audio-visual information using a deep network-based model.

In this work, we extend our audio-only HELM (AHELM) framework [14], by incorporating the visual cues alongside noisy speech signals for spectral mapping to further improve the system performance. The experimental results demonstrated that the audio-visual (AV) integration for HELM produced better performance as compared to the AHELM SE system under both matched and mismatched (stationery and non-stationery noises) testing conditions at severe signal-to-noise ratio (SNR) levels using limited training data in terms of the three standardized objective measures, including the perceptual evaluation of speech quality (PESQ) [23], hearing aid speech perception index (HASPI) [24], and segmental signal-to-noise ratio improvement (SSNRI) [25].

The remainder of this paper is organized as follows: Section II introduces the proposed AVHELM SE system. Section III presents the experiential setup and evaluation results. Section IV provides concluding remarks of this work.

II. Audio-Visual Speech Enhancement System

A. HELM-based Speech Enhancement System

The hierarchical framework of the ELM was proposed by Tang et al. [26] to further exploit the universal capability of the ELM, and extract more abstract information in a multi-layer manner. In an HELM system, the sparse representation of the input data is extracted using the ELM-based sparse autoencoders (i.e., unsupervised stage). The output of the unsupervised stage is fed subsequently to the ELM-based supervised regression/classification stage for the ultimate decision making. Fig. 1 shows the conventional HELM-based speech enhancement framework. The aim is to learn the spectral mapping from the noisy speech to the clean speech. During the training phase, the short-time Fourier transform (STFT) is applied to the speech signals. The logarithmic power spectral (LPS) features of the noisy and clean speech spectra are subsequently estimated. The noisy-clean speech spectra are later processed by the HELM framework to learn the spectral mapping from the noisy to clean speech signals.

In the testing stage, the noisy LPS features are processed by the HELM model to obtain enhanced LPS features. The phase of the original noisy speech is used to obtain the denoised speech waveforms alongwith the inverse STFT operations. More details can be found in [14].

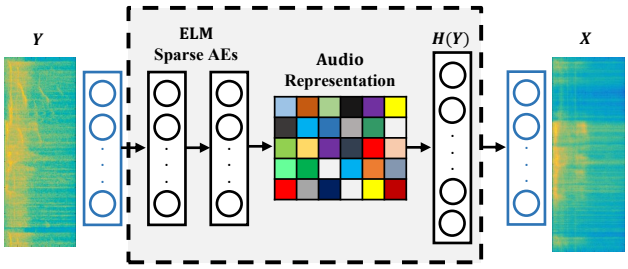

Fig. 1: Audio-only HELM-based SE framework.

B. HELM-based Audio-Visual Speech Enhancement System

In this section, we extend the AHELM framework by considering the visual information and propose an HELM-based audio-visual speech enhancement (AVHELM) framework. In the AVHELM framework the visual information is incorporated with the speech information to further improve the enhancement capability of the system. Fig. 2 illustrates the proposed AVHELM framework. In this system, the AV features are processed independently through the HELM sparse autoencoders (i.e., unsupervised stage) to learn the sparse representation of the noisy audio features and visual information individually. The outputs of the two modalities from the unsupervised stage are subsequently combined to form an integrated input to the supervised regression stage, as shown in Fig. 2. For AHELM, the relationship between input (noisy) and output (enhanced) can be written as:

$$\boldsymbol{X} = \boldsymbol{H}(\boldsymbol{Y})\ \boldsymbol{B}_a \tag{1}$$

where $\boldsymbol{H}(\boldsymbol{Y})$ is the hidden layer output matrix for input noisy speech signal $\boldsymbol{Y}$, $\boldsymbol{B}_a$ is the output weight matrix for audio-only HELM, and $\boldsymbol{X}$ is the estimated clean speech signal as shown in Fig. 1. Similarly, the estimated clean speech signal for the AVHELM framework can be computed by integrating the audio and visual information such as:

$$\boldsymbol{X} = [\boldsymbol{H}(\boldsymbol{Y}),\ \boldsymbol{H}(\boldsymbol{V})]\ \boldsymbol{B}_{av} \tag{2}$$

where $\boldsymbol{H}(\boldsymbol{Y})$ and $\boldsymbol{H}(\boldsymbol{V})$ are the corresponding hidden layer output matrices for audio and visual modality, $\boldsymbol{B}_{av}$ is the output weight matrix for integrated audio-visual information, and $\boldsymbol{X}$ is the estimated speech signal as shown in Fig. 2.

III. Experiments and Results

A. Experimental Setup

1) Dataset Description: The audio and visual information were recorded and prepared by Hou et al. [19] based on the transcript of the Taiwan Mandarin hearing in noise test (TMHINT) sentences [27]. The dataset contained audio-visual recordings of 320 Mandarin utterances spoken by a native speaker and were recorded in a quiet room with sufficient lighting, where the speaker was filmed facing towards the

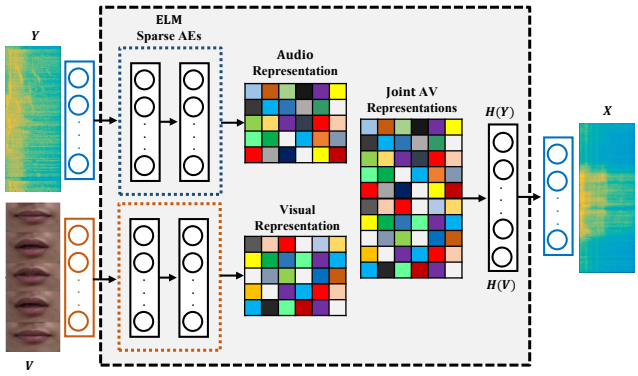

Fig. 2: Proposed AVHELM SE framework.

camera. The length of each utterance was approximately 3-4 seconds. The visual information was recorded at 30 frames per second (fps), at a resolution of 1920 pixels $\times$ 1080 pixels. The audio data were recorded at 48 kHz and resampled to 16 kHz for further processing. Among the 320 utterances, we randomly selected 100 utterances for the training set and 40 random utterances for the testing set, noting that no overlap occurred between the training and testing utterances. We used nine different stationery and non-stationery background noises, namely *machine*, *pink noise*, *babble*, *baby cry*, *party crowd*, *applause*, *grocery store*, *restaurant*, and *vacuum cleaner* as background noises to prepare the training and testing data. For the training set, the clean utterances were artificially contaminated with one stationery and four non-stationery noises, namely *machine*, *babble*, *party crowd*, *restaurant*, *vacuum cleaner* at 5 different signal-to-noise ratios (SNRs) $\in$ {-6, -3, 3, 6, 10 dB} to generate 100 $\times$ 5 (noise types) $\times$ 5 (SNRs) = 2500 noisy training utterances. To confirm the effectiveness of our proposed AVHELM system, two evaluation scenarios were adopted to design the testing sets: Matched noise type, and mismatched noise type. In the matched case, the clean testing utterances were contaminated with two matched noises namely, *party crowd* and *babble* at the matched and mismatched SNRs $\in$ {-6, -2, 0, 2, and 6 dB}, to that used in the training set. In the mismatched case, the clean testing utterances were contaminated with four mismatch noise types namely, *applause*, *baby cry*, *pink noise*, and *grocery store* at the matched and mismatched SNRs $\in$ {-6, -2, 0, 2, and 6 dB}, respectively.

The proposed system was evaluated using three standard objective evaluation metrics: PESQ, HASPI, and SSNRI.

2) Audio-visual Feature Extraction: The audio speech signals were processed using STFT with a frame length of 512 samples, and a frame shift of 256 samples. We varied the size of the input speech vector by considering more contexts at the input layer. In this work, we used $\pm$ 2 neighbouring speech vectors in the left and right alongside the central speech vector, similar to [19], generating LPS features of dimensions 257 $\times$ 5 (= 257 $\times$ (ws $\times$ 2 + 1), where ws is the contextual window size, and $ws = 2$ was used in our experiments).

For the visual information, we used the same visual features as that in [19], by converting each video of the corresponding utterance into a sequence of images at a frame rate of 50 fps. The mouth part of each image was subsequently detected using the Viola—Jones method [28] and was cropped into a 16 pixels $\times$ 24 pixels region, thus resulting in visual features of dimensions 16 $\times$ 24 $\times$ 3 $\times$ 5, where 3 is its RGB channel and 5 is the neighboring visual vectors including the left and right alongside the central visual vector.

3) AVHELM vs AHELM: In this section, we first evaluate the overall performance of the proposed AVHELM against AHELM. Table 1 compares the average PESQ results between the proposed AVHELM and AHELM structures under matched and mismatched testing conditions. For fair comparison, the two frameworks were trained using 1000, 1000, and 8000 neurons ([1000 1000 8000]). For the two HELM configurations, the sigmoidal activation function was employed with the regularization parameter equal to that used in [14]. In our preliminary experiments, we found that with such a small amount of training data, both the DNN- and CNN-based AVSE systems [19] cannot perform well. To focus our attention to HELM-based SE systems, the CNN-based results are not included in this study. For the AVHELM, each modality was processed independently by the unsupervised stage to convert low-level features to representative features. During the supervised stage, the representations learned by both modalities were integrated linearly to learn the multimodal transformation. By looking at Table 1, we can argue that the proposed AVHELM framework improved the speech quality (PESQ) for both matched and mismatched (stationery and non-stationery) noise types. The performance of AVHELM and AHELM is further compared against a traditional logarithmic minimum mean square error (logMMSE) [29] method. It is clear from Table 1, that the AVHELM and AHELM frameworks with similar configurations attained a significant performance improvement compared to logMMSE under testing conditions, except for most *pink noise*, where logMMSE as a powerful traditional method remains its advantage under mismatched stationery noise condition and performs better compared to AHELM and AVHELM frameworks.

Table 1 shows the average PESQ performance comparison between the two HELM systems under matched and

TABLE I: AVERAGE PESQ SCORES OF LOGMMSE, AHELM, AND AVHELM UNDER MATCHED AND MISMATCHED NOISE CONDITIONS.

Condition	Noise Type	logMMSE	AHELM	AVHELM
Matched	Babble	2.2138	2.2932	**2.4066**
	Crowd party	2.1473	2.3163	**2.4054**
Mismatched	Applause	1.8963	2.3873	**2.5104**
	Baby cry	1.9725	2.6812	**2.7633**
	Grocery store	2.0986	2.3083	**2.4948**
	Pink noise	**2.5774**	2.3639	2.5125

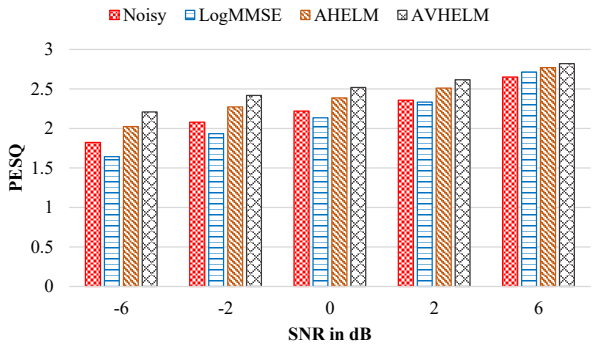

Fig. 3: Average PESQ scores over six noise types at different SNR levels.

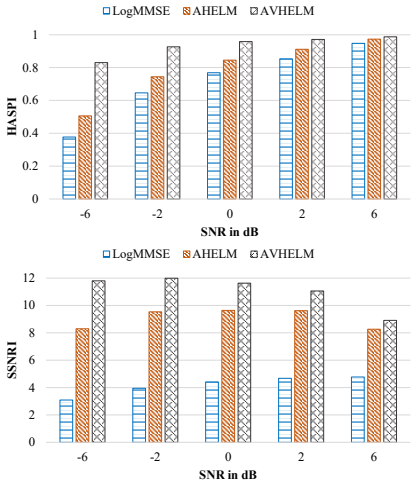

Fig. 4: Average HASPI and SSNRI scores over six noise types at different SNR levels.

mismatched testing conditions. However, the table failed to describe which modality weighted more while reconstructing the denoised speech signal at different SNR levels. Therefore, we plotted the average PESQ performance at specific SNR levels with the aim to further investigate the behaviors of the two HELM frameworks. Fig. 3 presents the average PESQ performance for the six noise types at different SNR levels for AHELM and AVHELM. The results of unprocessed speech (denoted as Noisy) and logMMSE are also listed for comparison. We observe that the proposed AVHELM framework obtained a significant performance improvement while handling low SNRs (i.e., -6 and -2 dB). The figure illustrates the behavior of the AVHELM framework by demonstrating that the framework relied more on the visual information by obtaining some guidance while handling low SNRs and making a decision. However, the visual information did not provide much help or guidance to the AVHELM system while handling high SNRs. The difference between the average PESQ score for AHELM and AVHELM gets smaller for high SNRs (2 dB and 6 dB), indicating that the audio modality played crucial role in high SNRs for decision making.

In addition to the PESQ scores, we also reported the average HASPI and SSNRI results for the AVHELM and AHELM frameworks beside logMMSE. Fig. 4 displays the average HASPI and SSNRI results for the six noise types at different SNR levels. Overall, the proposed AVHELM demonstrated better speech enhancement capabilities compared to the AHELM and logMMSE by maintaining high scores for HASPI and SSNRI. However, an obvious performance improvement can be seen at low SNR levels to that of at high SNR levels, again confirming that the visual modality played a crucial role while reconstructing a signal at low SNR levels.

To better appreciate the SE performance attained by the proposed model, we plotted the spectrogram of the enhanced speech signals yielded by the AHELM and AVHELM. Fig. 5 shows the spectrogram of a test utterance contaminated with a non-stationery noise *applause* at SNR = -2 dB. Fig. 5(c) and (d) display the spectrogram of the test utterance enhanced by AHELM and AVHELM frameworks. The spectrograms of clean and noisy speech signals are also illustrated in Fig. 5(a) and (b) for comparison. From Fig. 5, we note that although both AHELM and AVHELM perfectly restore clean speech under very challenging condition (non-stationery, -2 dB SNR), they can effectively suppress noise components from the noisy signal (Fig. 5(b)). Moreover, the AVHELM more effectively suppresses noise components and yields better speech quality (PESQ = 2.7784) compared to the AHELM (PESQ = 2.6496). In addition to spectrogram plots, we also plotted the waveforms to visually investigate the speech processed by AHELM and AVHELM. Fig. 6 (a), (b), (c), and (d) show the waveforms of Clean, Noisy, AHELM, and AVHELM speech, respectively, where the test utterance is contaminated with *applause* noise at SNR = -2 dB. From the figure, we observe that the waveform of the denoised speech yielded by AVHELM displays a similar pattern as clean speech with less distortion even at low SNR (SNR = -2 dB), illustrating that AVHELM can more effectively restore clean speech from the noisy counterpart.

IV. Conclusion

We herein proposed a novel AVHELM framework to improve the performance of the conventional AHELM framework. The results demonstrate that incorporating the visual modality/information increases the system performance under both matched and mismatched (also both stationery and non-stationery) noise conditions at severe SNR levels when limited training data is available. To the best of our knowledge, this is the first work that successfully applies HELM for audio-visual speech enhancement. In our future work, we aim to further enhance the system performance by considering noise- and SNR- aware training.

V. Acknowledgment

This work was partly supported by MOST Taiwan Grants 108-2634-F-008-004-.

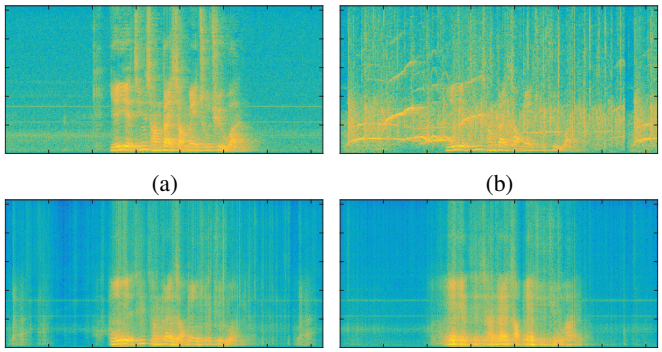

Fig. 5: Spectrograms of (a) Clean, (b) Noisy, (c) AHELM, and (d) AVHELM. The test utterance was contaminated with noise *applause* at SNR = -2 dB.

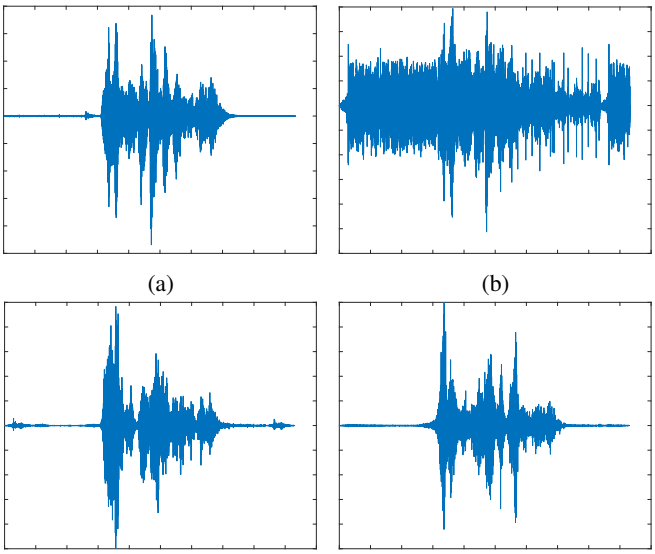

Fig. 6: Waveforms of (a) Clean, (b) Noisy, (c) AHELM, and (d) AVHELM. The test utterance was contaminated with noise *applause* at SNR = -2 dB.

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] 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.

[3] A. El-Solh, A. Cuhadar, and R. Goubran, "Evaluation of speech enhancement techniques for speaker identification in noisy environments," in *Proc. ISMW*, 2007, pp. 235–239.

[4] J. Ortega-García and J. González-Rodríguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Proc. ICSLP*, 1996, pp. 929–932.

[5] 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, no. 1, pp. 7–19, 2015.

[6] 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, no. 1, pp. 153–167, 2017.

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

[8] B. Xia and C. Bao, "Speech enhancement with weighted denoising auto-encoder." in *INTERSPEECH*, 2013, pp. 3444–3448.

[9] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. LVA/ICA*, 2015, pp. 91–99.

[10] 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, pp. 136–140.

[11] 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.

[12] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement." in *Interspeech*, 2016, pp. 3768–3772.

[13] B. O. Odelowo and D. V. Anderson, "A framework for speech enhancement using extreme learning machines," in *Proc. ACSSC*, 2017, pp. 1956–1960.

[14] T. Hussain, S. M. Siniscalchi, C.-C. Lee, S.-S. Wang, Y. Tsao, and W.-H. Liao, "Experimental study on extreme learning machine applications for speech enhancement," *IEEE Access*, vol. 5, pp. 25 542–25 554, 2017.

[15] T. Hussain, Y. Tsao, S. M. Siniscalchi, J.-C. Wang, H.-M. Wang, and W.-H. Liao, "Bone-conducted speech enhancement using hierarchical extreme learning machine,," in *Proc. IWSDS 2019, to be published*.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 689–696.

[17] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *Proc. ICASSP*, 2015, pp. 2130–2134.

[18] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, and S. Hayamizu, "Audio-visual speech recognition using deep bottleneck features and high-performance lipreading," in *Proc. APSIPA*, 2015, pp. 575–582.

[19] 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.

[20] A. Gabbay, A. Shamir, and S. Peleg, "Visual speech enhancement," in *Proc. INTERSPEECH*, 2018, pp. 1170–1174.

[21] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Effects of lombard reflex on the performance of deep-learning-based audio-visual speech enhancement systems," in *Proc. ICASSP*, 2019, pp. 6615–6619.

[22] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *arXiv preprint arXiv:1804.03619*, 2018.

[23] 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, pp. 749–752.

[24] J. M. Kates and K. H. Arehart, "The hearing-aid speech perception index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, 2014.

[25] J. Chen, J. Benesty, Y. Huang, and E. Diethorn, "Fundamentals of noise reduction in spring handbook of speech processing," *Springer*, 2008.

[26] J. Tang, C. Deng, and G.-B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 4, pp. 809–821, 2016.

[27] 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.

[28] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[29] 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.