Compressed Multimodal Hierarchical Extreme Learning Machine for Speech Enhancement

Tassadaq Hussain*‖ Yu Tsao† Hsin-Min Wang‡ Jia-Ching Wang§ Sabato Marco Siniscalchi¶ and Wen-Hung Liao‖
* Taiwan International Graduate Program in Social Network and Human-Centered Computing
Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: tass.hussain@iis.sinica.edu.tw
† Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
E-mail: yu.tsao@citi.sinica.edu.tw
‡ Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: whm@iis.sinica.edu.tw
§ Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan
E-mail: jcw@csie.ncu.edu.tw
¶ Department of Computer Engineering, Kore University of Enna, Enna, Italy
E-mail: marco.siniscalchi@unikore.it
‖ Department of Computer Science, National Chengchi University, Taipei, Taiwan
E-mail: whliao@cs.nccu.edu.tw

***Abstract*—Recently, model compression that aims to facilitate the use of deep models in real-world applications has attracted considerable attention. Several model compression techniques have been proposed to reduce computational costs without significantly degrading the achievable performance. In this paper, we propose a multimodal framework for speech enhancement (SE) by utilizing a hierarchical extreme learning machine (HELM) to enhance the performance of conventional HELM-based SE frameworks that consider audio information only. Furthermore, we investigate the performance of the HELM-based multimodal SE framework trained using binary weights and quantized input data to reduce the computational requirement. The experimental results show that the proposed multimodal SE framework outperforms the conventional HELM-based SE framework in terms of three standard objective evaluation metrics. The results also show that the performance of the proposed multimodal SE framework is only slightly degraded, when the model is compressed through model binarization and quantized input data.**

I. Introduction

In real-world conditions, background noise can severely degrade the quality and intelligibility of speech signals, thereby limiting the development of speech related applications [1]–[7]. Numerous signal processing-based speech enhancement (SE) methods have been proposed in the past to alleviate the background noise problem [8]–[11]. While these methods have been applied to improve the intelligibility for both human listening and machine recognition, the results have not always been satisfactory especially in regards to real acoustic conditions. Recently, approaches based on nonlinear spectral mapping have been proposed and confirmed to be effective in many SE tasks. The mapping function for these approaches aims to transform noisy speech to clean speech and is generally realized by a machine learning-based model. Several studies have been conducted to investigate the potential of deep-learning-based models with fine-tuned parameters for SE. For these approaches, a set of noisy and clean utterances is required to train the deep models. For example, the authors of [12] [13] proposed frameworks based on deep neural networks and deep denoising autoencoder (DDAE) to perform SE in non-stationary noise conditions. In [14] and [15], convolutional neural networks were used to transform noisy logarithmic power spectra (LPS) features and complex spectral features to their clean counterparts, respectively. Similarly, in [6] and [16], SE systems based on long short-term memory and recurrent neural networks were proposed to reduce the noise effects effectively. Although these deep-learning-based approaches have achieved state-of-the-art performance, they have the following limitations: (a) mismatched training/test conditions can severely deteriorate the system performance, and (b) a large amount of training data is required to achieve satisfactory generalization performance, which may limit the applicability of these frameworks in real-world scenarios.

To overcome the limitations of both conventional signal processing and deep-learning-based SE approaches, in our previous work [17] [18], we have proposed an alternative SE framework by adopting a hierarchical structure of the extreme learning machine (ELM) model. The parameters of the feature extraction layers of the hierarchical ELM (HELM) do not need to be fine-tuned using back propagation algorithms, thereby providing an extremely fast training phase with good generalization performance and general approximation capability.

Recent studies have shown that visual modalities, such as lip motions and mouth articulations, carry important information that can help distinguish similar speech sounds under noisy conditions [19]–[21]. Several audio-visual methods have been proposed recently to learn multimodal features for SE tasks using multimodal learning strategies. In [22], [23], feedforward and convolutional neural network models were used to build

an audio-visual SE system, which successfully improved the noise reduction performance compared with that of audio-only frameworks. In [24], a speech separation system was proposed, which used a deep-learning-based model to combine audio-visual information. Meanwhile, Li et al. proposed a cross-modal student-teacher learning framework to fully utilize the audio-visual information to attain improved speech recognition performance under challenging conditions [25].

In this work, we extend our previously proposed HELM-based SE framework [17], which adopts audio information only (thus termed HELM$_a$), by incorporating a visual modality to further improve the SE performance. The proposed HELM-based audio-visual SE framework, termed HELM$_{av}$, first processes the audio and visual modalities separately and then learns multimodal features and an output weight matrix. In addition to the state-of-the-art performance achieved by the deep-learning-based techniques in different classification and regression tasks, a considerable amount of research has been done on quantization-based model compression strategies to improve the computational capability of deep-learning-based systems for efficient online learning without degrading much of system's overall performance [26]–[28]. Motivated by the satisfactory performance achieved by the model compression strategies for back-propagation-based methods, we employ binarization and quantization schemes to train the feed-forward only framework (HELM$_a$ and HELM$_{av}$) for efficient learning using binary weights and quantized data. The experimental results demonstrate that the introduction of visual modality can improve the performance compared with that of HELM$_a$ in terms of three standardized objective measures: the perceptual evaluation of speech quality (PESQ) [29], hearing aid speech perception index (HASPI) [30], and segmental signal-to-noise ratio improvement (SSNRI) [31]. The results also show that by binarizing the weights (limiting the weights to +1 and -1) and quantizing the input data (representing the mantissa bits in a single floating-point number with fewer bits), the proposed framework still operates well and the overall SE performance of the system is only marginally affected.

The remainder of this paper is organized as follows: Section 2 introduces the proposed HELM$_{av}$ SE system as well as the model binarization and input data quantization schemes. Section 3 presents the experiential setup and results. Section 4 provides concluding remarks.

II. Proposed Method for Speech Enhancement

A. HELM-based Multimodal System for Speech Enhancement

To improve the learning capability of ELM further, Tang et al. introduced HELM by maintaining the unique and effective characteristics of the ELM [32]. The HELM has two stages: an unsupervised stage and a supervised regression/classification stage. In the unsupervised stage, a stack of ELM-based AEs is used to extract sparse and informative representations from the input data. The output of the unsupervised stage is subsequently processed by the supervised regression/classification stage for making the ultimate decision.

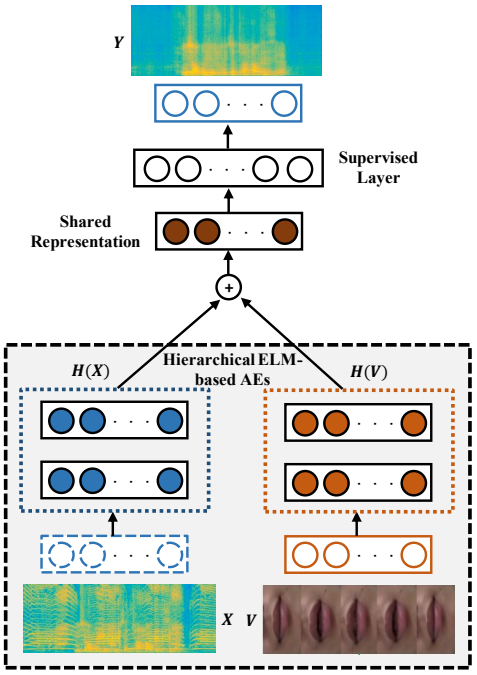

Fig. 1: The HELM-based multimodal SE framework.

In our previous HELM-based SE (HELM$_a$) framework [17], during the offline phase, the LPS features of the noisy and clean speech spectra were initially estimated and subsequently processed by HELM to learn the mapping function. In the testing phase, the noisy LPS features were processed by the HELM model to generate the enhanced LPS features. The phase of the original noisy speech was used to obtain the denoised speech waveform. Fig. 1 presents the architecture of the proposed HELM$_{av}$ framework, where the outputs of the two modalities from the two independent hierarchical ELM-based AEs are subsequently combined and fed into the supervised regression stage to learn the joint multimodal representation and output weight matrix. For HELM$_a$, the relationship between the noisy and estimated speech signals is written as

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

where $\boldsymbol{H}(\boldsymbol{X})$ is the hidden layer output matrix for the input noisy speech signal $\boldsymbol{X}$, $\boldsymbol{Y}$ is the estimated speech signal, and $\boldsymbol{B}_a$ is the output weight matrix for HELM$_a$.

On the other hand, the estimated speech signal for the HELM$_{av}$ framework can be computed by combining the audio and visual information as

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

where $\boldsymbol{H}(\boldsymbol{X})$ and $\boldsymbol{H}(\boldsymbol{V})$ are the hidden layer output matrices of the audio and visual modalities, respectively, $\boldsymbol{B}_{av}$ is the output weight matrix for the integrated audio-visual information, and $\boldsymbol{Y}$ is the estimated speech signal.

B. Binarization and Quantization

Although HELM already provides an extremely fast training phase with good generalization performance and a universal approximation capability, we can apply model compression strategies, namely binarization and quantization, to reduce the computational requirement further. In this study, we train our frameworks ($HELM_a$ and $HELM_{av}$) by limiting the real-valued weights to either binary values, i.e., -1 or +1 ({-1, +1}) or ternary values, i.e., {-1, 0, +1}. To transform the real-valued weights to binary or ternary weights, we use the criteria suggested in [33] and [34] as follows

$$w_b = \begin{cases} +1 & if\ w \geq 0 \\ -1 & otherwise, \end{cases} \tag{3}$$

where w_b is the binary weight, and w is the real-valued weight. In our implementation, we used the "*hard sigmoid*" activation function rather than the "*soft sigmoid*" activation function:

$$\sigma(x) = clip(0.5 * x + 0.5, 0, 1). \tag{4}$$

For ternary weight generation, we require a third quantized value to represent the weight. Accordingly, we used a threshold Δ to quantize the weight into $\{-1, 0, +1\}$:

$$w_t = \begin{cases} +1: & if\ w_t \geq \Delta \\ 0: & if\ \mid w_t \mid \leq \Delta \\ -1: & if\ w_t < -\Delta, \end{cases} \tag{5}$$

where w_t is the ternary weight. In our experiments, the threshold Δ was set to 0.5.

Subsequently, we quantize the input data with fewer precision bits. The objective is to construct a computationally efficient multimodal HELM framework with low-precision input data in order to reduce the computational requirement of HELM without affecting the performance. Typically, the value of a parameter of the input data is represented in IEEE 754 [35] single-precision floating-point format. The IEEE 754 binary format consists of 32 bits: the most significant bit or the sign bit (i.e., the bit at position 31) which represents the sign (0 indicates positive and 1 indicates negative), 8 exponent bits (bit positions 30 to 23) which represent the exponent part, and 23 fraction (or mantissa) bits (bit positions 22 to 0). More information about the quantization can be found in the experimental evaluation section.

III. Experimental Evaluations

A. Experimental Setup

The dataset used to evaluate the performance of the proposed multimodal HELM framework is the same as that prepared and used by Hou et al. [22], which contains the video recordings of 320 Mandarin utterances spoken by a native speaker. The recordings were based on the transcript of the sentences from the Taiwan Mandarin hearing in noise test (TMHINT) sentences [36]. The video was recorded at a frame rate of 30 frames per second (fps) and at a resolution of 1920 pixels × 1080 pixels whereas the audio was recorded at a sampling rate of 48 kHz, which was subsequently down-sampled to 16kHz for further processing. We selected 100 utterances from the corpus as the training set, and 40 utterances as the testing set. There was no overlap between the training and testing utterances. The training and testing utterances were subsequently contaminated with stationary and non-stationary noise types at different signal-to-noise ratio (SNR) levels. To verify the effectiveness of the proposed multimodal HELM framework, we used three noise types, i.e., *restaurant*, *babble*, and *party crowd*. The clean training utterances were artificially contaminated with these three noise types at four different SNR levels (SNR ∈ {-6, -3, 3, 6 dB}) to generate 100 × 3 (noise types) × 4 (SNRs) = 1200 noisy utterances. Two scenarios were considered to prepare the test set: matched and mismatched conditions. In the matched condition scenario, the aforementioned 40 testing utterances were contaminated with two matched noise types, namely *babble* and *party crowd*, at two matched SNRs, i.e., SNR ∈ {-6, 6 dB}, and three mismatched SNRs, i.e., SNR ∈ {-2, 0, 2 dB}. In the mismatched condition scenario, the clean testing utterances were contaminated with three unseen non-stationary noises, namely *applause*, *baby cry*, and *grocery store*, and one unseen stationary *Pink* noise at two matched SNRs, i.e., SNR ∈ {-6, 6 dB}, and three mismatched SNRs, i.e., SNR ∈ {-2, 0, 2 dB}.

The performance of the proposed framework was evaluated based on three objective evaluation metrics: PESQ, SSNRI, and HASPI. Higher scores of these metrics indicate better speech quality, speech SNR, and intelligibility, respectively.

B. Audio-visual Feature Extraction

Our preliminary results showed that considering ± 2 neighboring speech vectors could achieve better performance, generating LPS features of dimensions 257 × (*ws* × 2 + 1), where *ws* is the contextual window size: *ws* = 2 was used in our experiments. The visual features were the same as those prepared by [22], in which the visual information was processed in the form of an image sequence at a frame rate of 50 fps to be synchronized with the speech utterance frames. The images were subsequently cropped into an area of 16 × 24 pixels to extract mouth shape features by detecting the mouth part using the Viola-Jones method [37]. The corresponding visual information thus provided visual features with dimensions 16 × 24 × 3 × 5, where 3 represents its RGB channel and 5 is the neighboring visual vectors.

1) $HELM_{av}$ vs $HELM_a$ Analysis: We first evaluate the performance of the proposed $HELM_{av}$ framework against those of the previous $HELM_a$ framework, existing conventional SE methods such as log minimum mean square error (log-MMSE) [9], subspace-based Karhunen Loeve transform (KLT) [38], and robust principal component analysis (RPCA) [39]. $HELM_a$ only utilizes the audio to learn the spectral mapping function. For a fair comparison, both $HELM_a$ and $HELM_{av}$ frameworks use the same sigmoidal activation function and

regularization parameters used in [17]. The numbers of hidden neurons of $HELM_a$ were ([1000 1000 4000]). For $HELM_{av}$, we employed a late integration strategy where the audio and visual modalities were handled separately in the unsupervised stage to transform the low-level features to representative features. The representations learned for both modalities were subsequently combined in the supervised stage to learn the multimodal transformation. The same HELM-based AE architectures (two layers, each layer consisting of 1000 hidden neurons) were used to process audio and visual data separately in the unsupervised stage, and 4000 hidden neurons were used in the integration module. Table I presents the average PESQ scores attained using these methods under matched and mismatched testing conditions. From the table, it can be observed that both HELM frameworks outperformed the three conventional SE methods with a reasonable margin. In the meanwhile, $HELM_{av}$ achieved superior average PESQ scores compared with $HELM_a$. Moreover, although $HELM_a$ and $HELM_{av}$ achieved significantly better speech quality demonstrated by higher PESQ scores for almost all matched and mismatched noise types, the three conventional SE frameworks performed relatively well for the mismatched stationary *Pink* noise, especially logMMSE. The results in Table I demonstrate that $HELM_{av}$ yielded superior performance in terms of the PESQ score among all the frameworks, thereby confirming the effectiveness of the multimodal structure under matched and mismatched testing conditions.

Subsequently, we compare the intelligibility and SNR improvements of the different SE frameworks. Fig. 2 shows the performance comparison of the different frameworks using HASPI and SSNRI evaluation metrics for different noise types. It can be observed that $HELM_a$ and $HELM_{av}$ achieved better generalization performance as demonstrated by higher average scores of HASPI and SSNRI for different noise types except for mismatched *Grocery store* and *Pink noise*, where $HELM_a$ performed slightly worse than logMMSE, KLT, and RPCA. Among all the frameworks, $HELM_{av}$ performed the best with providing better speech intelligibility and higher SNRs (i.e., higher HASPI and SSNRI scores).

2) HELM with Binary and Ternary Weights: Subsequently, we analyze the performance of the two HELM frameworks by adjusting the real-valued weights to either binary or ternary weights. The frameworks were trained by limiting

TABLE I: AVERAGE PESQ SCORES OF KLT, LOGMMSE, RPCA, $HELM_A$, AND $HELM_{AV}$ PROCESSED SPEECH SIGNALS UNDER MATCHED AND MISMATCHED NOISE CONDITIONS.

Framework	Noise Type						Avg.
	Babble	Crowd party	Applause	Baby cry	Grocery store	Pink noise	
KLT	1.8939	1.8521	1.7567	1.7666	1.8041	2.3862	1.9099
logMMSE	2.2138	2.1473	1.8963	1.9725	2.0986	**2.5774**	2.1510
RPCA	2.2588	2.2363	1.9979	2.0285	2.1831	2.4053	2.1850
$HELM_a$	2.2644	2.2850	2.2979	2.4413	2.2602	2.3572	2.3137
$HELM_{av}$	**2.3269**	**2.3116**	**2.4302**	**2.5923**	**2.3995**	2.4265	**2.4145**

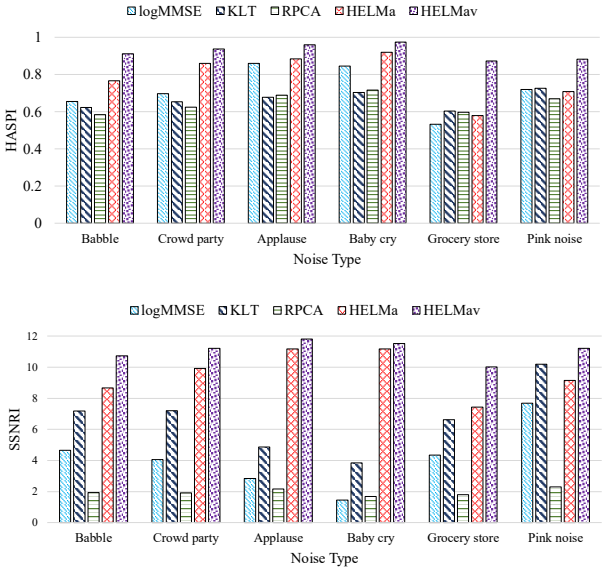

Fig. 2: Performance comparison of different frameworks using HASPI and SSNRI evaluation metrics for six noise types averaged across different SNRs.

the weights to either binary values ({+1, -1}) or ternary values ({-1, 0, +1}). In this study, we only exploit the deterministic binarization with *hard sigmoid* activation to compute the output weight of the supervised layer of the two frameworks. Table II lists the average PESQ scores of the two frameworks trained using binary and ternary weights. It can be observed that both $HELM_a$ and $HELM_{av}$ trained using binary and ternary weights achieved a slightly lower performance compared with the frameworks trained using real-valued weights (average PESQ scores for $HELM_a$ = 2.3137 and for $HELM_{av}$ = 2.4145, as shown in Table I). When using binary weights, the average PESQ scores for $HELM_a$ = 2.1852 and for $HELM_{av}$ = 2.3646; when using ternary weights, the average PESQ scores for $HELM_a$ = 2.1839 and for $HELM_{av}$ = 2.3691, as shown in Table II. It is also noted that the single-modality framework $HELM_a$ performed worse than the multimodal framework $HELM_{av}$ when binarizing or ternarizing the model parameters, which indirectly confirms the effectiveness of incorporation the visual information. Moreover, there is no significant difference in the performances of both frameworks trained using binary and ternary weights. Therefore, we only use binary weights in the following experiments.

TABLE II: AVERAGE PESQ SCORES OF $HELM_A$ AND $HELM_{AV}$ WITH BINARY AND TERNARY WEIGHTS UNDER MATCHED AND MISMATCHED NOISE CONDITIONS.

Weights	Framework	Noise Type						Avg.
		Babble	Crowd party	Applause	Baby cry	Grocery store	Pink noise	
Binary	**$HELM_a$**	2.1201	2.1208	2.1737	2.3930	2.1430	2.1603	2.1852
	$HELM_{av}$	2.2790	2.2895	2.3941	2.5284	2.3468	2.3499	**2.3646**
Ternary	**$HELM_a$**	2.1157	2.1220	2.1761	2.3901	2.1447	2.1550	2.1839
	$HELM_{av}$	2.2850	2.2972	2.3991	2.5276	2.3520	2.3539	**2.3691**

3) Quantized Input Data and HELM with Binary Weights: In this section, we further investigate the effectiveness of the compressed HELM frameworks using quantized input data. Here, we only quantized the mantissa (i.e., precision) bits. More specifically, we converted the input data into the IEEE 754 binary format and quantize the mantissa bits of the input data with fewer bits. There are 23 mantissa bits in the IEEE 754 single-precision format. In the experiments, we quantized the b least significant bits of the mantissa part such that 23 - b bits remained in the mantissa part. The last b bits of the mantissa part were removed, and the remaining bits were subsequently combined with the exponent bits and the sign bit to convert back to the floating point.

Table III shows the performance of the two HELM frameworks trained using 16-bit quantized input data with real-valued weights and binary weights under matched and mismatched noise conditions. It can be observed that the proposed HELM_{av} framework with either real-valued or binary weights maintained a satisfactory performance as demonstrated by a marginal reduction in the average PESQ score. However, the performance of HELM_{a} degraded notably when using 16-bit quantized input data. HELM_{av} trained using quantized input data and binary weights attained a slightly lower average PESQ score (2.3088, as shown in Table III) compared with HELM_{av} trained using original data with binary weights (average PESQ = 2.3646, as shown in Table II). There was only a small relative reduction of 2.37% in the average PESQ score when the precision of the data was reduced to 50% (from 32 bits to 16 bits). However, the audio-only framework (HELM_{a}) was unable to maintain a stable performance. When HELM_{a} was trained using binary weights, its average PESQ score was reduced from 2.1852 (using 32-bit data, as shown in Table II) to 2.0422 (using 16-bit quantized data, as shown in Table III). The relative PESQ reduction is approximately 6.54%, which is higher than that of HELM_{av} (2.37%) (from 2.3646 to 2.3088).

Fig. 3 presents the average HASPI and SSNRI scores of the different HELM frameworks across six noise types at different SNR levels. We compared the performances of the different HELM frameworks: HELM_{a} trained using real-valued weights (termed HELM_{a}(R)), HELM_{av} trained using real-valued weights (termed HELM_{av}(R)), HELM_{a} trained using binary weights (termed HELM_{a}(B)), HELM_{av} trained using binary weights (termed HELM_{av}(B)), HELM_{a} trained using quantized input data with binary weights (termed HELM_{a}(Q+B)), and HELM_{av} trained using quantized input data with binary weights (termed HELM_{av}(Q+B)). Notably, multimodal HELM frameworks with real-valued weights, binary weights, and even quantized input data with binary weights maintained stable HASPI and SSNRI scores even at low SNR levels. The results again confirm that the visual modality used in the HELM_{av} frameworks plays a crucial role in reconstructing a signal even when using quantized input data and binary weights at low SNR levels. The proposed framework maintained a stable performance with reduction in the computational requirement of HELM, enabling its use in the hardware implementation for multimodal environments to obtain an efficient regression ability.

TABLE III: AVERAGE PESQ SCORES OF HELM_{A} AND HELM_{AV} USING 16-BIT QUANTIZED INPUT WITH REAL-VALUED AND BINARY WEIGHTS UNDER MATCHED AND MISMATCHED NOISE CONDITIONS.

Weights	Framework	Noise Type						Avg.
		Babble	Crowd party	Applause	Baby cry	Grocery store	Pink noise	
Real-valued	**HELM_{a}**	2.1260	2.0947	2.1589	2.2662	2.0680	2.1397	2.1422
	HELM_{av}	2.2918	2.2354	2.3792	2.6043	2.3205	2.3158	**2.3578**
Binary	**HELM_{a}**	1.9898	1.9592	2.0859	2.3555	1.9280	1.9349	2.0422
	HELM_{av}	2.2341	2.2089	2.3425	2.5733	2.2559	2.2383	**2.3088**

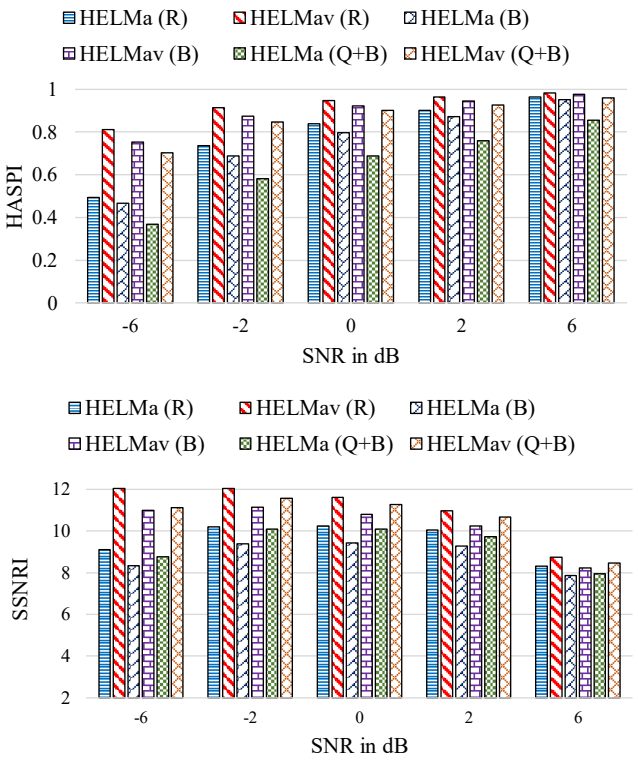

Fig. 3: Performance comparison of HELM_{a} and HELM_{av} using HASPI and SSNRI evaluation metrics for different SNRs averaged across six noise types.

IV. CONCLUSION

In this paper, we proposed a novel HELM_{av} framework for SE to improve the performance of our previous HELM_{a} framework. The main contribution of this study is threefold. First, we confirm that incorporating visual information with audio can enhance the SE performance under various noise conditions across different SNR levels when limited training data are available. Second, a compressed framework was employed for HELM to replace the real-valued weights with binary or ternary weights to reduce the model size. Third, the input data quantization was adopted to reduce

the computational requirement. Our experimental results demonstrate that visual information helps the framework retain most of the information lost owing to the model binarization and input data quantization. The proposed multimodal framework with the binarization and quantization processes can be very useful in real-time situations, where the data arrive in a sequential stream and under dynamically changing and non-stationary environments.

V. Acknowledgment

This work was partly supported by MOST Taiwan Grants 108-2634-F-008-004-, 107-2221-E-001-012-MY2 and 106-2221-E-001-017-MY2.

References

[1] D. L. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectrum*, vol. March Issue, pp. 32–37 (Cover Story), 2017.

[2] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional neural networks to enhance coded speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, 2019.

[3] S. Doclo, M. Moonen, T. Van den Bogaert, and J. Wouters, "Reduced-bandwidth and distributed mwf-based noise reduction algorithms for binaural hearing aids," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 38–51, 2009.

[4] Z.-Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[5] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Academic Press, 2015.

[6] 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.

[7] D. Michelsanti and Z.-H. Tan, "Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification," in *Proc. Interspeech*, 2017, pp. 2008–2012.

[8] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[9] 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.

[10] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 2. IEEE, 1996, pp. 629–632.

[11] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4029–4032.

[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 Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, 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, pp. 436–440.

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

[15] 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*, 2017, pp. 1–6.

[16] 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.

[17] 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.

[18] T. Hussain, Y. Tsao, S. M. Sinicalchi, 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.

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

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

[21] 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 ASC*, 2015, pp. 575–582.

[22] 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.

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

[24] 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.

[25] W. Li, S. Wang, M. Lei, S. M. Siniscalchi, and C.-H. Lee, "Improving audio-visual speech recognition performance with cross-modal student-teacher training," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6560–6564.

[26] K. Hwang and W. Sung, "Fixed-point feedforward deep neural network design using weights+ 1, 0, and- 1," in *Proc. SiPS*, 2014, pp. 1–6.

[27] R. Prabhavalkar, O. Alsharif, A. Bruguier, and L. McGraw, "On the compression of recurrent neural networks with an application to LVCSR acoustic modeling for embedded speech recognition," in *Proc. ICASSP*, 2016, pp. 5970–5974.

[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, pp. 566–573.

[29] 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.

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

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

[32] 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.

[33] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *Proc. NIPS*, 2015, pp. 3123–3131.

[34] C. Zhu, S. Han, H. Mao, and W. J. Dally, "Trained ternary quantization," *arXiv preprint arXiv:1612.01064*, 2016.

[35] N. USED, "IEEE standard for binary floating-point arithmetic, 1985, ansi," *IEEE Standard*, pp. 754–1985.

[36] 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.

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

[38] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[39] C. Sun, Q. Zhang, J. Wang, and J. Xie, "Noise reduction based on robust principal component analysis," *Journal of Computational Information Systems*, vol. 10, no. 10, pp. 4403–4410, 2014.