Bone-conducted Speech Enhancement using Hierarchical Extreme Learning Machine

Tassadaq Hussain, Yu Tsao, Sababto Marco Sinicalchi, Jia-Ching Wang, Hsin-Min Wang, and Wen-Hung Liao

Abstract Deep learning-based approaches have demonstrated promising performance for speech enhancement (SE) tasks. However, these approaches generally require large quantities of training data and computational resources for model training. An alternate hierarchical extreme learning machine (HELM) model has been previously reported to perform SE and has demonstrated satisfactory results with a limited amount of training data. In this study, we investigate application of the HELM model to improve the quality and intelligibility of bone-conducted speech. Our experimental results show that the proposed HELM-based bone-conducted SE framework can effectively enhance the original bone-conducted speech and outperform a deep denoising autoencoder-based bone-conducted SE system in terms of speech quality and intelligibility with improved recognition accuracy when a limited quantity of training data is available

1 Introduction

Speech enhancement (SE) refers to a technique to modify an input speech signal to a target signal with improved speech quality and intelligibility. The input speech signal is usually distorted by additive or convolutive noises or recording device constraints. In this work, we propose a hierarchical extreme learning machine learning (HELM) model to convert a speech utterance recorded with a bone-conducted microphone (BCM) to one from an air-conducted microphone (ACM). Unlike an ACM that records speech signals directly, a BCM captures speech signals based on the vibrations of the speaker's skull. The speech signals recorded with a BCM are robust against noise while some high frequency components may be missing compared to the speech signals recorded with an ACM.

Tassadaq Hussain
TIGP-SNHCC
Department of Computer Science,
National Chengchi University, Taiwan
email: tass.hussain@iis.sinica.edu.tw

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

Sabato Marco Siniscalchi
Department of Computer Science
Kore University of Enna, Italy

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

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

Wen-Hung Liao
Department of Computer Science
National Chengchi University, Taiwan
e-mail: whliao@cs.nccu.edu.tw

A number of filtering-based and probabilistic solutions have been proposed in the past to convert BCM utterances to ACM utterances. In [12], the BCM utterances were passed through a designed reconstruction filter to improve quality. In [19] and [20], BCM and ACM utterances were combined for SE and automatic speech recognition (ASR) in non-stationary noisy environments. In [4], a probabilistic optimum filter (POF)-based algorithm was used to estimate the clean features from the combination of standard and throat microphone signals. Thang et al. [16] restored bone conducted speech in noisy environments based on a modulation transfer function (MTF) and a linear prediction (LP) model. Later, Tajiri et al. [14] proposed a noise suppression technique based on non-negative tensor factorization using a body-conducted microphone known as a nonaudible murmur (NAM) microphone.

Recently, neural network based approaches have shown tremendous performance for SE. In these approaches, non-linear mapping functions are estimated to transform the source speech utterances to target speech utterances using a learning-based model. The models are generally trained with source–target utterance pairs. Numerous neural network frameworks have been used and have demonstrated exceptional performance for speech processing. For example, in [9], the authors applied a deep denoising autoencoder (DDAE) framework by stacking multiple denoising autoencoders and demonstrated state-of-the-art performance for SE. Meanwhile, in [18], a deep neural network (DNN) was used to handle a wide range of additive noises for the SE task. In addition, a signal-to-noise ratio (SNR)-aware convolutional neural network was proposed by Fu et al. [2] for SE. Similarly, in [3], a fully convolutional network (FCN)-based architecture is used to optimize the intelligibility of speech, along with the model parameters, during SE. In [1], a bidirectional long short-term memory (BLSTM)-based framework was utilized for SE and ASR. Furthermore, generative adversarial networks have also been deployed for SE [10] [17]. More recently, the DDAE framework has been applied to BCM–ACM SE and has been noted to provide satisfactory generalization and speech recognition performance [8]. Despite notable improvements achieved by the deep learning models over the conventional approaches, deeper structure-based methods typically require large computational resources and adequate quantities of training data for effective learning.

In addition to the deep learning models that require large quantities of data to train the models and learn the mapping functions, an HELM model has been proposed to perform SE and has been confirmed to achieve very good performance with limited quantities of training data [7]. In this study, we further investigate the HELM model for enhancing BCM speech. The experimental results verify that the proposed SE framework notably improves the original BCM speech and outperforms the previous DDAE-based SE framework in terms of two standardized objective measures, namely perceptual evaluation of speech quality (PESQ) [11] and short-time objective intelligibility (STOI) [13], as well as ASR performance.

2 Proposed HELM-based BCM Speech Enhancement System

The extreme learning machine (ELM) was proposed by Huang et al. [6] for shallow neural feed-forward architectures to address the slow gradient issue. In addition to ELM, the hierarchical structure of ELM, termed as HELM, was proposed by Tang et al. [15] to extract feature representation in a hierarchical manner. HELM consists of both unsupervised and supervised (semi-supervised) stages, in which the compact feature representations are extracted using the unsupervised ELM-based sparse autoencoders followed by the supervised regression/classification stage. Fig. 1 shows the offline and online phases of the proposed HELM-based SE system. During the offline phase, the speech utterances recorded using both BCM and ACM are divided into short segments of frames and subsequently converted into the frequency domain using short-time Fourier transform (STFT). The frame-based 160 dimensional Mel spectral features of the BCM–ACM utterance pairs are estimated and analyzed by the HELM to learn the mapping function to convert the BCM Mel spectral features to the corresponding ACM Mel spectral features

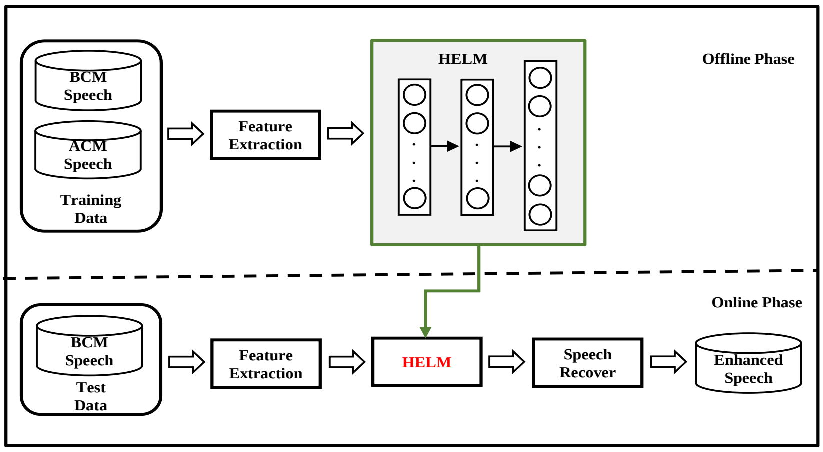

Fig.1 HELM-based SE Architecture

In the testing phase, the incoming BCM test utterance is initially segmented into short frames using STFT and subsequently converted into 160 dimensional Mel spectral features. The BCM Mel spectral features are further converted by the trained HELM to acquire the enhanced Mel spectral features. The inverse STFT is then applied to the enhanced Mel spectral features together with the phase of the original BCM utterance to obtain the restored speech. More details for the online and offline phases of the HELM framework can be found in [7].

3 Experiments and Results

3.1 Experimental Setup and Dataset Description

The dataset used in our experiments is the same as the one used in [8]. The utterances were spoken by a native speaker and recorded using ACM and BCM microphones simultaneously with the transcript of the Taiwan Mandarin hearing in noise test (TMHINT) [8]. The dataset was originally recorded using a sampling rate of 44.1 kHz and was further resampled to 16 kHz for processing. In this study, we selected 270 utterances from the complete dataset, among which 200 utterances were randomly selected and used as the training data, and the remaining 70 utterances were selected as the testing data. There was no overlap between the training and testing utterances.

We evaluated the proposed approach using two standardized objective measures: PESQ and STOI. The PESQ measures the speech quality of the estimated speech signal with the clean speech signal as a reference. The PESQ score ranges between -0.5 and 4.5, where a higher score denotes better speech quality. The STOI measures the intelligibility of the estimated speech with the clean speech signal as a reference. The STOI score ranges between 0.0 and 1.0, and a higher score indicates better speech intelligibility. In this study, we used 160 dimensional Mel spectral features along with 80000 patches for the BCM and ACM training samples. In the experiments, we did not use the context information of the input speech vectors, i.e., no neighboring input speech vectors were considered.

3.2 HELM-based SE System

We first utilize the performance of the proposed HELM system for converting the BCM speech to ACM speech. Table 1 demonstrates the average PESQ and STOI scores of the unprocessed BCM and HELM enhanced speech. The utterances recorded with the BCM only contain the lower frequency components of the utterances. The HELM framework was trained using the BCM/ACM training utterance pairs with small numbers of hidden neurons (200 200, and 500), where 200, 200, and 500 are the number of hidden neurons for the first, second, and third layer of HELM, with a sigmoid activation function. The regularization parameters in HELM were the same as those used in [7].

From Table 1, it is clear that HELM improves the speech quality and intelligibility by obtaining a remarkable improvement in the PESQ and STOI scores compared to the unprocessed BCM utterances. Next, we applied a speech intelligibility index (SII)-based post-filter [8] on the utterances recorded with the ACM to train our HELM framework. That is, HELM was trained using the BCM/ ACM(IE) utterance pairs, where ACM(IE) represents the ACM utterances processed by the SII post-filter. The purpose of SII-based post-filtering is to consider the critical band importance function, which correlates speech intelligibility for humans. Table 2 shows the performance comparison between the

Table 1 Average PESQ and STOI scores of the unprocessed BCM speech and the HELM enhanced speech trained with the BCM/ACM utterance pairs.

Method	**PESQ**	**STOI**
Unprocessed	2.5775	0.7254
HELM(BCM/ACM)	2.7903	0. 8329

DDAE- and HELM-based enhanced utterances. For fair comparison, we employed a three-layer DDAE structure, where each layer had 300 hidden neurons (300 × 3 = 900 hidden neurons), with a sigmoid activation function. The table shows that HELM outperformed the unprocessed BCM utterances and DDAE-based enhanced utterances in terms of PESQ score with a reasonable margin when the systems are trained using BCM-ACM(IE) utterances. However, DDAE provided a higher intelligibility (STOI) score as compared with the proposed HELM framework. In addition, comparing the results in Tables 1 and 2, we can see that the PESQ performance of HELM can be improved by using the BCM/ACM(IE) training utterances.

Table 2. Average PESQ and STOI scores of the unprocessed BCM speech and DDAE- and HELM- enhanced speech trained with the BCM/ACM(IE) utterance pairs.

Method	**PESQ**	**STOI**
DDAE(BCM/ACM-IE)	2.7145	0.8470
HELM(BCM/ACM-IE)	2.8454	0.8306

3.3 Spectrogram Analysis

In this section, we analyze the spectrogram of the converted speech signal yielded by HELM and DDAE. Fig. 2(a) and (b) show the spectrograms of the BCM and corresponding ACM utterances. Fig. 2(c) and (d) display the spectrograms of the DDAE- and HELM- enhanced speech. From Fig. 2(c) and (d), we observe that both of the frameworks (DDAE and HELM) have successfully converted the BCM utterance to the enhanced utterance closer to the original ACM (Fig. 2(a)) utterance. Meanwhile, HELM achieves better speech quality with PESQ = 2.9633 as compared to DDAE whose PESQ = 2.6027, where the original BCM utterance has PESQ = 2.2619.

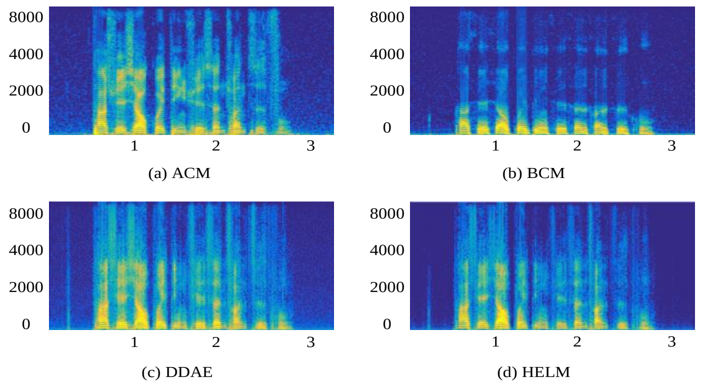

Fig.2 Spectrograms of the enhanced test utterances using the (c) DDAE and (d) HELM of the (a) ACM and (b) BCM utterances. For each figure, the x-axis denotes the time in seconds, and the y-axis represents the frequency in Hertz.

3.4 Automatic Speech Recognition

In addition to the speech quality and sound intelligibility measures, we also compared the recognition results of the speech processed by the HELM and DDAE frameworks. As discussed in the previous section, the original BCM utterances only contain the low frequency components of the speech signals, which could result in a poor recognition performance. We computed the ASR performance of the DDAE- and HELM- enhanced speech and the original BCM test utterances using Google ASR [5]. The testing results of the ACM utterances were also listed as the upper bound. Table 3 presents the character error rate (CER) evaluated on the 70 test utterances. From Table 3, we first note that BCM speech achieved higher CERs compared to ACM speech. Next, we can see that both frameworks (DDAE and HELM) provide a lower CER compared to the unprocessed BCM speech. Moreover, HELM clearly outperforms DDAE in the ASR experiments.

Table 3 CERs of the original ACM and BCM test utterances and DDAE- and HELM- enhanced speech.

	ACM	**BCM**	**DDAE**	**HELM**
CER	1.0%	12.13%	10.98%	8.27%

3.5 Sensitivity/Stability Towards the Training Data

Next, we intend to analyze the sensitivity of the proposed HELM framework with the quantity of training data. In our previous sections, we used 80000 Mel spectral patches (MSP) of the samples, randomly selected from the training set, to train the enhancement models. In this set of experiments, we investigated the performance using different sizes for the training data from 200 to 500, 1000, 5000, 10000, 20000, 40000, and 80000 MSPs.

Figs. 3 and 4 show the brief summaries of the two frameworks in terms of average PESQ and STOI, respectively. From Fig. 3, we can note that the PESQ scores of the DDAE framework has improved from 1.4004 to 1.4732 when the size of the training data increases from 200 to 500 MSP. The same trend of improvement can be seen when the size of the training data is increased from 500 to 80000 MSP for the DDAE framework. On the other hand, HELM presents less sensitivity towards the reduction in the size of the training data. The PESQ score for HELM configuration escalated from 2.1867 to 2.4605 when the size of the training data only increased from 200 to 500 MSP, as shown in Fig. 3. Similarly, the PESQ score further escalated to 2.8454 (with 80000 MSP) from 2.6018 (with 1000 MSP). It is very interesting to observe that DDAE framework require more than 10000 MSP to increase the PESQ score beyond 2.0 (PESQ = 1.9517 with 10000 MSP), whereas HELM can achieve PESQ = 2.1867 even with 200 MSP.

Next from Fig. 4, the STOI score for the DDAE framework was low when the patches were 200 MSP. In contrast, HELM produced comparable STOI scores to the unprocessed BCM speech when the training data size was 200 MSP. The STOI for DDAE dropped from 0.8470 to 0.5066 while the drop was not that severe for HELM as it only declined from 0.8306 to 0.7116 when the amount of MSP was reduced from 80000 to 200. The results from Figs 3 and 4 show that both DDAE and HELM can yield improved speech quality and intelligibility when sufficient training data are available (more than 80000MSP), while HELM gave improved performance even with 2000 MSP.

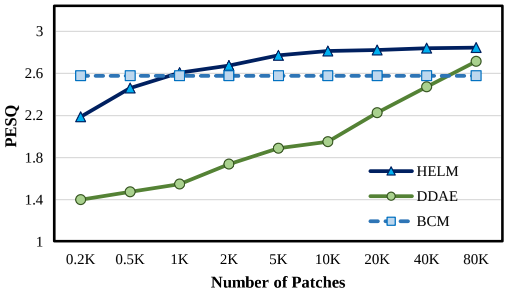

Fig. 3 Average PESQ scores for DDAE and HELM SE frameworks using different amounts of training data.

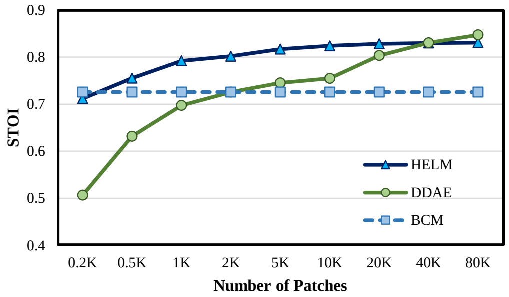

Fig. 4 Average STOI scores for DDAE and HELM SE frameworks using different amounts of training data.

Moreover, we evaluated the ASR results of the DDAE and HELM frameworks for different amounts of training data. Fig. 5 shows the impact of the number of MSPs on the CER results of the DDAE and HELM frameworks. From the figure, we can observe a similar trend for the performance of CER as that observed for PESQ and STOI. Overall, the CERs of the two frameworks are reduced when the MSP increases from 1000 to 80000. Moreover, HELM yields a consistent and stable recognition performance with small CER as compared with DDAE for different numbers of MSP. The CERs of the DDAE decreased consistently and later reduced dramatically when the MSP increased beyond 10000 MSP, indicating the impact of the training data on the performance of the DDAE framework. As compared to DDAE, HELM can yield reduced CER as compared to the BCM even with 5000 MSP. The results again confirm the advantage of HELM over the DDAE when the amount of training data is limited.

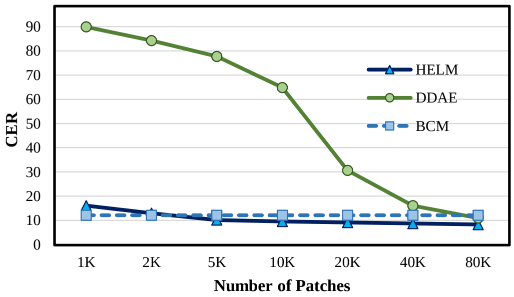

Fig. 5 CER results of DDAE and HELM frameworks using different amounts of training data.

4 Conclusion

In this study, an HELM-based BCM SE approach was proposed. We evaluated the proposed approach using two standardized objective measures, i.e., PESQ and STOI. For comparison, a DDAE-based SE system was established and used to test performance. Meanwhile, the CERs were tested to compare the performance of the DDAE and HELM enhanced speech. The experimental results have confirmed the effectiveness of the proposed HELM-based SE framework by maintaining high speech quality and intelligibility with high recognition accuracy. Since deep learning based approaches generally require large quantities of training data to learn complex non-linear relationships between the input and output, we examined the sensitivities of the DDAE and HELM frameworks with different quantities of training samples. The performance of the HELM-based SE framework proved to be consistent and provided superior performance compared with the DDAE-based SE framework. This is the first attempt that has successfully applied HELM framework for bone-conducted SE. Because of no fine-tuning or adjustment of parameters during the training phase, HELM-based models are highly suitable for applications in embedded and mobile devices.

References

[1] Chen Z, Watanabe S, Erdogan H, Hershey JR (2015) Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks. In: Proc. ISCA, pp. 3274-3278

[2] Fu SW, Tsao Y, Lu X (2016) SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement. In: Proc. INTERSPEECH, pp. 3768-3772

[3] Fu SW, Wang TW, Tsao Y, Lu X, Kawai H (2018) 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, 26:1570-1584. doi:10.1109/TASLP.2018.2821903

[4] Graciarena M, Franco H, Sonmez K, Bratt H (2003) Combining standard and throat microphones for robust speech recognition. IEEE Signal Processing Letters 10:72-74. doi:10.1109/LSP.2003.808549

[5] Google (2017), Cloud Speech API, https://cloud.google.com/speech/

[6] Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing, 70:489-501.

[7] Hussain T, Siniscalchi SM, Lee CC, Wang SS, Tsao Y, Liao WH (2017) Experimental study on extreme learning machine applications for speech enhancement. IEEE Access, 5:25542-25554. doi:10.1109/ACCESS.2017.2766675

[8] Liu HP, Tsao Y, Fuh CS (2018) Bone-conducted speech enhancement using deep denoising autoencoder. Speech Communication, 104:106-112

[9] Lu X, Tsao Y, Matsuda S, Hori C (2013) Speech enhancement based on deep denoising autoencoder. In: Proc. INTERSPEECH, pp. 436-440

[10] Pascual S, Bonafonte A, Serra J (2017) SEGAN: Speech enhancement generative adversarial network. arXiv preprint arXiv:1703.09452.

[11] Rix AW, Beerends JG, Hollier MP, Hekstra AP (2001) Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: Proc. ICASSP, pp. 749-752

[12] Shinamura T, Tomikura T (2005) Quality improvement of bone-conducted speech. In: Proc. ECCTD, pp. 1-4

[13] Taal C H, Hendriks R C, Heusdens R, Jensen J (2010) A short-time objective intelligibility measure for time-frequency weighted noisy speech. In: Proc. ICASSP, pp. 4214-4217

[14] Tajiri Y, Kameoka H, Toda T (2017) A noise suppression method for body-conducted soft speech based on non-negative tensor factorization of air-and body-conducted signals. In: Proc. ICASSP, pp. 4960-4964

[15] Tang J, Deng C, Huang GB (2016) Extreme learning machine for multilayer perceptron. IEEE Transactions on Neural Networks and Learning Systems, 27:809-821. doi:10.1109/TNNLS.2015.2424995

[16] Thang T V, Kimura K, Unoki M, Akagi M (2006) A study on restoration of bone-conducted speech with MTF-based and LP-based models. Journal of Signal Processing. 10:407-417

[17] Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kavukcuoglu K (2016) WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499

[18] Xu Y, Du J, Dai LR, Lee CH (2015). A regression approach to speech enhancement based on deep neural networks. IEEE/ACM Transactions on Audio, Speech and Language Processing, 23:7-19. doi:10.1109/TASLP.2014.2364452

[19] Zhang Z, Liu Z, Sinclair M, Acero A, Deng L, Droppo J, Zheng Y (2004) Multi-sensory microphones for robust speech detection, enhancement, and recognition. In: Proc. ICASSP, pp. 781-784

[20] Zheng Y, Liu Z, Zhang Z, Sinclair M, Droppo J, Deng L, Huang X (2003) Air- and bone-conductive integrated microphones for robust speech detection and enhancement. In: Proc. ASRU, pp. 249-254