Improving Speech Enhancement Performance by Leveraging Contextual Broad Phonetic Class Information

Yen-Ju Lu, Chia-Yu Chang, Cheng Yu, Ching-Feng Liu, Jeih-weih Hung, *Member, IEEE*, Shinji Watanabe, *Fellow, IEEE*, and Yu Tsao, *Senior Member, IEEE*

***Abstract*—Previous studies have confirmed that by augmenting acoustic features with the place/manner of articulatory features, the speech enhancement (SE) process can be guided to consider the broad phonetic properties of the input speech when performing enhancement to attain performance improvements. In this article, we explore the contextual information of articulatory attributes as additional information to further benefit SE. More specifically, we propose to improve the SE performance by leveraging losses from an end-to-end automatic speech recognition (E2E-ASR) model that predicts the sequence of broad phonetic classes (BPCs). We also developed multi-objective training with ASR and perceptual losses to train the SE system based on a BPC-based E2E-ASR. Experimental results from speech denoising, speech dereverberation, and impaired speech enhancement tasks confirmed that contextual BPC information improves SE performance. Moreover, the SE model trained with the BPC-based E2E-ASR outperforms that with the phoneme-based E2E-ASR. The results suggest that objectives with misclassification of phonemes by the ASR system may lead to imperfect feedback, and BPC could be a potentially better choice. Finally, it is noted that combining the most-confusable phonetic targets into the same BPC when calculating the additional objective can effectively improve the SE performance.**

***Index Terms*—Speech enhancement, broad phonetic classes, articulatory attribute, robust automatic speech recognition, end-to-end.**

I. Introduction

SPEECH enhancement (SE) systems improve the intelligibility and quality of speech signals by searching for mapping between distorted speech signals and their clean counterparts. SE has been widely adopted as a front-end processor in various real-world applications such as assistive listening devices [1], [2], speech coding [3], [4], speaker recognition [5], and automatic speech recognition (ASR) [6], [7], [8], [9], [10]. With recent breakthroughs in deep learning (DL), DL-based SE methods have been extensively investigated and have exhibited outstanding performance [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. A DL-based SE model with nonlinear processing can learn complex transformations for the denoising process. Accordingly, they can considerably outperform conventional SE methods, particularly in extremely low-SNR scenarios and/or non-stationary noise environments. Furthermore, alternative signal processing approaches allow end-to-end DL-wise neural networks to incorporate speech signals with heterogeneous data. For instance, previous studies have confirmed the effectiveness of leveraging face/lip images [24] and symbolic sequences for acoustic signals [25] in SE systems.

To enhance the performance of SE systems, researchers have explored the use of additional losses, including perceptual losses [26], [27], [28], [29] and losses from acoustic models (AM) [30], [31], [32]. The former group of studies typically computes losses between clean speech and the corresponding enhanced speech in subsequent networks. For instance, in [26], an audio classification network is constructed, and the feature loss function is computed based on the difference between the feature activations of the clean reference signal and the denoised output in an intermediate layer of the classification network. In [27], the dynamic perceptual loss is calculated as the difference between the classification results of clean speech and enhanced speech from the discriminative network, which is then used to update the SE model. Moreover, [28], [29] employ pre-trained models that provide phonetic and acoustic information of audio signals to compute the loss between the acoustic features of clean speech and enhanced speech.

Phoneme information in acoustic models (AM) has also been used to improve SE performance. In [30], a variety of perceptual losses were tested using pre-trained AMs for different tasks, including acoustic event detection, automatic speech, speaker, and emotion recognition. Both [31] and [32] add a perceptual loss by passing both clean and denoised spectral features to a pre-trained AM and computing the L2 distance of the respective outputs. With the emergence of end-to-end speech recognition (E2E-ASR), joint training of SE and E2E-ASR has been studied for the development of robust speech recognition systems [33],

Manuscript received 7 July 2022; revised 6 April 2023 and 1 June 2023; accepted 9 June 2023. Date of publication 27 June 2023; date of current version 26 July 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Stefan Goetze. *(Yen-Ju Lu and Chia-Yu Chang contributed equally to this work.) (Corresponding author: Yu Tsao.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Chi Mei Medical Center under Application No. 10311-008, and performed in line with the Declaration of Helsinki.

Yen-Ju Lu, Chia-Yu Chang, Cheng Yu, and Yu Tsao are with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan (e-mail: neilyenjulu@gmail.com; r07522520@ntu.edu.tw; victor41899@gmail.com; yu.tsao@sinica.edu.tw).

Ching-Feng Liu is with the Chi Mei Hospital, Tainan 71004, Taiwan (e-mail: wtcen@stust.edu.tw).

Jeih-weih Hung is with the Department of Electrical Engineering, National Chi Nan University, Puli 454301, Taiwan (e-mail: jwhung@ncnu.edu.tw).

Shinji Watanabe is with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: shinjiw@ieee.org).

Digital Object Identifier 10.1109/TASLP.2023.3288418

[34]. The E2E-ASR loss has also been shown to help improve SE metrics in [35].

These works have demonstrated the benefits of deep features and phonetic features for the SE task. However, the performances are limited by the accuracy of phoneme identification task. This motivated our prior work using broad phonetic class (BPC) posteriorgrams [36]. We also demonstrated that speech signals within the same BPC share the same noisy-to-clean transformation. Moreover, the SE model to be learned may combat the noise effect when the original training set has contextual information (desired redundancy) among speech signals. Based on these observations, the contextual information of BPCs, which was not explicitly used in our earlier work [36], will benefit SE. Therefore, we proposed an SE architecture incorporating contextual articulatory information acquired by an end-to-end BPC–ASR system, which is expected to further boost SE performance.

The proposed architecture has an SE model and a BPC-based end-to-end ASR (E2E-ASR) module. The bidirectional long short-term memory (BLSTM) encoder and connectionist temporal classification (CTC)/attention hybrid decoder served as the E2E-ASR module, with its phoneme targets replaced by BPCs. Compared with characters and mono-phonemes, selecting BPCs as labels reduced the prediction difficulty. This study derives two losses from the pre-trained AM, which are used for multi-objective training to train the SE system: the ASR classification loss and the perceptual loss. To compute the losses, we selected ESPnet as the toolkit for the E2E-ASR model and connected it to the output of the SE model to establish an end-to-end SE–ASR system.

The deep-feature embeddings in the ASR model were extracted for perceptual loss training, and the distance between the deep features of clean and enhanced speech signals was used as the extra loss for learning the SE model. We considered three types of BPCs: the manner of articulation (BPC(M)), place of articulation (BPC(P)), and data-driven BPC (BPC(D)).

The proposed system was first evaluated on a speech denoising task with TIMIT (English) and TMHINT (Mandarin Chinese) datasets with two standard objective evaluation metrics: perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI), subjective listening tests, and ASR performance. Further experiments were conducted on the denoising-and-dereverberation task using the TMHINT dataset. Finally, the system was evaluated on an impaired speech task, and a set of subjective evaluations was conducted to test its performance.

The rest of this article is organized as follows: Section II introduces the criteria used to define the BPC and E2E-ASR systems in ESPnet. Section III describes the proposed end-to-end BPC SE–ASR system. Section IV presents the experimental setup, evaluation results, and analyses. Finally, Section V concludes this article.

II. Backgrounds

This section introduces the BPC and end-to-end speech recognition model, which serve as the primary components of our proposed architecture.

A. Broad Phonetic Classes

BPC categorizes all phonemes into several groups according to the articulation of each phoneme. Acoustic characteristics are similar among phonemes of the same group, such as the manner/place of obstruction of airflow that passes through the mouth. This study adopted three BPC grouping methods: knowledge-based and data-driven.

1) Knowledge-Based BPC: This study used two knowledge-based BPCs: place and manner of articulation. The place of articulation indicates where the air stream is obstructed/modified in the vocal tract when a sound is uttered. The manner of articulation indicates how the air stream is obstructed/modified. Similar characteristics appear in phonemes with the same manner and place of articulation, and the type of articulation manner can be discerned by observing the shape of its waveform [37]. Phonemes with the same manner/place of articulation have similar spectral characteristics and may generate significant confusion when performing ASR [38]. Nevertheless, this problem can be alleviated using contextual articulatory information [39], [40], [41], [42].

Different native languages have different articulatory characteristics, and divergence of the manner/place of articulation enables them to produce distinct sounds and prosody. This study used the International Pronunciation Alphabets (IPAs) to represent the target sentences in any language and characterize each IPA label into BPCs. Eighty-seven phones in IPAs were clustered into nine articulatory manner classes: vowel, plosive, nasal, trill, trap or flap, fricative, lateral fricative, approximant, and lateral approximant. The vowel class includes diphthongs and semi-vowels, as suggested by [38]. Each language uses only a portion of the IPAs to represent all its phonemes. For instance, in the TIMIT dataset, 60 IPAs were used to represent all phonemes in English, and they were clustered into five groups based on articulation manner: vowels, stops, fricatives, nasals, and silence.

For the place of articulation, we used 10 clusters in Mandarin by classifying 87 IPAs: vowels, bilabial, labiodental, dental alveolar posteroalveolar, retroflex, palatal, velar, ular, pharyngeal, and glottal. Comparatively, only nine clusters have been used in English [37]. In both manner and place articulations, vowels constituted a distinct group from the others. This is because these two classification criteria mainly focus on consonants, whereas uttering vowels do not block the air stream as much as pronouncing consonants.

2) Data-Driven BPC: The similarity between phonemes can also be evaluated in a data-driven manner, derived from the recognition result of a pre-trained AM. In a previous study [43], a confusion matrix $\mathbf{M}$ for phonemes was created, where its entry $\mathbf{M}_{ij}$ was defined by the number of events for phoneme i to be misidentified as phoneme j. This matrix was assumed to reflect the similarities between each pair of phonemes. When clustering phonemes through the similarity metric, a merging process was performed until the cluster number reached nine, as recommended in [43].

B. CTC/attention E2E-ASR

The applied recognition model adopted two major E2E-ASR implementations: CTC and attention. It provided a single neural network architecture to perform speech recognition in an E2E manner [44]. The attention-based method used an attention mechanism to align acoustic frames and recognize symbols. CTC uses Markov assumptions to solve sequential problems efficiently using dynamic programming. Multi-task learning based on CTC and attention allows E2E-ASR to resolve the misalignment issues encountered in ordinary attention-based E2E-ASR.

Compared to conventional ASR models that require various modules, such as AMs, language models, and lexicons, the E2E network eliminates the need for linguistic resources. It enables an optimization of front-end processors that precede the ASR component in an end-to-end manner. Furthermore, the complexity of the E2E-ASR building process is notably reduced, as it does not require GMM/HMM construction, DNN pre-training, lattice generation, and complex searches during decoding, compared to the conventional ASR. Simplifying a unified deep learning framework in E2E-ASR enables researchers to develop or use an ASR system for new applications without extra efforts, such as preparing the linguistic resources of new languages.

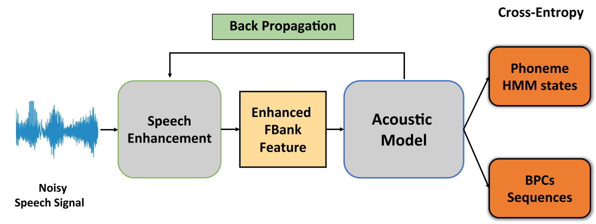
In the E2E-ASR model, a shared BLSTM encoder transforms the input sequence into high-level features and undergoes multi-objective learning. When training the E2E-ASR model, the objective functions for the attention and CTC frameworks were applied to improve the robustness and achieve fast convergence. The CTC objective function is an auxiliary task to train the encoder of the attention model. Compared with the single attention model, combining the forward-backward algorithm in CTC accelerates the process of finding the desired alignment in a monotonic manner and mitigates the prediction from a letter-wise attention objective to a sequence-level CTC objective. Through joint decoding, attention- and CTC-based scores were combined in a one-pass beam search algorithm to obtain the ASR results and further eliminate irregular alignments.

III. Proposed Model

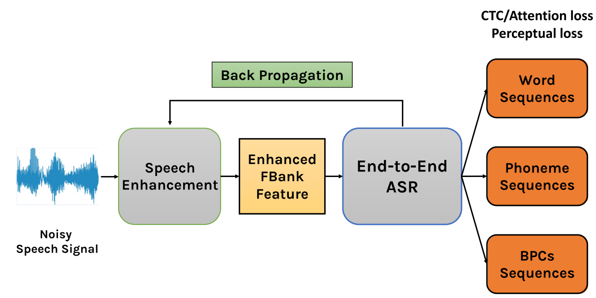
The proposed model connected and trained the CTC/attention E2E-ASR for ASR and a transformer for SE in an E2E manner to improve SE performance with BPC recognition.

A. SE With AM and E2E-ASR Multi-Objective Training

We hypothesize that the SE model can be further promoted through learning to generate enhanced speech signals with a more precise transformation guided by BPC information. To validate this hypothesis, we set up two SE systems that are estimated by multi-objective training with two different losses. Fig. 1(a) connects the SE model with a DNN-based AM from the conventional hybrid DNN-HMM ASR system, and both BPCs and phonemes can be HMM states in the AM. Here, the SE model was independently updated in each time frame because the AM only predicts one HMM state at one forward step without considering the long-term context. By contrast, the newly proposed architecture shown in Fig. 1(b) connects the SE model with an E2E-ASR model, predicting consecutive BPC labels instead of phoneme/word sequences. Since the E2E model predicts all the



(a) The training approach with DNN-HMM hybrid AM model.



(b) The training approaches with E2E-ASR model.

Fig. 1. Training approach with DNN-HMM hybrid AM model and E2E-ASR model. Training targets of DNN-HMM hybrid AM are phonemes or BPCs, and the training targets of E2E-ASR model are words, phonemes, or BPCs.

BPC labels at once, the SE model can learn information for a longer period and generate results with better transformation of articulatory information. We also conducted SE-E2E-ASR multi-objective training for the sequences of phonemes and words instead of BPCs for comparison, as shown in Fig. 1(b).

The training of conventional DNN-HMM hybrid acoustic models typically has two steps. First, the GMM-HMM acoustic models for different tri-phones were trained using the expectation-maximization (EM) algorithm, which infers the state emission probabilities in the HMM. We then used DNN instead of GMM and trained it to model the HMM states more precisely. We conducted our experiments based on the TIMIT recipe in Kaldi [45] and built a four-layered DNN using PyTorch-Kaldi [46] with dimensions $[1024, 1024, 1024, \text{states}]$, where "states" denotes the number of states in the HMM. We connected the DNN-based AM after the SE model and used cross-entropy loss from the predicted results of the HMM states to update the SE model.

B. SE-E2E-ASR Model Architecture

Fig. 2 shows the proposed SE-E2E-ASR architecture, which has a feature extractor, SE model, and BPC-based pre-trained ASR model, which uses the overall loss of both models for back-propagation in training. The SE model is arranged as a transformer, which is an attention-based deep neural network originally proposed for machine translation [47] and later applied in numerous other natural language processing tasks. The transformer has shown considerable improvements over common recurrent neural networks (RNNs) and has been further

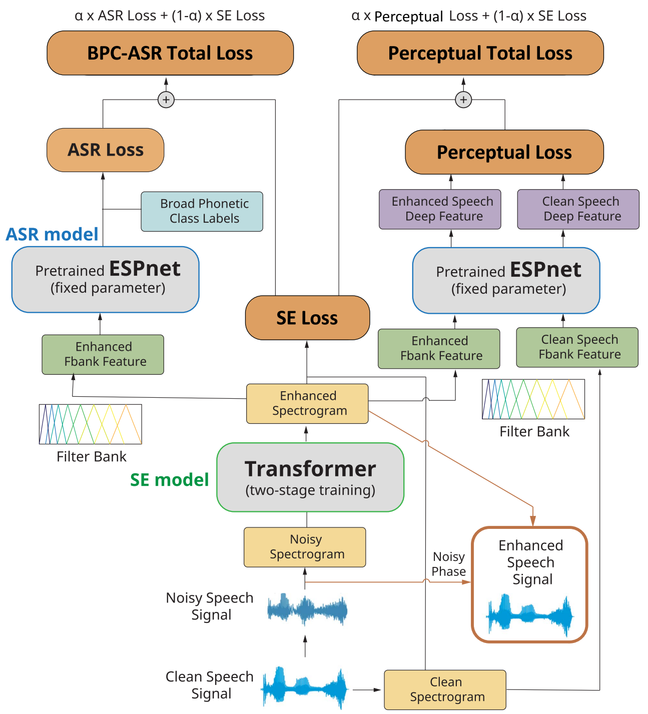


Fig. 2. Architecture of the proposed combination of SE model and E2E BPC-ASR Network.

exploited in SE tasks. The E2E-ASR model in ESPnet was initially set to recognize the acoustic waveform into character-level sequences, while its output labels were modified into the desired BPC labels in our ASR model.

1) Transformer-Based SE Model: Transformers have been investigated extensively in SE studies [48], [49], [50]. Following a sequence-to-sequence learning structure, the transformer comprises encoder and decoder networks. This method preserved only the encoder part of the transformer for the SE process because the input distorted signal and output enhanced signal share the same length. The transformer has four convolutional layers for encoding the spectrum of the input signals and eight attention blocks. Each attention block comprises multi-headed self-attention and two fully connected layers as the feed-forward network. Residual connections and layer normalization were performed in each layer [51].

2) Transformer-Based ISE Network: In addition to speech denoising, we also enhanced impaired speech as an alternative SE task. We adopted a voice conversion (VC) model [52] based on a transformer as the fundamental architecture of the ISE model. We applied a sequence-to-sequence (seqASRseq) model based on transformer architecture with text-to-speech (TTS) pre-training. By transferring knowledge from learned TTS models trained by the large TTS dataset, we could satisfy the need for large-scale corpora to train the transformer. This ISE model was primarily based on the transformer-TTS model, consisting of encoder and decoder stacks. The encoder layer has a multi-head self-attention sub-layer followed by a fully connected feed-forward network, and the decoder layer contains another sub-layer, which performs multi-head attention over the output of the encoder stack. Each layer comprises a residual connection and layer normalization. This ISE model takes the source log-Mel spectrogram as input and outputs a converted log-Mel spectrogram. Before training with the ISE objective, the decoder in the transformer was first pre-trained using the TTS-objective tasks. Subsequently, the decoder was fixed to preserve its ability to robustly capture speech features, such as articulation and prosody, and the encoder was trained with TTS speech input to learn the effective hidden representation. Finally, the ISE model was trained by initializing the model using the pre-trained parameters of the encoder and decoder. The models initialized with the TTS-pre-trained model parameters generated effectively hidden representations for high-fidelity and highly intelligible converted speech.

3) Mel-Filters Processing: To make the whole SE-E2E-ASR system differentiable, we replaced the Kaldi feature extractor in the original ESPnet with a filter-bank extractor, creating speech features of the enhanced waveforms from the SE module. The original ESPnet uses Kaldi feature extraction, for which most recipes use 80-dimensional logarithmic Mel-spectra with the pitch feature (83 dimensions in total). By contrast, the filter-bank extractor applies 26 triangular Mel-scaled filters to the power spectrum of an input signal to extract the filter-bank feature. Compared with the original Kaldi feature extractor, the filter-bank extractor connects the SE module with the ASR model and ensures that the whole SE-E2E-ASR system is differentiable. After feature extraction, we prepared the data for the SE-E2E-ASR system with all the information included in the Kaldi data directory (transcriptions, speaker and language IDs, and input and output lengths) and pre-trained the E2E-ASR model using the clean filter-bank features as input. The ISE task used the Kaldi feature extractor and preserved the processes in the ESPnet toolkit, where 80-dimensional logarithmic Mel-spectra features (with 1024 DFT points and 144-point frameshift) are used.

4) E2E BPC-ASR Network: To obtain BPC context information, we changed the original output word labels of E2E-ASR to the desired BPC labels. Accordingly, the BPC–ASR model predicted the BPC label sequence that corresponds to each utterance. We pre-trained the BPC–ASR model using the filter-bank features of clean utterances as the input for the overall E2E model training. In order to avoid the SE model from producing distorted signals while being used as new features for training the ASR model during joint training, the parameters of the pre-trained BPC-ASR model were kept fixed or "frozen" during the training of the connected SE model.

C. Multi-Objective Training Methods

1) SE-ASR Multi-Objective Training: Optimization of the E2E-ASR objective function was treated as an auxiliary task to train the transformer-based SE model. We combined the losses of both the SE and E2E-ASR models with a tuning parameter α to form the total loss for multi-objective training, as shown in (1).

$$\text{loss}_{\text{total}} = (1-\alpha) \times \text{loss}_{\text{SE}} + \alpha \times \text{loss}_{\text{ASR}}, \qquad (1)$$

The L1 loss function was used to calculate the SE loss and the combined CTC and attention losses from the E2E-ASR model was used as the ASR loss. Parameter α was tuned to make the two losses (SE and ASR losses) contribute almost equally to the total loss in (1) and was set in the range $[0.001, 0.002]$ because the ASR loss is usually exceedingly larger than the SE loss. Furthermore, since the ASR model was pre-trained, the parameter α was set to 0 in the first stage of the multi-objective training to learn the SE model alone without considering the ASR loss. This arrangement yielded better SE performance in our preliminary experiments.

2) SE Optimization Integrating Perceptual Loss: In addition to the recognition error used as the ASR loss, we exploited the ASR model to create another objective function that considered the perceptual loss to train the SE model. The respective model architectures are shown on the right side of Fig. 2. The E2E-ASR model extracted the features in the last layer (320) of the encoder from both clean and enhanced utterances, known as deep features. Subsequently, the L1 loss between the clean and enhanced deep features was set as another form of loss ($\mathbf{loss}_{\mathrm{PL}}$), which was combined with the spectrum-wise SE loss ($\mathbf{loss}_{\mathrm{SE}}$) for multi-objective training, as shown in (2). The perceptual loss evaluated the distance between clean and enhanced signals at the layer level inside the ASR model; thus, it is considered as another SE loss for the ASR. During training with the perceptual loss, only the pre-trained ASR model is required without the corresponding transcriptions for the enhancement training data.

$$\mathbf{loss}_{\mathrm{total}} = (1-\alpha) \times \mathbf{loss}_{\mathrm{SE}} + \alpha \times \mathbf{loss}_{\mathrm{PL}}, \tag{2}$$

where α is a tuning parameter.

3) SE Optimization Integrating Three Losses: Both the ASR loss ($\mathbf{loss}_{\mathrm{ASR}}$) and the perceptual loss ($\mathbf{loss}_{\mathrm{PL}}$) can be used for training simultaneously. $\mathbf{loss}_{\mathrm{ASR}}$ guides the SE model to generate enhanced speech with better prediction results in the recognition model, while $\mathbf{loss}_{\mathrm{PL}}$ can make the model generate enhanced speech with prediction results that are closer to the clean speech by leveraging a pre-trained ASR. The combined loss of all three losses is:

$$\begin{aligned}\mathbf{loss}_{\mathrm{total}} = (1-\alpha_1-\alpha_2) \times \mathbf{loss}_{\mathrm{SE}} + \alpha_1 \\ \times \mathbf{loss}_{\mathrm{ASR}} + \alpha_2 \times \mathbf{loss}_{\mathrm{PL}},\end{aligned} \tag{3}$$

where α_1 and α_2 are the weights for $\mathbf{loss}_{\mathrm{ASR}}$ and $\mathbf{loss}_{\mathrm{PL}}$.

IV. Experiment

Two datasets were used to evaluate the proposed architecture: the TIMIT corpus [53] and the Taiwan Mandarin version of hearing in noise test (TMHINT) sentences [54]. The following section introduces the experimental setups and presents the evaluation results and respective analyses and discussion.

A. Experiments on the TIMIT Dataset

For the experiments conducted on the TIMIT database with multiple noise sources, 10,000 noisy-clean paired training utterances were used, comprising 3,696 utterances in the training set with an average duration of 4 seconds and their noisy counterparts containing 102 noise types from [55] at six different SNR levels (20, 15, 10, 5, 0, and −5 dB). The core test set of TIMIT (including 192 utterances) was mixed with five unseen noise types at four SNR levels (5, 0, −5, and −10 dB) to build the test set in our experiments. The training and test sets did not share common speakers.

TABLE I
Averaged PESQ and STOI Scores for SE-AM Systems With BPC(M) and Phoneme, and SE-E2E–ASR Systems With BPC(M), Phoneme, and Word. The Scores for the Unprocessed and SE Baseline are Listed for Comparison

		PESQ	STOI
Noisy		1.563	0.650
SE Baseline		1.689	0.662
SE-AM	Phoneme	1.730	0.669
	BPC(M)	1.732	0.667
	BPC(D)	**1.744**	**0.674**
SE E2E-ASR	Word	1.688	0.667
	Phoneme	1.753	0.669
	BPC(D)	**1.803**	**0.681**

The speech waveforms were recorded at a 16 kHz sampling rate and converted into 257-dimensional spectrograms using the short-time Fourier transform with a Hamming window size of 32 ms and a hop size of 16 ms. The log1p function ($\mathrm{log1p(x)} = \log(1+x)$) was adopted on the magnitude spectrogram to ensure non-negative outputs [56], and normalization was performed on the waveform. The test stage combined the enhanced magnitude spectral features and original phases from the noisy signals to synthesize the enhanced signals.

As previously mentioned, two-stage training was applied to train the SE model. The SE model was first trained for 70 epochs without considering the ASR results (by setting $\alpha = 0$ in Eqs. (1)) and (2)), and was further updated with the combined objective loss, as in (1)) and 2 (by setting $\alpha = 0.001$ for the ASR loss and $\alpha = 0.05$ for the perceptual loss) for the next 80 epochs. For the experiment that combines all three losses, we have reduced the weightage of α to half for both the ASR loss and perceptual loss. Specifically, we set $\alpha_1 = 0.0005$ and $\alpha_2 = 0.025$. This was done to balance the contribution of each loss and ensure that the SE model is trained to produce enhanced speech that is not only better for ASR tasks but also closer to clean speech. The ASR model was pre-trained with the clean dataset, and its parameters were fixed (frozen) when learning the SE model. To prevent overfitting, we performed early stopping based on validation performance. We used the Adam optimizer [57] with a fixed learning rate of 5×10^{-5} for all of our experiments. The SE model and ASR model had sizes of 33.8 MB and 30.4 MB, respectively. Our experiments were conducted on a GeForce RTX 2080 Ti, with an average training time of 0.88 milliseconds per frame. Please note that the ASR model was only used in the training phase but not in the testing phase.

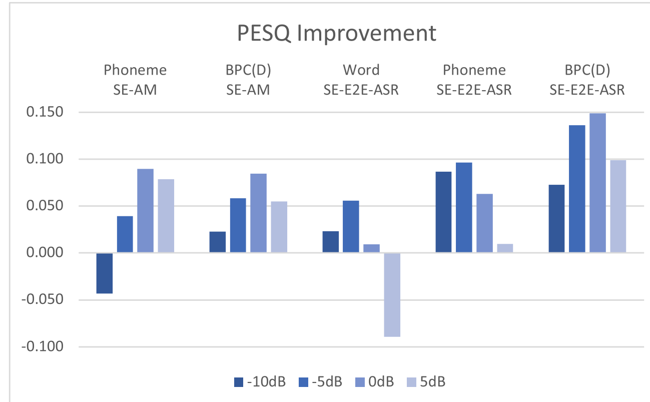
Based on the model architecture illustrated in Fig. 2, we implemented three systems using three types of BPCs, namely BPC(M), BPC(P), and BPC(D), as the acoustic units in the ASR model. In the following, we refer SE by leveraging AM only and E2E-ASR as SE-AM and SE-E2E-ASR, respectively.

1) Results for the Multi-Objective Training Models: Table I presents the PESQ and STOI scores of some baseline models and the proposed SE scenario. In the following experiments, unless otherwise specified, we train SE systems using a combined E2E-ASR and L1 loss ((1)). In addition to the unprocessed baseline (denoted as Noisy), we considered the transformer-based SE without ASR loss and perceptual loss as the first advanced baseline for comparison (denoted as SE baseline). Subsequently, the SE-AM applied the multi-objective training of the SE model and the DNN-HMM-based AM model with phoneme, BPC(M), and BPC(D) as the acoustic unit individually. For the proposed SE-E2E-ASR system, three kinds of acoustic units, namely word, phoneme, and BPC(D), were individually used for the E2E-ASR module to evaluate the corresponding SE results.

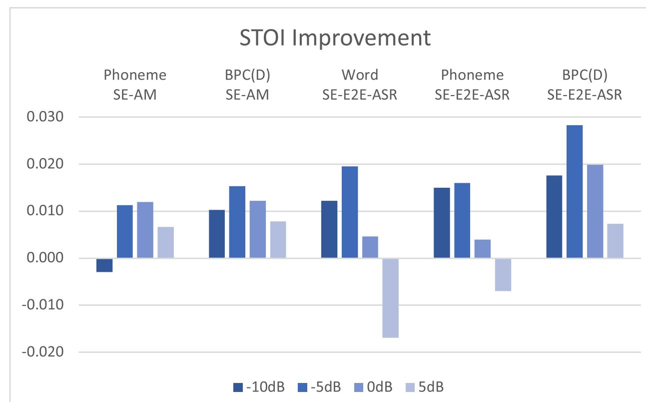
The initial experiments compared the performance of the SE-AM system using BPC(M) and BPC(D) as acoustic units. The results, as shown in Table I, revealed that while BPC(M) provided scores similar to phonemes, BPC(D) outperformed phonemes in terms of both PESQ and STOI scores, indicating that using BPC(D) as the target for the multi-objective training in the SE model is a better choice. Additionally, when using phonemes as the acoustic unit, SE-E2E-ASR outperformed SE-AM in PESQ but had a similar performance in STOI. The word-level SE-E2E-ASR model outperformed the SE baseline in STOI but performed worse in PESQ. Overall, the phoneme-level approaches improved the SE baseline, BPCs-level approaches outperformed the phoneme-level approaches, and the SE-E2E-ASR approaches outperformed the SE-AM approaches.

In contrast to the other methods evaluated, the proposed BPC(D)-level SE-E2E-ASR model showed the best performance, with improvements of 0.114 and 0.019 for PESQ and STOI, respectively, compared to the SE baseline. Wilcoxon sign rank tests were performed to measure the performance difference between the BPC(D)-level SE-E2E-ASR and all other methods listed in Table I, including the SE Baseline, three SE-AM strategies, and the Word/Phoneme SE-E2E-ASR approaches. These comparisons were conducted using all 3840 test samples (made up of 192 utterances, 5 unseen noise types, and 4 SNR levels). All the computed p-values were less than the adjusted significance level of 0.01 after applying a Bonferroni correction for multiple tests ($0.05/5 = 0.01$), providing strong evidence of the superiority of the BPC(D) approach. These results support the hypothesis that incorporating contextual information about the articulation transition between consecutive BPC labels can effectively enhance the quality and intelligibility of processed speech.

Furthermore, we compared these systems in different SNR scenarios, which are shown in Fig. 3. Based on the SE baseline as a reference, Fig. 3 illustrates the improvements in PESQ and STOI achieved by these systems. From this figure, we observe that both the phoneme- and BPC-level SE-AMs work well in PESQ and STOI for the three higher SNR cases (-5 dB, 0 dB, and 5 dB), but perform worse when the SNR is as low as -10 dB. This is because it is difficult for the DNN-HMM AM to recognize severely distorted speech and hence it fails to guide the connected SE model. However, the three SE-E2E-ASR systems can enhance the -10 dB SNR utterances exceedingly well, showing that E2E-ASR performs better than DNN-HMM AM in promoting SE in low-SNR cases. The BPC-level SE-E2E-ASR provided higher STOI scores, verifying our hypothesis that contextual broad phonetic information helps in the learning of the SE model.

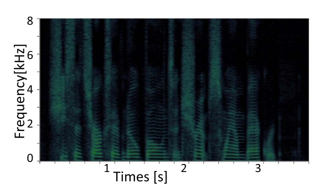


(a) PESQ Improvements over the SE baseline

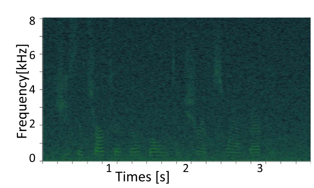


(b) STOI Improvements over the SE baseline

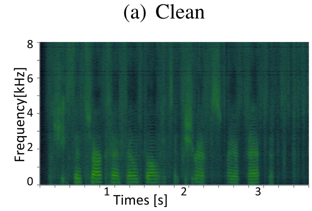
Fig. 3. Average PESQ and STOI improvements of the two SE-AM and three SE-E2E-ASR systems over the SE baseline on different SNR ranges of the TIMIT corpus.



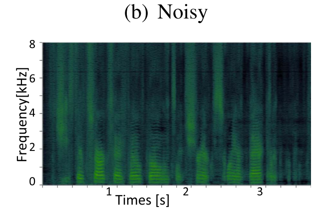
(a) Clean



(b) Noisy



(c) SE Baseline



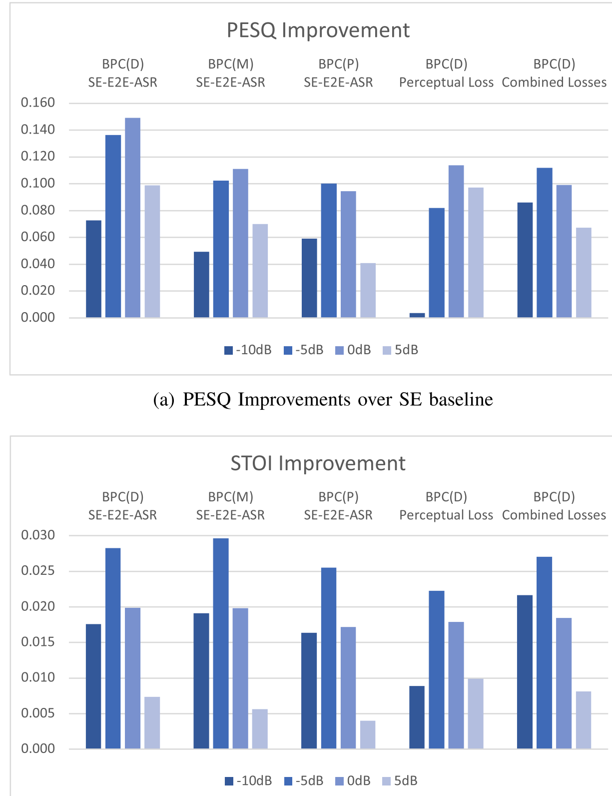
(d) BPC(D)

Fig. 4. Spectrogram plots of an utterance at different situations: (a) clean noise-free, (b) noise-corrupted, (c) noise-corrupted and enhanced by the SE baseline, (d) noise-corrupted and enhanced by the BPC(D)-level SE-E2E-ASR.

2) Spectrogram for SE-E2E-ASR: In addition to the quantitative evaluations, we presented the spectrogram plots for the tested utterance in Fig. 4 to demonstrate the differences in the SE methods. It shows that the spectrogram of clean speech is

TABLE II
AVERAGE PESQ AND STOI SCORES FOR THE SAME BPC-LEVEL SE-E2E-ASR SYSTEMS WITH THREE LOSSES: ASR, PERCEPTUAL AND COMBINED LOSSES. SCORES OF THE SE BASELINE ARE LISTED FOR COMPARISON

		PESQ	STOI
SE Baseline		1.689	0.662
ASR Loss + L1 ((1))	BPC(M)	1.772	**0.681**
	BPC(P)	1.762	0.678
	BPC(D)	**1.803**	**0.681**
Perceptual Loss + L1 ((2))	BPC(M)	**1.764**	**0.677**
	BPC(P)	1.722	0.663
	BPC(D)	1.763	**0.677**
Combined Losses ((3))	BPC(M)	1.777	0.677
	BPC(P)	1.746	0.667
	BPC(D)	**1.780**	**0.681**



(a) PESQ Improvements over SE baseline

(b) STOI Improvements over SE baseline

Fig. 5. Averaged PESQ and STOI improvements of various BPC SE-E2E-ASR systems with ASR and perceptual losses over the SE baseline on different SNR sets for the TIMIT corpus.

severely distorted by noise, while it can be markedly enhanced by the two SE methods. However, there are still moderate residual noises/artifacts in the case of the SE baseline, while they are considerably suppressed by the presented BPC(D)-level SE-E2E-ASR.

3) Results for the SE-E2E-ASR With ASR Loss and Perceptual Loss: In this subsection, we intend to explore the effects of different losses reported in (1), (2), and (3) and thus use the same SE model architecture throughout the experiments. Table II presents the PESQ and STOI scores of the proposed BPC-level SE-E2E-ASR models employing the ASR and perceptual losses, as shown in Fig. 2. Three types of BPC, namely BPC(M), BPC(P), and BPC(D), were individually used as the target for the E2E-ASR model. When using (1) and (2) to train the SE model, we denoted the results as "ASR Loss + L1" and "Perceptual Loss + L1", respectively. For the "Combined Losses" case, the two losses (perceptual loss and ASR loss) from ASR were both used during training as (3). Based on the Table I and II, the following observations were made:

1) Compared with the results of the phoneme-level SE-AM and SE baseline shown in Table I, almost all the BPC-level SE-E2E-ASR systems achieve better PESQ and STOI scores (except for the BPC(P)-level system with perceptual loss). Integrating SE loss with either ASR loss or perceptual loss (as in (2) and (3)) exhibited superior SE performance. Thus, we verified the effectiveness of these contextual and articulatory features in SE.
2) As for the three types of BPCs used in the SE-E2E-ASR with the ASR loss, BPC(P) performed worse than BPC(M) and BPC(D), BPC(D) achieved the optimal PESQ score, and BPC(D) and BPC(M) achieved similar STOI scores. Therefore, the combination of confusion phonemes performed in BPC(D) facilitates the SE module to provide better speech quality, and the clustering methods used in BPC(M) and BPC(D) help to improve speech intelligibility.
3) Regarding the system using the perceptual loss and combined losses, BPC(D) and BPC(M) performed better than BPC(P) in PESQ and STOI metrics, which agreed with the results for the system with the ASR loss. Accordingly, we verified that the performance of the proposed SE-E2E-ASR depends on how we cluster the phonemes.
4) The experiments with combined losses showed that while the majority of the results outperformed the perceptual loss experiments, they still fell short of the results obtained from the ASR loss for all types of BPCs. Surprisingly, the PESQ score for the BPC(M) combined losses experiment even outperformed the BPC(M) ASR loss experiment, suggesting that the combination of different loss functions can complement each other and has the potential to further enhance speech quality.

Even though the systems with perceptual loss achieved lower average PESQ and STOI scores than those with ASR loss, this does not necessarily apply to the individual SNR situation. Fig. 5 shows the PESQ and STOI improvements over the SE baseline for several BPC-level systems at the four SNRs. From this figure, we observe that the three systems with ASR loss exhibit similar trends of improvement with the different SNRs, whereas the BPC(D)-level system with perceptual loss performs quite well in STOI for the high-SNR case (5 dB), outperforming the three systems with ASR loss. On the other hand, the BPC(D)-level system with combined losses performs the best in both PESQ and STOI for the low-SNR case (−10 dB). Furthermore, Note that the ASR loss and the perceptual loss are two types of losses that serve different purposes. The ASR loss aims to improve the accuracy of ASR results, while the perceptual loss measures the difference between clean and enhanced speech at an intermediate level of the ASR model. The results from Table II suggest that when training SE models, the ASR loss might offer additional and valuable information that complements the L1 loss.

B. Overall Discussion of the TIMIT Experiments

1) Misclassification of Phonemes Causes Poor Feedback: From the experiments on the TIMIT dataset, we observed that misclassification of phonemes by the ASR system can lead to poor feedback, as discussed below:

a) Our experiments shows that BPC-level objectives outperformed phoneme-level objectives, suggesting that distinguishing between confusable phonemes may not be as helpful as correctly classifying groups of phonemes. When the SE model learns to generate speech that overly emphasizes the difference between similar phonemes, the generated speech may not necessarily be an improvement.

b) We also found that correct objective feedback from the ASR loss performs better than the soft objective from the perceptual loss. This indicates that misclassification results from clean speech reduce the improvement from the ASR feedback, highlighting the importance of accurate feedback for effective model training.

c) BPC(D) performs the best in almost all the experiments, indicating that combining the most-confusable targets is the most helpful for the additional objective. On the other hand, the place of articulation is not as critical as the manner of articulation for the shape of the audio waveform [37], meaning that the phonemes in the same group of BPC(P) are not confusable and lead to the worst performance among all the BPCs.

Based on the above observations, we conclude that objectives with misclassification of phonemes by the ASR system may lead to inadequate feedback for SE models. Although the experiments were conducted on a relatively small set of training data (3,696 clean utterances), the performance improvement of low-resource training conditions is still valuable for practical application. However, it is worth noting that the advantages of the knowledge-based approach (BPC(M) and BPC(P)) may be reduced as the amount of data increases.

2) Contextual Acoustic Feedback From the E2E-ASR Model: Most previous studies that apply feedback from the ASR objective for SE use losses from AM, which provides frame-wise feedback [35], or contextual E2E-ASR with word-level objective feedback [31], [32]. In contrast, our approach applies phoneme-level E2E-ASR feedback for the SE system. This approach has benefits in that the ASR model learns to predict phonemes as a sequence instead of individually for each time segment. The benefits of this approach are listed below:

a) Compared to AM feedback like [31], [32], phoneme/BPCs-level E2E-ASR feedback can guide the SE model with the level of the whole utterance instead of individual time segments. One of the advantages of using phoneme-level E2E-ASR feedback for SE is that it allows for better modeling of the temporal relationships between speech features and phonemes. In traditional ASR systems, phonemes are typically modeled using hidden Markov models (HMMs), which may not take into account the temporal structure of the entire speech utterance. However, in phoneme-level E2E-ASR, the ASR model learns to predict phonemes as a sequence, which allows for better modeling of the dynamic relationships between speech features and phonemes. The consistent results of Phoneme/BPC(M)/BPC(D)-level E2E-ASR feedback outperform corresponding AM feedbacks in Tables I and II, supporting this statement.

b) Another advantage of using phoneme/BPCs-level E2E-ASR feedback for SE is that it provides more direct and informative feedback to the SE model compared with word-level E2E-ASR feedback like [35]. As we mentioned earlier, phonemes are closer to the audio features compared to words, which makes the phoneme-level feedback more relevant and useful for guiding the SE model. Moreover, phoneme-level feedback is more fine-grained than word-level feedback, which allows for better differentiation of the different phonemes and their acoustic characteristics. The results of Word-level E2E-ASR feedback perform the worst among the Word/Phoneme/BPC(D)-levels in Table I, supporting this statement.

These observations show that incorporating contextual broad phonetic information to learn the SE model, as in BPC-level SE-E2E-ASR, is most helpful in reconstructing the original clean signal and removing the interference. It's worth noting that while phoneme/BPCs-level E2E-ASR feedback has several advantages for SE, it also has some potential limitations. For example, phoneme/BPCs-level ASR feedback may require transcription training labels compared to other types of feedback such as perceptual loss. This can be a limitation in low-resource settings where obtaining large amounts of labeled data is challenging. Additionally, the phoneme-level feedback may not be as effective for languages with complex phonetic systems. However, the consistent improvement of the Phoneme/BPC(M)/BPC(D)-level E2E-ASR feedback over the AM feedback and the Word-level E2E-ASR feedback in our experiments suggests that phoneme/BPCs-level E2E-ASR feedback is a promising approach for improving the performance of SE systems.

C. Experiments on the TMHINT Dataset

International Phonetic Alphabet (IPA) presents the phones used in all languages. Therefore, articulation feature classification methods using BPCs can also be applied to other languages. For the TMHINT corpus, we transferred the Chinese characters into IPA phone sequences and categorized these phones into BPC clusters.

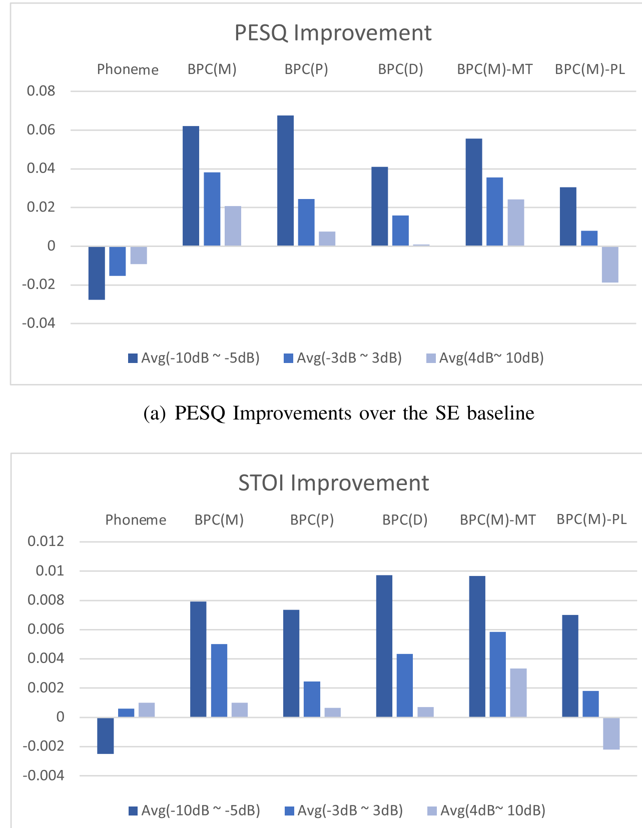
For the training set, we used 10,000 noisy-clean paired training utterances. The paired training set contained 1,200 clean utterances with an average duration of 3.5 seconds and their noisy mixed speech using 104 noises with multiple noise sources [55] (at 31 SNR levels from -10 to 20 dB). For the test set, 640 utterances were mixed with seven unseen noise types at 14 SNR levels (ranging from -10 to 10 dB). The training set included three male and three female speakers, and the testing set contained one male and one female. The SE-E2E-ASR experiments used various ASR labels, such as phoneme, BPC(M), and BPC(P).

TABLE III
AVERAGED PESQ AND STOI SCORES FOR SE-E2E-ASR SYSTEM ON TMHINT CORPUS

		PESQ	STOI
Noisy		1.572	0.684
SE Baseline		2.029	0.725
SE E2E-ASR	Phoneme	2.012	0.724
	BPC(M)	**2.068**	**0.729**
	BPC(P)	2.060	0.728
	BPC(D)	2.047	**0.729**
	BPC(M)-MT	**2.066**	**0.731**
Perceptual Loss	BPC(M)	2.034	0.726

Furthermore, we evaluate the ability of the clustering approach in English used in BPC(D) to generalize to other languages. Specifically, we apply the BPC(D) clusters trained on English data to a denoising experiment on TMHINT by mapping the Chinese phonemes to their corresponding IPA symbols. We made some adjustments to the original BPC(D) clusters by removing redundant phonemes and grouping new phonemes based on their manner of articulation, resulting in a total of nine groups, as in the English experiment. This evaluation allows us to test the generalizability of the data-driven approach across different languages and to assess whether the phoneme clusters learned from one language can be applied to another language. The transformer was set as the baseline SE model and learned jointly with the connected ASR model with extra BPC semantic information. A BPC(M)-ASR classification model trained with noisy speech was also examined. To test the SE-E2E-ASR methods in the experiments for noisy-reverberant utterances, we selected BPC(M) as the recognition unit.

1) Results for Speech Denoising: The model structure used here was similar to that described in Section IV-A. As the value of the ASR loss was larger than the SE loss in this task, we lowered the parameter α in (1) to 0.0001 to equally weight the two losses in the total loss function. The resulting PESQ and STOI scores are presented in Table III. From this table, we first observe that the BPC-level SE-E2E-ASR model can improve both the PESQ and STOI scores compared to the SE baseline, while the mono-phoneme SE-E2E-ASR model compromises the SE improvement. These results differ from those obtained for the TIMIT task described in Section IV-A. This might be because Mandarin Chinese is a tonal language, where the classification of respective phonemes may be less helpful for the SE model. Second, when the English data-driven cluster BPC(D) is applied to the TMHINT corpus, it produces the lowest PESQ score among the BPC-level SE-E2E-ASR approaches. Our investigation in IV-B1 suggests that the inaccurate classification of phonemes can lead to unsatisfactory results. Considering the distinctive acoustic properties of Chinese, it is reasonable to assume that BPC(D) needs further customization, including the incorporation of tonal features, for Chinese corpora. To validate this hypothesis, future studies can explore various phoneme groups specifically designed for the Chinese language. Comparatively, using BPC(M) as the acoustic unit for SE-E2E-ASR resulted in the best performance among all selections, including monophonic, BPC(P), and BPC(D). Since BPC(M) (and BPC(P)) is designed based on the property of IPA-level phones, the cluster is identical to the BPC(M) we use in English and could potentially be used for cross-language SE training in the future.



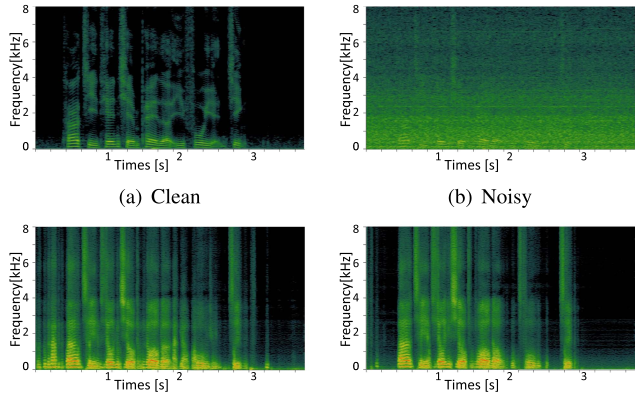
(a) PESQ Improvements over the SE baseline

(b) STOI Improvements over the SE baseline

Fig. 6. PESQ and STOI improvements of SE-E2E-ASR system over the SE baseline averaged on different SNR sets for the TMHINT corpus.

In addition to the experiments, wherein the E2E-ASR model was trained with clean, noise-free utterances, we used noisy utterances to train the E2E-ASR model and then conducted the respective SE-E2E-ASR experiments. We randomly selected the noise signals within 104 different noise types mixed with 1,200 utterances as the training set to train the CTC/attention E2E-ASR model. The subsequent SE-E2E-ASR experiments with the multi-condition trained E2E-ASR model adopted the same training configuration as those with the clean E2E-ASR model mentioned above, with BPC(M) as the acoustic unit. The obtained PESQ and STOI scores, which are listed in Table III with the label "BPC(M)-MT," are quite close to those of the clean BPC(M)-level E2E-ASR model. These results clearly show that whether the ASR model is trained by clean or noisy utterances does not considerably influence the average SE performance of SE-E2E-ASR if BPC(M) is used.

Fig. 6 shows the PESQ and STOI improvements of different SE-E2E-ASR systems over the SE baseline at different SNR sets (-10 to -5 dB, -3 to 3 dB, and 4 to 10 dB). As shown in the figure, almost all systems outperform the SE baseline, except for the phoneme-level system and the perceptual loss system (at high SNRs). These results demonstrated the effectiveness of articulatory features of BPCs for SE particularly for exceedingly noisy (low-SNR) situations. By contrast, the phoneme-level E2E-ASR



(a) Clean (b) Noisy

(c) Baseline (d) BPC(M)

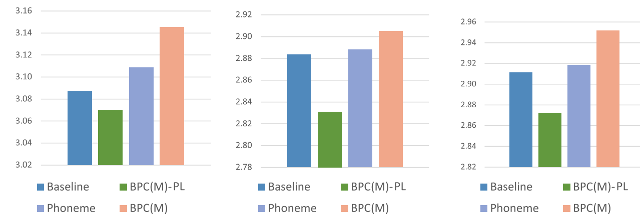
Fig. 7. Spectrogram plots of an utterance at different situations: (a) clean noise-free, (b) noise-corrupted, (c) noise-corrupted and enhanced by the SE baseline, (d) noise-corrupted and enhanced by the BPC(M)-level SE-E2E-ASR.

may not benefit the connected SE at low SNRs probably due to its poor recognition accuracy.

Additionally, in Fig. 7, we display the spectrograms of the clean speech, its noisy counterpart, and their enhanced versions at an engine noise SNR of -5 dB for qualitative comparison. It is evident from the figure that the SE baseline fails to entirely remove the noise in non-speech regions, whereas the proposed BPC(M)-level SE-E2E-ASR model better suppresses the noise in these areas, resulting in a spectrogram that more closely resembles clean speech. This finding reaffirms that the contextual information of the BPCs enhances SE performance.

2) Listening Test: To further evaluate the effectiveness of the proposed approach, a listening test was conducted for the TMHINT experiments.The test set included two challenging noise types - engine and street noises - with two different SNR levels (-5 and 5 dB). The four processing approaches - the SE baseline, BPC(M) with perceptual loss (with loss of (2)), Phoneme, and BPC(M) (with loss of (1))- were tested with each noise type and SNR level as a total of 16 conditions: 2 SNR level $\times$ 2 noise types $\times$ 4 processing approaches, each containing ten randomly selected sentences. The order of the conditions was also randomized. The subjective quality of the enhanced speech utterances was evaluated using mean opinion score (MOS) tests, with 17 subjects asked to judge the quality of the audio for signal distortion (SIG), background intrusiveness (BAK), and overall quality (OVL) using a five-point scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent) [58].

In the SIG test, the subjects were asked to rate the natural level of the speech signals after listening to an enhanced speech utterance processed by the four different methods. A higher score indicates that the speech signals are more natural. The SIG results are shown in Fig. 8(a). We found that although the perceptual loss approach (BPC(M)-PL) performs worse than the results from the baseline, the ASR objectives (Phoneme and BPC(M)) improve the scores, and BPC(M) provides the best result. This shows that the E2E-ASR objective alleviates the distortion of the recovered signal and provides better speech quality. For the BAK test, the subjects were asked to judge the level of noise artifact perceived after listening to an utterance, and a higher score



(a) scores of signal distortion (SIG) (b) scores of background intrusiveness (BAK) (c) scores of overall quality (OVL)

Fig. 8. Listening test results in terms of SIG, BAK, and OVL scores. The "Phoneme" and "BPC(M)" are the approaches with E2E-ASR objectives, and the "BPC(M)-PL" is the approach with the perceptual loss.

TABLE IV
WERs (%) of Three SE-E2E-ASR Systems (Phoneme, BPC(M), and BPC(M) Perceptual) and SE Baseline in Different SNRs With Four Noise Types on the TMHINT Dataset

SNR	SE Baseline	BPC(M)-PL	Phoneme	BPC(M)
−5	85.3%	81.4%	84.3%	82.6%
−1	58.7%	53.7%	57.9%	55.4%
1	44.3%	40.2%	42.3%	42.2%
5	20.8%	19.2%	20.5%	19.5%
Avg.	52.3%	48.6%	51.3%	49.9%

indicated a lower level of noise artifact perceived. In Fig. 8(b), BPC(M) was the only method that outperformed the baseline in BAK scores, while Phoneme performed close to the baseline and the BPC(M)-PL approach decreased the score. Finally, for the overall quality (OVL) in Fig. 8(c), while the BPC(M)-PL still compromises the results, the results of the Phoneme approach are slightly better than the results from the baseline, and the BPC(M) approach performs the best in the overall quality. The results indicate that while the perceptual loss approach did not perform well in subjective evaluations, both Phoneme-level and BPC-level E2E-ASR objectives yielded less distorted outcomes, and BPC targets appeared to be more effective in eliminating noise and providing better speech quality. To supplement our analysis, we also conducted the Wilcoxon sign rank tests to compare the four approaches in terms of overall quality (OVL), background intrusiveness (BAK), and signal distortion (SIG) scores. The test results show that in terms of OVL and BAK scores, there is a statistically significant improvement when comparing the BPC(M) method with the SE baseline and BPC(M)-PL approaches ($p < 0.05$). However, for SIG scores and the Phoneme approach, no significant differences were found. One possible reason for the lack of significant difference in SIG scores and the Phoneme approach might be due to a high variation among the listeners' responses, combined with a relatively small sample size. These factors make it more challenging to reach statistically significant results.

3) ASR Results: We compared the performance of different SE approaches on automatic speech recognition (ASR) using Google Speech-to-Text [59] to compute word error rates (WER) compared with the SE baseline. We conducted experiments with four real-world noise types - engine, babble, street, and three-talker - to simulate practical environments. Table IV summarizes our findings. Specifically, we evaluated the WERs of

TABLE V
AVERAGE PESQ AND STOI SCORES FOR SE-E2E-ASR SYSTEM WITH BPC(M) ON TMHINT CORPUS WITH NOISE AND REVERBERATION

		PESQ	STOI
Noisy Reverberation		1.373	0.482
SE Baseline		1.284	0.536
SE-E2E-ASR	Phoneme	1.288	0.540
	BPC(M)	**1.384**	**0.555**

speech enhanced using Phoneme, BPC(M), and BPC(M)-PL objectives. Our results indicate that all Phoneme/BPC-level objectives (i.e., Phoneme, BPC(M), and BPC(M)-PL) resulted in a reduced WER compared to the SE baseline. Notebally, the BPC-based enhancement methods (BPC(M) and BPC(M)-PL) surpassed the Phoneme-level SE-E2E-ASR approach. Among all, BPC(M)-PL yielded the best results across all signal-to-noise ratio (SNR) conditions.

Our experiments demonstrate that BPC-based enhancement approaches are promising pre-processing modules for speech recognition in practical settings. Our findings indicate that the perceptual loss approach, despite some compromises in subjective evaluation, yields more accurate recognition results than the Phoneme and baseline approaches. Meanwhile, the Phoneme/BPC-level E2E-ASR objective approaches strike a balance between subjective evaluation and recognition results, achieving improved performance across all evaluation metrics. Specifically, the BPC(M)-level E2E-ASR objective approach outperformed all the other approaches in all experiments.

4) Results for Speech Denoising and Dereverberation: In addition to the denoising task, we evaluated the proposed methods on utterances further corrupted by reverberation, which is a more challenging task. For the training set, the clean utterances from the TMHINT corpus were first mixed with 104 noise types at 31 SNR levels from -10 to 20 dB and then distorted by reverberation. The test set had reverberated and noise-corrupted utterances, where seven unseen noise types were added at 14 SNRs (from -10 to 10 dB). We used the room impulse response (RIR) generator to create the reverberation, and room impulse responses were generated using the image method and applied to the noisy utterances. The reverberation time (T60) was randomly selected from 0.3, 0.6, and 0.9 s to generate impulse responses for the training data, and T60 was set to 0.4 s for the testing data. The RIR had a total of 4,096 samples. The receiver position was [2 m, 1.5 m, 2 m], the source position was [2 m, 3.5 m, 2 m], and the room dimensions were [5 m, 4 m, 6 m]. The experiment aimed to remove both the noise and reverberation; thus, the clean data was set as the target when training the SE model.

Table V shows the average PESQ and STOI scores of the original reverberant noisy data and their two enhanced versions. It is observed from this table that the SE baseline improves STOI by 0.054 but worsens PESQ by 0.089. The phoneme-level objective provides slightly better but still degrades the PESQ score. According to [61], [62], a single DNN-based SE model may produce limited performance in a composite of noisy and reverberant condition: the STOI scores can be improved, but PESQ scores show no improvements, which matches our results. On the other hand, our proposed approach, BPC(M)-level SE-E2E-ASR objective, promotes both STOI and PESQ by 0.073 and 0.011, respectively, on average. Therefore, we have demonstrated that the proposed system can effectively handle noise and reverberation issues simultaneously.

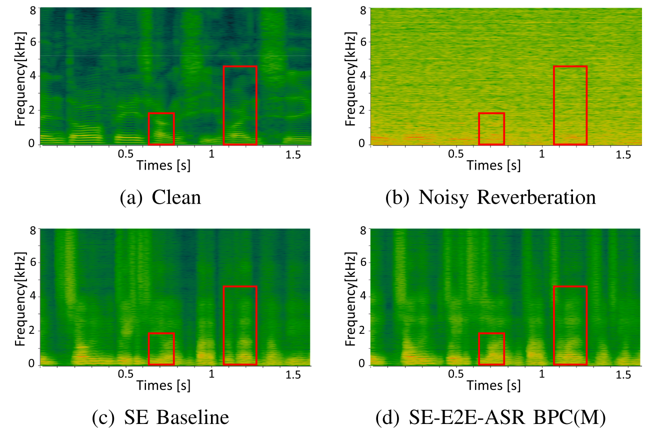


Fig. 9. Spectrogram plots of an utterance at different situations: (a) clean noise-rev-free, (b) noise-rev-corrupted, (c) noise-rev-corrupted and enhanced by the SE baseline, (d) noise-rev-corrupted and enhanced by the BPC(M)-level SE-E2E-ASR.

The spectrograms of a clean utterance, its noisy-reverberant counterpart, and their enhanced signals are shown in Fig. 9. We highlighted two speech regions in the clean utterance and used them to compare the enhanced versions from the SE baseline and BPC(M)-level SE-E2E-ASR. We observe from this figure that the SE baseline does not recover the speech signals in the regions of the red box, whereas the speech is preserved for the proposed BPC(M)-level SE-E2E-ASR case.

5) Results for Impaired SE: The impaired utterances used in this study were based on the TMHINT corpus, which has 1,200 Mandarin Chinese utterances. We used non-impaired utterances from the TMSV corpus [63] as the target speech data and dysarthric utterances as the input-impaired data to train the SE model. The TMSV corpus included 13 males and five female speakers, from which utterances from 13 male and four female speakers with better pronunciation accuracy were used in our task. The speech content in the TMSV corpus is similar to that in the TMHINT corpus. For each speaker, 240 utterances were used for training, 40 for validation, and 40 for testing. The 80-dimensional mel-spectrograms with 1024 DFT points and a 144-point frameshift for the utterances were extracted using the open-source ESPnet toolkit. The TTS-transformer model was pre-trained and used for the ISE task. Referring to (1), the transformer-based SE model was trained for 1000 epochs in advance with the SE loss ($\mathbf{loss}_{SE}$) and then further trained for another 1000 epochs by adding the BPC(M)-level E2E-ASR loss ($\mathbf{loss}_{ASR}$) to the SE loss. Parameter α was set to 0.001. Early stopping was implemented during training.

Subjective evaluations were conducted to evaluate the performance of impaired speech experiments. We randomly selected four utterances produced by 10 speakers (eight males and two females), which were processed by either the SE baseline or SE-E2E-ASR. A total of 21 subjects performed the evaluation, each given 120 samples (four samples $\times$ 10 speakers $\times$ three

TABLE VI
SUBJECTIVE PREFERENCE EVALUATION FOR THE IMPAIRED SE

	Preference
Unprocessed	16.8%
SE Baseline	36.8%
SE-E2E-ASR BPC(M)	**46.4%**

approaches) and requested to choose his/her preference among the three types of utterances. The listening test was conducted in a quiet environment, with an SNR level of approximately 55 dB. A Sennheiser HD599 headset was used, and the audio was played through a Samsung Tab S6 device. The resulting preference rates are presented in Table VI. It is surprising that the unprocessed utterances were least preferable at 16.8%, while both the SE baseline and SE-E2E-ASR with BPC(M) achieved more than twice the preference rate than the unprocessed speech, revealing the effectiveness of the SE models in amending impaired speech. In particular, the proposed SE-E2E-ASR obtained a 9.6% higher preference rate than the SE baseline, showing that the multi-objective training of SE and BPC(M)-level E2E-ASR can further improve the speech quality of utterances converted from impaired speech.

V. CONCLUSION

This study proposed a novel architecture that applies a BPC-based E2E-ASR to guide the SE process with contextual broad phonetic information to achieve superior speech quality and intelligibility. Three BPC clustering methods were investigated for the English corpus, and the evaluation results confirmed the context information of the BPC SE considerably over a wide range of SNR conditions. Furthermore, with the word-to-IPA transformation, we have extended the use of this novel approach to the Mandarin corpus with similar BPC clusters as in the English corpus experiments. Experimental results on three tasks, namely speech denoising, speech denoising-and-dereverberation, and impaired speech enhancement, verified the effectiveness of incorporating contextual broad phonetic information into SE to improve enhancement results. The main contributions of this study are as follows: (1) This is the first study that employed the context information of broad phonetic/articulatory phonemes classes for an end-to-end SE–ASR system. (2) We demonstrated that using both knowledge-based and data-driven BPCs as enhancement targets can further improve the quality and intelligibility of enhanced speech for both English and Mandarin. (3) We validated that knowledge-based BPCs are generally more flexible than data-driven BPCs and mono-phonemes, as they can be used in a wider range of scenarios. The main focus of this study is to examine the losses prepared by various pre-trained models, including AM and E2E-ASR systems, to leverage the SE performance. Our experimental results have validated the efficacy of including contextual broad phonetic information in SE training. In the future, we will further utilize the findings of this study to enhance other DL-based speech processing tasks.

REFERENCES

[1] Y.-H. Lai, F. Chen, S.-S. Wang, X. Lu, Y. Tsao, and C.-H. Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 7, pp. 1568–1578, Jul. 2017.

[2] D. Wang, "Deep learning reinvents the hearing aid," *IEEE Spectr.*, vol. 54, no. 3, pp. 32–37, Mar. 2017.

[3] A. J. Accardi and R. V. Cox, "A modular approach to speech enhancement with an application to speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 1999, pp. 201–204.

[4] R. Martin and R. V. Cox, "New speech enhancement techniques for low bit rate speech coding," in *Proc. IEEE Workshop Speech Coding Proc. Model, Coders, Error Criteria*, 1999, pp. 165–167.

[5] 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.

[6] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[7] F. Weninger et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Latent Variable Anal. Signal Separation, 12th Int. Conf.*, 2015, pp. 91–99.

[8] 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 708–712.

[9] F.-A. Chao, S.-W. F. Jiang, B.-C. Yan, J.-W. Hung, and B. Chen, "TENET: A time-reversal enhancement network for noise-robust ASR," 2021, pp. 55–61.

[10] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7009–7013.

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

[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 Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[13] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[14] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5751–5755.

[15] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 153–167, Jan. 2017.

[16] J. Qi, J. Du, S. M. Siniscalchi, and C.-H. Lee, "A theory on deep neural network based vector-to-vector regression with an illustration of its expressive power in speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 1932–1943, Dec. 2019.

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

[18] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. Interspeech*, 2014, pp. 2685–2689.

[19] P. G. Shivakumar and P. G. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement.," in *Proc. Interspeech*, 2016, pp. 3743–3747.

[20] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[21] J. Qi, H. Hu, Y. Wang, C.-H. H. Yang, S. M. Siniscalchi, and C.-H. Lee, "Tensor-to-vector regression for multi-channel speech enhancement based on tensor-train network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7504–7508.

[22] Y. Xian, Y. Sun, W. Wang, and S. M. Naqvi, "Convolutional fusion network for monaural speech enhancement," *Neural Netw.*, vol. 143, pp. 97–107, 2021.

[23] Q. Zhang, A. Nicolson, M. Wang, K. K. Paliwal, and C. Wang, "Deep-MMSE: A deep learning approach to MMSE-based noise power spectral density estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 1404–1415, 2020.

[24] 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 Trans. Emerg. Topics Comput. Intell.*, vol. 2, no. 2, pp. 117–128, Apr. 2018.

[25] C.-F. Liao, Y. Tsao, X. Lu, and H. Kawai, "Incorporating symbolic sequential modeling for speech enhancement," in *Proc. Interspeech*, 2019, pp. 2733–2737.

[26] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," in *Proc. Interspeech*, 2019, pp. 2723–2727.

[27] J. Yao and A. Al-Dahle, "Coarse-to-fine optimization for speech enhancement," in *Proc. Interspeech*, 2019, pp. 2743–2747.

[28] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement," in *Proc. Interspeech*, 2020, pp. 196–200.

[29] M. Yang, J. Konan, D. Bick, A. Kumar, S. Watanabe, and B. Raj, "Improving speech enhancement through fine-grained speech characteristics," in *Proc. Interspeech*, 2022, pp. 2953–2957.

[30] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7118–7122.

[31] D. Bagchi, P. Plantinga, A. Stiff, and E.F. Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5609–5613.

[32] P. Plantinga, D. Bagchi, and E. Fosler-Lussier, "Perceptual loss with recognition model for single-channel enhancement and robust asr," 2021, *arXiv:2112.06068*.

[33] C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu, and Z. Wen, "Gated recurrent fusion with joint training framework for robust end-to-end speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 198–209, 2021.

[34] B. Liu et al., "Jointly adversarial enhancement training for robust end-to-end speech recognition.," in *Proc. Interspeech*, 2019, pp. 491–495.

[35] A. S. Subramanian et al., "Speech enhancement using end-to-end speech recognition objectives," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2019, pp. 234–238.

[36] Y.-J. Lu, C.-F. Liao, X. Lu, J.-W. Hung, and Y. Tsao, "Incorporating broad phonetic information for speech enhancement," in *Proc. Interspeech*, 2020, pp. 2417–2421.

[37] P. Ladefoged and K. Johnson. *A Course in Phonetics*. Toronto, ON, Canada: Nelson Education, 2014.

[38] P. Scanlon, D. P. W. Ellis, and R. B. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 803–812, Mar. 2007.

[39] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "An artificial neural network approach to automatic speech processing," *Neurocomputing*, vol. 140, pp. 326–338, 2014.

[40] J. Lin et al., "Improving mandarin tone recognition based on dnn by combining acoustic and articulatory features using extended recognition networks," *J. Signal Process. Syst.*, vol. 90, pp. 1077–1087, 2018.

[41] S. M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee, "Exploiting deep neural networks for detection-based speech recognition," *Neurocomputing*, vol. 106, pp. 148–157, 2013.

[42] A. S. Shahrebabaki, N. Olfati, S. M. Siniscalchi, G. Salvi, and T. Svendsen, "Transfer learning of articulatory information through phone information," in *Proc. Interspeech*, 2020, pp. 2877–2881.

[43] C. Lopes and F. Perdigão, "Broad phonetic class definition driven by phone confusions," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, 2012, Art. no. 158.

[44] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1240–1253, Dec. 2017.

[45] D. Povey et al., "The kaldi speech recognition toolkit," in *Proc. Autom. Speech Recognit. Understanding*, 2011.

[46] M. Ravanelli, T. Parcollet, and Y. Bengio, "The pytorch-kaldi speech recognition toolkit," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6465–6469.

[47] A. Vaswani et al., "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, 2017.

[48] X. Hao, C. Shan, Y. Xu, S. Sun, and L. Xie, "An attention-based neural network approach for single channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6895–6899.

[49] C.-H. Yang, J. Qi, P.-Y. Chen, X. Ma, and C.-H. Lee, "Characterizing speech adversarial examples using self-attention u-net enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3107–3111.

[50] Y. Koizumi, K. Yaiabe, M. Delcroix, Y. Maxuxama, and D. Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 181–185.

[51] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *Stat*, vol. 1050, p. 21, 2016.

[52] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech*, 2020, pp. 4676–4680.

[53] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," in *Proc. Linguistic Data Consortium*, 1993.

[54] 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.

[55] G. Hu, "100 nonspeech environmental sounds," The Ohio State University, Department of Computer Science and Engineering, 2004.

[56] S.-W. Fu et al., "Boosting objective scores of a speech enhancement model by metricGAN post-processing," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2020, pp. 455–459.

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[58] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. audio, speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[59] A. Zhang, "Speech recognition (version 3.8)[software]," in *Proc. Int. Conf. Commun. China*, 2017.

[60] Z.- Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 796–806, Apr. 2016.

[61] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 6, pp. 982–992, Jun. 2015.

[62] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 53–62, Jan. 2019.

[63] S.-Y. Chuang, H.-M. Wang, and Y. Tsao, "Improved lite audio-visual speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1345–1359, 2022, .