A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement

Li Chai , Jun Du , Qing-Feng Liu, and Chin-Hui Lee , *Fellow, IEEE*

Abstract—A new cross-entropy-guided measure (CEGM) is proposed to indirectly assess accuracies of automatic speech recognition (ASR) of degraded speech with a speech enhancement front-end and without directly performing ASR experiments. The proposed CEGM is calculated in three steps, namely: (1) a low-level representations via feature extraction, (2) a high-level nonlinear mapping using an acoustic model, and (3) a final CEGM calculation between the high-level representations of clean and enhanced speech. Specifically, state posterior probabilities from outputs of conventional hybrid acoustic model of the target ASR system are adopted as the high-level representations and a cross-entropy criterion is used to calculate the CEGM. Due to CEGM's differentiability, it can also be used to replace the conventional minimum mean squared error (MMSE) criterion as an objective function for deep neural network (DNN)-based speech enhancement. Therefore, the front-end enhancement model can be optimized towards improving the accuracies of the back-end ASR system. Experiments on single-channel CHiME-4 Challenge show that CEGM yields consistently the highest correlations with word error rate (WER) which is often costly to calculate, and achieves the most accurate assessment of ASR performance when compared to the perceptual evaluation metrics commonly used for assessing speech enhancement performance. Furthermore, CEGM-optimized speech enhancement could effectively reduce the WER on the CHiME-4 real test set when compared to unprocessed noisy speech and enhanced speech obtained with MMSE-optimized enhancement for ASR systems with fixed multi-condition acoustic models in various deep architectures.

Index Terms—Acoustic model, cross entropy, deep neural network (DNN), robust automatic speech recognition, speech enhancement.

I. INTRODUCTION

AUTOMATIC speech recognition (ASR) has recently achieved improved accuracies by introducing deep neural network (DNN) based acoustic models [1]. However, modern ASR systems still suffer from performance degradations in real world environments due to adverse acoustic conditions, such as channel distortion, ambient noise and reverberation, etc.

Speech enhancement is a commonly used technique to increase speech quality and increase ASR robustness. It acts as front-end processing attempting to remove the corrupting noise from noisy speech prior to back-end speech recognition. Features extracted from clean speech often contain more discriminant information than those of noise-corrupted speech. Therefore, if clean features can be recovered to some degree, better ASR performance could be obtained. Various speech enhancement frameworks have been proposed for achieving this by removing background noise from observed speech, including conventional methods, such as spectral subtraction [2], Wiener filtering [3], minimum mean squared error (MMSE) estimation [4], and an optimally-modified log-spectral amplitude (OM-LSA) speech estimator [5], [6]. These approaches are designed based on some assumptions of speech and noise characteristics. In general, they can yield a satisfactory performance in terms of perceptual quality but may not be directly beneficial to the improvement of ASR performance [7]–[9]. Recently, data-driven approaches based on deep learning have received increased attention, and have been shown to improve noise robustness for ASR when they are combined with beamforming techniques [10]–[12]. However, DNN-based single-channel speech enhancement often leads to performance degradation when it is used as a pre-processing step for ASR systems with multi-condition training due to the spectral distortion induced by speech enhancement [13]–[15].

Quality evaluation for speech enhancement is a very important step in the development of advanced algorithms. However, quality evaluation of the resulting enhanced speech is a complex problem that depends on the application field [16]. In many cases, the main objective of speech enhancement is preserving some characteristics that are required for the downstream task concerned. For example, a good listening quality and intelligibility in terms of human auditory perception is highest priority for speech communication. However, for other applications that require front-end speech processing, the resulting enhanced speech with good human perceptual quality could not guarantee satisfactory performance for back-end processing, such as speech and speaker recognition. For ASR systems, the de facto standard metric to evaluate the performance is word error rate

(WER). Nevertheless, the calculation of WER needs to perform a series of ASR experiments. This is often time-consuming and requires a large amount of computation and manual transcription costs. In addition, evaluating ASR accuracies in many different acoustic conditions is usually not complete that some unseen noise could still degrade ASR system in operating environments. Therefore, objective error evaluation measures, such as those commonly used in traditional pattern recognition systems [17], [18], are often desired. This is beneficial to research in assessing noisy speech recognition and in optimizing speech enhancement algorithms for robust ASR. An example for error rate estimation without evaluating ASR systems can be found in [19]. As far as using a speech enhancement front-end, most techniques use known objective quality evaluation metrics for speech enhancement to predict ASR error rates. Specifically, studies in [20]–[24] used perceptual evaluation speech quality (PESQ) [25] scores to predict WERs. It was found that there is a good correlation between source-to-distortion ratio (SDR) [26] and WER [27]. Moreover, it has been shown that the correlation between WER and short-time objective intelligibility (STOI) [28] is stronger than that between other metrics (e.g., PESQ) [29]–[31]. However, many previous papers [13], [15], [32], [33] have indicated that improvements in objective perceptual quality did not necessarily lead to WER reductions. Since a machine listener, i.e., an ASR system might be more sensitive to speech distortion and noise interferences than a human listener, distortion measures that only focus on partial distortions of degraded speech cannot completely capture the various speech distortions that lead to ASR performance degradation. Accordingly, it is not reliable enough to just employ the conventional enhancement distortion measures to assess ASR. Furthermore, the distortion measures are usually calculated based on speech waveforms and thus are not easily applicable to ASR systems that only require speech enhancement front-end to directly provide speech features to the ASR back-end.

DNN has been successfully applied to single-channel speech enhancement and shown to provide a good improvement in perceptual quality when compared to classical approaches. Due to the powerful modeling capability of deep structures, DNN-based speech enhancement can effectively model the complicated relationship between noisy and clean speech [34]. It can also estimate a mask to suppress noise from noisy speech [35], [36]. The MMSE is usually used as the objective function to optimize the DNN-based speech enhancement. However, the MMSE-based objective function is not closely related to ASR performance and thus cannot guarantee good ASR results [31]. Formulating consistent training objectives that meet specific evaluation criteria has always been a challenging task for speech enhancement [37]. Clearly, WER should be the optimal objective function to guide optimization of speech enhancement towards improving noise robustness in ASR. Nonetheless, it is difficult to directly use WER to optimize the enhancement models because there are multiple complex ASR modules, such as feature exaction, acoustic and language probability calculation, that are usually not easily differentiable to be directly incorporated into the enhancement objective functions [31]. The commonly used MMSE-based objective function tends to generate over-smoothed spectra that fails to directly optimize

towards minimizing ASR WERs [38]. To optimize the speech enhancement model by using an ASR-oriented criterion for noise robust ASR, many techniques have been proposed to jointly train a single DNN for both speech enhancement and acoustic modeling [14], [39], [40]. However, in many real-world applications, the ASR system could be supplied by a third party, and speech enhancement is used to generate suitable and discriminative inputs to the ASR system. Accordingly, advanced objective functions for DNN-based speech enhancement for robust ASR without retraining or joint-training are of great interest and worth exploring.

In this article, we propose a cross-entropy-guided measure (CEGM) to assess ASR performance and optimize speech enhancement. It is defined as the cross entropy of the state posterior densities between enhanced and clean speech from the outputs of the conventional DNN-HMM (hidden Markov model) acoustic model. As shown in the upper dashed box in Fig. 1, unlike WER, the calculation of CEGM does not require transcribed data, language model and ASR decoding. It only needs a parallel corpus of degraded and clean speech utterances and the acoustic model of the target ASR system. Moreover, the acoustic characteristics directly correlated with ASR are incorporated into CEGM via the introduction of high-level representations, i.e., the state posteriors. This makes CEGM closely related to ASR performance. Experiments with single-channel track data of the CHiME-4 Challenge [41] demonstrate that CEGM yields a consistently highest degree of correlation with WER comparing to STOI, PESQ and SDR from various aspects, including acoustic models, language models, speech enhancement algorithms, signal-to-noise-ratio (SNR) levels and noise types. Consequently, CEGM provides a more accurate guideline at the time of choosing a suitable speech enhancement algorithm as a mean to introduce robustness into the recognizer. More importantly, CEGM is differentiable and thus can be easily used to replace the conventional MMSE criterion as the objective function for DNN-based speech enhancement for improving ASR noise robustness. Automatic differentiation [42] can be applied to optimize the front-end DNN model, which makes training easy. CEGM is calculated on high-level representations, i.e., state posterior probabilities from the outputs of the DNN-HMM acoustic model which allows CEGM to take account of useful acoustic knowledge. Thus when CEGM is used to guide optimization of the speech enhancement for improving the ASR accuracy, it allows linguistic information from the acoustic model that is critical for state discrimination to be back-propagated to the front-end enhancement model and the enhancement model can be informed by the acoustic model to provide discriminative input features to the back-end ASR system. Experiments demonstrate that CEGM-optimized speech enhancement can effectively improve the ASR accuracy on the CHiME-4 real test set when compared to those obtained with both original noisy speech and MMSE-enhanced speech for fixed back-end ASR systems with different deep structures trained using multi-condition data.

This article is extended from our earlier work [43] with the following new contributions. First, to make a more comprehensive correlation comparison, SDR is included as one of the competing evaluation metrics in addition to PESQ and STOI. Second, we
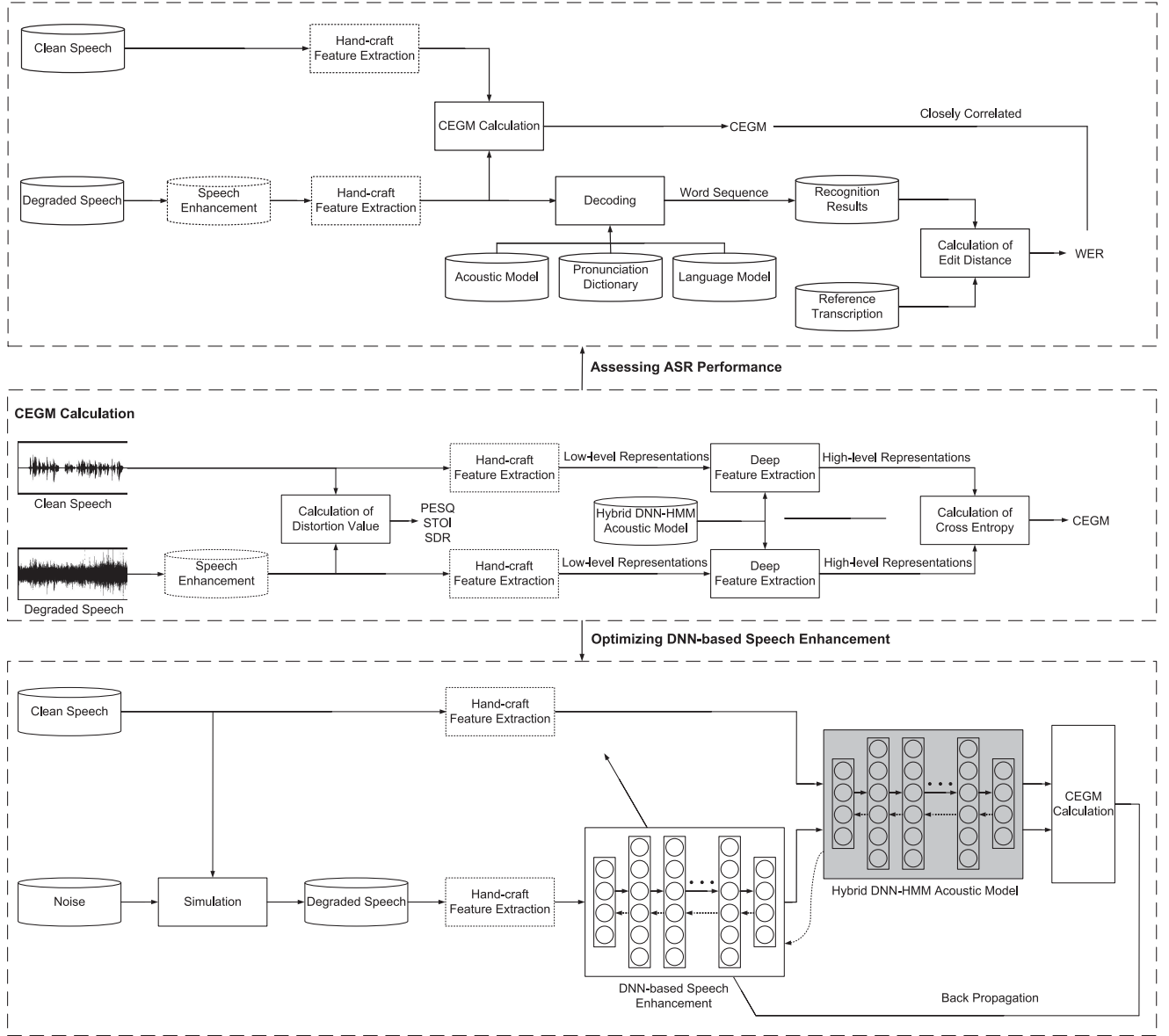
Fig. 1. The overall framework of CEGM for assessing ASR performance and optimizing the model parameters of DNN-based speech enhancement.

provide a simple explanation on every comparative evaluation metric. Third, we propose that CEGM can be directly used to guide estimation of the DNN-based speech enhancement models for robust ASR.

The rest of the paper is organized as follows. CEGM is presented in Section II. Using CEGM for assessing ASR performance is described in Sections III. The optimization process of the front-end enhancement model based on the CEGM objective function is presented in Section IV. New experiments beyond those in [43] are also designed with detailed result analysis in Section V. Finally, we conclude our findings in Section VI.

II. CROSS-ENTROPY-GUIDED MEASURE (CEGM)

A. Background

In the context of human auditory perception, the aim of speech enhancement can be defined as to achieve a higher perceived similarity between the enhanced signal and the clean signal than between the unprocessed and the clean signal. In the past few decades, many objective perceptual quality assessment measures for speech enhancement have been proposed to reduce time-consuming and cost-intensive subjective listening test. They measure the trade-off between the noise reduction and the speech distortion in the enhanced signal via a parallel corpus of enhanced-clean speech utterances. In the context of ASR, the aim of speech enhancement is to convert an observed speech signal to a set of input features of the ASR system that are insensitive to environmental distortion while simultaneously containing a sufficient amount of discriminant information [44].

Given that features extracted from clean speech signals contain much more discriminant information than those of corrupted speech, the aim of speech enhancement can also be defined as to achieve a higher discriminant information similarity contained in the enhanced features and the clean features than in the

unprocessed features and the clean features. In this study, we propose a CEGM for assessing ASR performance which compares the amount of the discriminant information contained in the degraded features and clean features. Its calculation does not require reference transcription and ASR decoding. This greatly reduces calculation costs and time consumption when compared to WER. Note that the reference clean speech is needed in both our proposed ASR performance assessment measure and objective perceptual quality assessment measures. The original clean speech is usually available since an assumption is made in the process of research and development of speech enhancement algorithms that the noisy speech is generated by recording the noises in different environments and artificially adding them to the clean speech. This assumption is reasonable from the viewpoint of reducing the recording cost.

B. Definition of CEGM

Clean speech usually enjoys a good ASR performance as it contains sufficient discriminant information, which indicates that the outputs of the acoustic model ("soft targets") can be considered as the ground truth. This motivates us to propose an evaluation measure for assessing ASR performance by comparing the amount of the discriminant information contained in the degraded features and clean features via the outputs of the acoustic model. The middle dashed box in Fig. 1 shows the calculation framework of our proposed CEGM. Generally, CEGM is a function of high-level representations of clean and enhanced speech computed by a nonlinear operations between the acoustic model and the inputs to the ASR system being tested. This process mainly includes three steps. The first step is the extraction of low-level representations. They correspond to the handcrafted features that can be the raw time signals, mel-frequency cepstral coefficients (MFCCs), log-mel-filterbank (FBANK) features or feature-space maximum likelihood linear regression (fMLLR) [45]. The low-level representations are not directly correlated with speech recognition. The amount of the discriminant information cannot be well reflected by the low-level representations. To incorporate acoustic information from the ASR back-end into the calculation of CEGM, we utilize the acoustic model to map the low-level representations to high-level representations in the second step. This allows CEGM to take account of useful acoustic knowledge for better assessment of ASR performance. The state posterior probabilities from the outputs of the hybrid DNN-HMM acoustic model are used as the high-level representations in this study, which are generated by deep feature extraction from the low-level representations. Therefore, CEGM currently aims to work with the conventional hybrid DNN-HMM ASR system which is also one of the main streams. In the future, we will explore high-level representations for other types of ASR system, e.g., the acoustic models with end-to-end optimization [46]. The high-level representations, i.e., the state posterior probabilities reflect the amount of the discriminant information, where the larger amount of the discriminant information is contained, the closer the state posterior probabilities approach the ground truth, i.e., forced alignment state labels. The last step is the calculation of CEGM which

compares the amount of the discriminant information contained in the clean and enhanced speech by measuring the difference between the high-level representations of them via a criterion, e.g. cross entropy, Kullback-Leibler divergence and MSE. Motivated by the training criterion of hybrid DNN-HMM acoustic models [1], cross entropy is adopted here. Accordingly, CEGM is defined as follows:

$$\text{CEGM} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I} p(s_i|\boldsymbol{x}_n^{\text{C}}, \boldsymbol{W}) \log p(s_i|\boldsymbol{x}_n^{\text{E}}, \boldsymbol{W}) \quad (1)$$

where $p(s_i|\boldsymbol{x}_n^{\text{C}}, \boldsymbol{W})$ and $p(s_i|\boldsymbol{x}_n^{\text{E}}, \boldsymbol{W})$ are the state posteriors of clean speech and enhanced speech from the outputs of the hybrid DNN-HMM acoustic model, respectively, I is the number of output classes or senones, N is the number of frames, $\boldsymbol{x}_n^{\text{C}}$ and $\boldsymbol{x}_n^{\text{E}}$ are the input vectors from the clean speech and enhanced speech for the n-th frame, respectively and $\boldsymbol{W}$ is the parameter set of the DNN-HMM acoustic model of the ASR system.

III. CEGM FOR ASSESSING ASR PERFORMANCE

A. Competing Measures

In this article, correlations between WER and evaluations metrics, such as CEGM, PESQ, STOI and SDR, are compared and contrasted. A detailed explanation of these commonly used distortion measures is given below.

1) PESQ: PESQ is a quality metric recommended by the International Telecommunication Union. It applies an auditory transform to produce a loudness spectrum, and compares the loudness spectra of clean and enhanced speech to generate a score. It ranges between 1 and 4.5 with higher values corresponding to better quality. It has a high correlation with the perceptual mean opinion score [47]. More detailed descriptions of PESQ can be found in [25].

2) STOI: STOI is a function of a time-frequency-dependent intermediate intelligibility measure, which compares the temporal envelopes of clean and enhanced speech in short-time regions by means of a correlation coefficient. It has been found to be highly correlated with intelligibility as measured in human listening tests. The STOI value typically ranges from 0 to 1, which can be interpreted as percent correct. Higher value of STOI represents better speech intelligibility. More detailed descriptions of STOI can be found in [28].

3) SDR: SDR is inspired by SNR and defined as a ratio of the power of clean speech and the power of the difference between clean and enhanced speech. It is calculated as follows,

$$\text{SDR} = 10\log_{10}\frac{||\boldsymbol{s}_{\text{clean}}||^2}{||\hat{\boldsymbol{s}} - \boldsymbol{s}_{\text{clean}}||^2}. \quad (2)$$

SDR shows SNR via decomposing the enhanced signal into the clean plus the residual error parts and taking ratio of the two. More detailed descriptions of SDR can be found in [26].

These distortion metrics are calculated using low-level features as shown in the left part of the middle dashed box in Fig. 1, which clearly do not include acoustic and linguistic information from the ASR back-end and thus are not closely correlated to the ASR accuracies. In contrast, CEGM incorporates information

from the acoustic model into its calculation via introducing high-level representations and thus is closely correlated with the ASR error rates. Although some extra information from ASR acoustic models is needed to calculate CEGM, it could usually be made available since it is a common practice to design front-end speech enhancement for an existing ASR back-end.

B. Correlation Evaluation Procedure

There are three kinds of commonly used correlation coefficients, namely the Pearson correlation coefficient (PCC), the Spearman rank correlation coefficient and the Kendall rank correlation coefficient, while the PCC is the most commonly used one [48] which represents a linear correlation between two data sets and can be calculated as follows:

$$\rho_{xy} = \frac{\sum_{n=1}^{N}(x_n - \overline{x})(y_n - \overline{y})}{\sqrt{\sum_{n=1}^{N}(x_n - \overline{x})^2}\sqrt{\sum_{n=1}^{N}(y_n - \overline{y})^2}}. \qquad (3)$$

This equation can be considered as an expression of a ratio of how much the two data sets $\boldsymbol{x} = [x_1 x_2 \ldots x_N]^\top$ and $\boldsymbol{y} = [y_1 y_2 \ldots y_N]^\top$ vary together as oppose to how much they vary separately [48]. The magnitude of the correlation coefficient indicates the strength of the correlation and the sign indicates if the correlation is positive or negative.

We would like to establish a monotonic relation between the evaluation metrics (CEGM, PESQ, STOI, SDR) and WER. Accordingly, a mapping is used to account for the nonlinear relationship in order to linearize the data such that we can use the PCC to evaluate the correlation. Motivated by [22], [28], [29], [49], a logistic function is used here:

$$f(m) = \frac{100}{1 + \exp(am + b)} \qquad (4)$$

where a and b are constants to be determined by data-fitting using a least-squares method, with m representing the evaluation metric score (CEGM, PESQ, STOI, SDR) and $f(m)$ could be considered as an estimator of the WER which is between 0 and 100. Note that a logistic function is also monotonic and will therefore not influence the ranking of the mapping. Then the performance of the evaluation metric is represented by means of the PCC, which is applied to the mapped objective scores, i.e., $f(m)$. The WER, CEGM, PESQ, STOI and SDR were all computed per utterance. All the correlation coefficients in this study were computed by the aforementioned procedure. Since we are interested in the strength of the correlation, only the magnitudes of the PCC are shown in the experimental results.

IV. CEGM for Optimizing DNN-Based Speech Enhancement

MMSE is the most commonly used objective function for DNN-based speech enhancement [34]. However, it is not directly correlated with either the perceptual quality or ASR performance so that good results for the desired targets could not be guaranteed. In the context of DNN-based enhancement for robust ASR, in order to utilize the back-end information to guide enhancement model optimization towards improving the ASR accuracies, various strategies, such as joint training [14], [39],

[40] or a combination of teacher-student learning or adversarial training [50]–[53], have been proposed. However, these techniques are not applicable to the scenario that a well-trained ASR system is provided by a third party, and speech enhancement is used to generate suitable inputs to the ASR system. More recently, reinforcement learning was utilized as an ASR-oriented objective [31] for DNN-based enhancement to improve ASR noise robustness.

Our proposed CEGM is differentiable and thus can be easily used to replace the conventional MMSE criterion as the objective function to optimize DNN-based speech enhancement. CEGM considers the acoustic information by the introduction of high-level representations and thus it is directly correlated with ASR system. Clearly, CEGM allows the speech enhancement front-end to take into account the way in which the enhanced features are processed by the back-end recognizer. Automatic differentiation can then be applied to optimize the front-end DNN model. Accordingly, the CEGM-optimized speech enhancement model is easier to train as compared to reinforcement learning based speech enhancement in [31]. Moreover, CEGM makes it possible to customize speech enhancement models for different ASR systems by incorporating the corresponding acoustic model into the metric calculation.

The framework of CEGM for guiding the front-end DNN-based speech enhancement is shown in the lower dashed box in Fig. 1. The model is trained with gradient descent by performing backpropagation with the CEGM criterion defined as the cross entropy of the state posteriors between clean and enhanced speech from the outputs of the conventional DNN-HMM acoustic model. Notably, the parameters of the back-end acoustic model are used in the forward propagation stage to calculate CEGM and fixed during backpropagation. To avoid extra middle-stage post-processing and dynamic feature calculation operations, front-end enhancement is designed to directly map the input noisy features to the desired input features of the back-end acoustic model. This seamless connection between the front-end and back-end models simplifies DNN training.

A regression DNN model is used as the speech enhancement model to directly map the noisy features to the desired discriminative input features to the ASR system. The acoustic model of the ASR system works as a fixed nonlinear function mapping the paired enhanced and clean acoustic features to the state posteriors. CEGM is obtained by calculating the cross entropy between the paired state probabilities and then guides the speech enhancement model optimization process which is an error backpropagation algorithm. In the feedforward computation, the speech enhancement model maps the input noisy feature vector $\boldsymbol{y}_n$ to the desired acoustic feature vector $\boldsymbol{x}_n^{\mathrm{E}}$ fed to the acoustic model at time frame n as follows:

$$\boldsymbol{x}_n^{\mathrm{E}} = \mathcal{F}(\boldsymbol{y}_n; \boldsymbol{W}_{\mathrm{SE}}), \qquad (5)$$

where $\mathcal{F}$ denotes the regression DNN and $\boldsymbol{W}_{\mathrm{SE}}$ is the corresponding parameter set. Then, the paired clean and enhanced feature vectors, $\boldsymbol{x}_n^{\mathrm{C}}$ and $\boldsymbol{x}_n^{\mathrm{E}}$, are fed into the DNN-HMM acoustic model of the target ASR system to output the paired state posteriors of the i-th context-dependent HMM state s_i generated by the nonlinear function $\mathcal{G}_i$ at the output layer of the acoustic

model as described as follows,

$$p(s_i|\boldsymbol{x}_n^{\mathrm{E}}, \boldsymbol{W}_{\mathrm{AM}}) = \mathcal{G}_i(\boldsymbol{x}_n^{\mathrm{E}}; \boldsymbol{W}_{\mathrm{AM}});$$
$$p(s_i|\boldsymbol{x}_n^{\mathrm{C}}, \boldsymbol{W}_{\mathrm{AM}}) = \mathcal{G}_i(\boldsymbol{x}_n^{\mathrm{C}}; \boldsymbol{W}_{\mathrm{AM}}), \tag{6}$$

where $\boldsymbol{W}_{\mathrm{AM}}$ denotes the parameter set of the acoustic model. Finally, similar to Eq. (1), the CEGM objective is calculated via the paired state posteriorss as follows:

$$\mathrm{CEGM} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I}\mathcal{G}_i(\boldsymbol{x}_n^{\mathrm{C}}; \boldsymbol{W}_{\mathrm{AM}})\log\mathcal{G}_i(\boldsymbol{x}_n^{\mathrm{E}}; \boldsymbol{W}_{\mathrm{AM}}), \tag{7}$$

where I is the number of the HMM states and N is the number of frames. After the feedforward computation, error backpropagation with stochastic gradient learning is developed to estimate the enhancement model parameter set $\boldsymbol{W}_{\mathrm{SE}}$ by setting the acoustic model parameter set $\boldsymbol{W}_{\mathrm{AM}}$ fixed. Therefore, $\boldsymbol{W}_{\mathrm{SE}}$ can be updated with gradient descent as follows,

$$\boldsymbol{W}_{\mathrm{SE}} \leftarrow \boldsymbol{W}_{\mathrm{SE}} - \alpha\frac{\partial\mathrm{CEGM}}{\partial\boldsymbol{W}_{\mathrm{SE}}}, \tag{8}$$

where α is the learning rate. By applying the chain rule, the gradient of the objective function CEGM with respect to the speech enhancement model parameter set can be obtained

$$\frac{\partial\mathrm{CEGM}}{\partial\boldsymbol{W}_{\mathrm{SE}}} =$$
$$-\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{I}\frac{\mathcal{G}_i(\boldsymbol{x}_n^{\mathrm{C}}; \boldsymbol{W}_{\mathrm{AM}})}{\mathcal{G}_i(\boldsymbol{x}_n^{\mathrm{E}}; \boldsymbol{W}_{\mathrm{AM}})}\frac{\partial\mathcal{G}_i(\boldsymbol{x}_n^{\mathrm{E}}; \boldsymbol{W}_{\mathrm{AM}})}{\partial\boldsymbol{x}_n^{\mathrm{E}}}\frac{\partial\mathcal{F}(\boldsymbol{y}_n; \boldsymbol{W}_{\mathrm{SE}})}{\partial\boldsymbol{W}_{\mathrm{SE}}} \tag{9}$$

V. EXPERIMENTS

A. Database Description

Experiments were conducted on the CHiME-4 speech separation and recognition challenge [41], which targets distant-talking ASR based on the speaker-independent 5K-word subset of the Wall Street Journal (WSJ0) tasks [54]. Real and simulation data are given for each of the training, development and test sets. The real data was recorded by a 6-channel tablet-based microphone array spoken by talkers situated in four recording locations, including cafe (CAF), public transport (BUS), pedestrian area (PED) and street junction (STR). The simulation data was constructed by mixing clean utterances from the WSJ0 corpus with the abovementioned four noise conditions. There are three test scenarios, i.e., single-channel (1-channel) and multi-channel (2-channel and 6-channel) tasks. In this study, the single-channel task is selected to evaluate our proposed CEGM framework.

Because the calculations of the comparative distortion measures (i.e., PESQ, STOI and SDR) and our proposed CEGM needs time-synchronized speech pairs, we evaluated their correlations with WER by the simulated data. The correlations were investigated in five situations, with varying acoustic models, language models, speech enhancement algorithms as a pre-processing stage of ASR, noise types, and SNRs. Experiments of the first four situations were conducted on the simulated data from the official development and test sets consisting of 1640 and 1320 utterances respectively. Experiments of the last
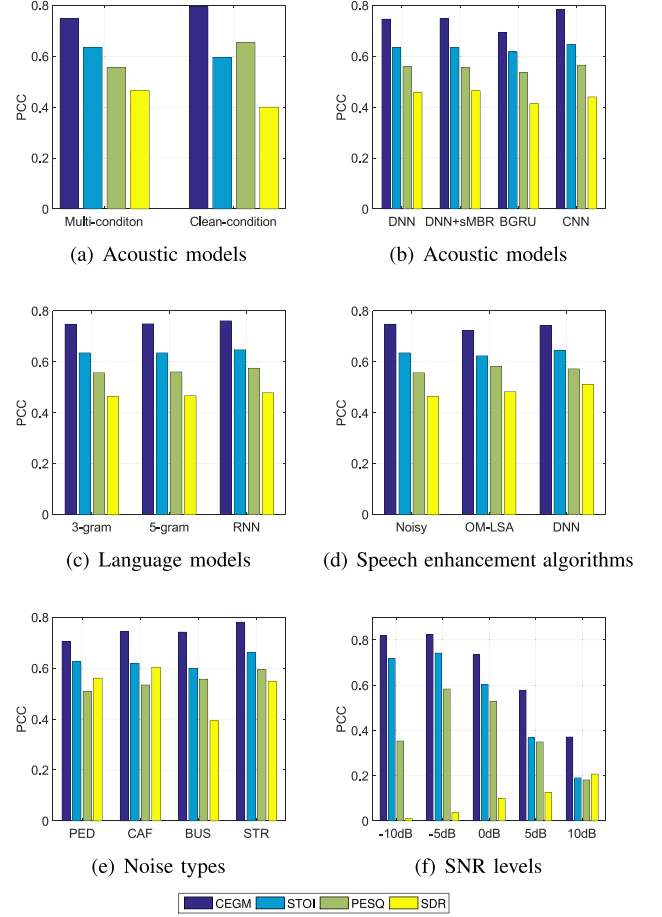

Fig. 2. The PCC comparison in various situations.

situation were conducted on the simulated data consisting of 2960 utterances at respective SNR level constructed by mixing the clean speech data from the development and test sets of WSJ0 corpus with the four types of background noises from all six channels at five levels of SNRs at −10 dB, −5 dB, 0 dB, 5 dB, and 10 dB, respectively.

Here, the output layers of all the acoustic models have 1987 shared state output units (i.e., $I = 1987$). The tied HMM states were generated by forced-alignment via a GMM (Gaussian mixture model)-HMM system [1].

B. Correlation Comparisons in Different Situations

The correlations of WER with our proposed CEGM and the competing measures are explored from five aspects, including acoustic models, language models, speech enhancement algorithms, noise types and SNR levels. The results are summarized in Fig. 2.

1) Acoustic Models: Concerning the acoustic models, the correlations with WER are compared based on different training modes, input features, DNN structures and optimization criteria. First, we evaluate the correlations in two training modes, namely clean-condition training and multi-condition training. The adopted acoustic model in the multi-condition training mode was a feed-forward DNN, discriminatively trained with a state-level minimum Bayes risk (sMBR) criterion [55] following

the cross entropy criterion provided by the official CHiME-4 baseline [41]. Its training set consists of 1600 real utterances recorded in the abovementioned four noisy environments from four speakers and 7138 simulated noisy utterances based on the clean utterances in the WSJ0 SI-84 training set. The input of the acoustic model was a 440-dimensional feature vector consisting of a 40-dimensional fMLLR with an 11-frame expansion. The DNN had 7 layers and each layer had 2048 neurons with sigmoid activation functions between the linear transformation layers. The only difference between the adopted clean-condition trained acoustic model and the abovementioned multi-condition trained acoustic model is the training data. 7138 clean utterances from the WSJ0 SI-84 training set were used for training the clean-condition acoustic model. From Fig. 2(a), it is observed that CEGM had the highest degree of correlation with WER for both clean-condition and multi-condition training. Moreover, the correlation of either STOI or PESQ with WER was stronger than that of SDR. More interestingly, the correlation scores between STOI and WER tended to be larger than those between PESQ and WER for multi-condition training and a contrary conclusion could be drawn for clean-condition training. This indicates that the WER depends more largely on speech quality for clean-condition training and on speech intelligibility for multi-condition training. Our observations somehow differs from those in [29]–[31] claiming that the correlation between STOI and WER is stronger than that between PESQ and WER. In the following experiments, the multi-condition training mode is adopted as a default.

Next, we make a simple correlation comparison for the sMBR and cross entropy criteria. As clearly shown in Fig. 2(b), CEGM had the highest PCC scores for the two optimization criteria, implying its robustness towards the acoustic model optimization criterion.

Finally, we compare the correlation in different acoustic model structures. In addition to the official acoustic models, including "DNN" and "DNN+sMBR," acoustic models based on the bidirectional gated recurrent unit (BGRU) [56] and deep convolutional neural network (CNN) [57] are presented. The BGRU had 3 hidden layers with 512 cells in each hidden layer. A VGG16 network [58] was adopted as the deep structure of the CNN. All the convolutional layers used a kernel size of 3×3. There were 2 Conv-64 layers, 2 Conv-128 layers, 3 Conv-256 layers, 3 Conv-512 layers, 3 Conv-512 layers and 3 fully connected layers with 2048 neurons per layer. Pooling [59] was utilized on the frequency dimension and performed after consecutive convolutional layers using 1×2 max pooling. Batch normalization was applied to the input of each rectified linear unit based hidden activation function. The two acoustic models shared the same training data as the official CHiME-4 baseline. A 40-dimensional FBANK feature vector with no frame expansion instead of the fMLLR vector was used as their input. Both of them were trained using the cross entropy criterion. Fig. 2(b) shows a consistently highest degree of correlation between CEGM and WER compared with other distortion measures for different deep structures of acoustic models. Please note that the acoustic model "DNN+sMBR" is used in the following correlation comparison experiments.

TABLE I
THE AVERAGE EVALUATION METRIC SCORES AND WERs OF DIFFERENT SPEECH ENHANCEMENT ALGORITHMS FOR ROBUST ASR ON THE OFFICIAL SIMULATED DEVELOPMENT AND TEST SETS (2960 UTTERANCES) FOR MULTI-CONDITION TRAINING

	Noisy	OM-LSA	DNN
WER(%)	19.39	25.70	24.46
CEGM	4.23	4.45	4.39
STOI	0.82	0.81	0.85
PESQ	2.00	2.25	2.30
SDR	4.44	8.71	6.84

2) Language Models: The language model used in the above experiments was the official 3-gram model provided by Kaldi [45]. The official CHiME-4 baseline also provides 5-gram and recurrent neural network (RNN)-based language models. In Fig. 2(c), we compare the correlations for three language models. The same was observed that CEGM achieves the highest correlation with WER, which is also robust towards different language models. The 3-gram language model is adopted by default in all the following experiments.

3) Speech Enhancement Algorithms: ASR performance varies with different speech enhancement front-ends. Therefore, we investigate the correlations with WER by using two representative noise reduction algorithms (i.e., an OM-LSA speech estimator [6] and a masking-based DNN enhancement model [36]) in addition to the reference case of unprocessed noisy speech. From Fig. 2(d), we can observe that CEGM was most closely correlated with WER using either enhancement algorithm.

4) Noise Types: We also present some detailed correlation comparisons in the context of different noise conditions as shown in Fig. 2(e). Consistent highest PCCs were observed for the proposed CEGM under all four noise types, although the results were mixed for other three distortion measures (STOI, PESQ and SDR).

5) SNR Levels: Correlation comparisons were also conducted under different SNR levels. Clearly, we can observe that CEGM still had the strongest correlation with WER under each SNR level from Fig. 2(f). Notably, the correlation became weak in very high SNR levels where most WER scores were 0% due to the negative influence of these points which had the same WER score but different evaluation metric scores. Similarly, the same phenomenon also occurs at extremely low SNR levels.

C. Evaluation of CEGM for Assessing ASR Performance

Table I shows the average evaluation metric scores and WERs of unprocessed noisy and enhanced speech processed by the OM-LSA and masking-based DNN methods, respectively, on the official simulated development and test sets. Several observations can be made. The variation tendency of the CEGM scores with respect to the speech enhancement algorithms was completely consistent with that of the WERs, where the smaller the CEGM was, the lower the WER was. This implies that CEGM can well assess the ASR WERs. In contrast, other measures could not accurately evaluate the ASR performance of different speech enhancement algorithms. For example, better

STOI, PESQ or SDR of enhanced speech processed by either OM-LSA or DNN-based enhancement did not bring reductions to ASR error rates. Furthermore, high SDR of OM-LSA did not lead to a decline in WER when compared with that of DNN-based speech enhancement.

Clearly, it is not accurate and reliable enough to apply conventional distortion measures to assess the ASR performance. Motivated by the conclusion in [29], [30] that STOI has a stronger correlation with WER when compared with other distortion measures, researchers [37], [60] have designed a speech enhancement front-end to specifically for improving STOI in order to achieve better ASR performance. Accordingly based on the abovementioned experimental results, CEGM seems to be a better metric than STOI to optimize the parameters of enhancement models. It is thus adopted for the comparison with the conventional MMSE in the subsequent experiments.

D. Evaluation of CEGM for Optimizing Enhancement Models

1) Size of the Training Dataset: Training of DNN-based speech enhancement model only requires a parallel corpus comprising pairs of clean and noisy speech samples without the need of any information about the content of speech, i.e., no need with transcription. The training set can be easily obtained by artificially adding recording noises in different noisy environments to clean speech at various SNR levels. Therefore, we could explore how much the WER can be further reduced by gradually incorporating larger number of utterances into the training set to build the speech enhancement model. The officially provided feed-forward DNN acoustic model trained using the cross entropy criterion was adopted to evaluate the effect of the training data size on the ASR performance of DNN-based speech enhancement. Three sets of training data were compared. The first set consisted of 7138 utterances (about 12 hours) from the simulated data of officially provided DNN acoustic model training set. The second one consisted of 7138×3 utterances (about 36 hours) constructed by mixing the clean speech data from the WSJ0 SI-84 training set with the four types of background noise at the SNR level of -5 dB, 0 dB and 5 dB respectively. The last one consisted of 7138×5 utterances (about 60 hours) constructed by mixing the clean speech data from the WSJ0 SI-84 training set with the four types of background noise at the SNR level of -10 dB, -5 dB, 0 dB, 5 dB and 10 dB respectively. They were denoted as "TS1," "TS2," and "TS3," respectively. A BGRU model with 3 hidden layers and 512 units per layer was used for speech enhancement, which directly mapped the noisy fMLLR features with an 11-frame expansion to the clean version fed to the acoustic model. It was trained using the MMSE criterion and the proposed CEGM, respectively. The recognition performance comparison of the two optimization criteria and the variation tendency of the WER with respect to the training data size are evaluated on the real test set across the four environments as shown in Fig. 3. Obviously the CEGM criterion achieved significant WER reductions when compared to both the MMSE criterion and the unprocessed case. As expected, the WER decreased with the increasing amount of training data. Specifically, including a large SNR range in the training set for the DNN-based speech enhancement
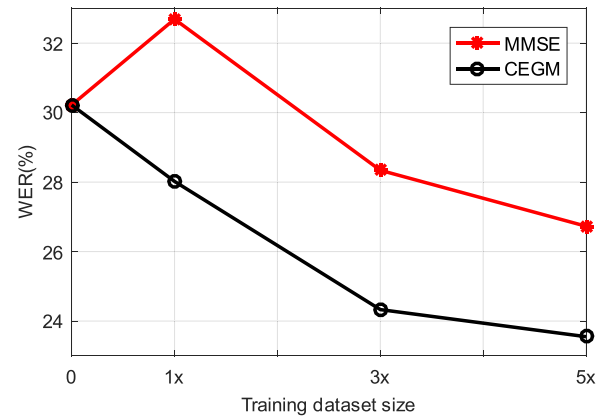

Fig. 3. Average WER (%) comparison between MMSE and CEGM for optimizing enhancement models of recognition system using different training data sizes on the real test set across the four environments. The point at "0" corresponds to the recognition result of unprocessed noisy speech. "1×," "3×" and "5×" represent the settings of "TS1," "TS2" and "TS3," respectively.

TABLE II
WER (%) COMPARISON BETWEEN MMSE AND CEGM FOR OPTIMIZING ENHANCEMENT MODELS USING DIFFERENT ACOUSTIC MODELS OF RECOGNITION SYSTEM ON THE REAL TEST SET

Acoustic Model	Criterion	BUS	CAF	PED	STR	AVG
DNN	Noisy	43.67	32.91	25.93	18.36	30.22
	MMSE	34.25	30.43	24.57	17.67	26.73
	CEGM	29.82	27.25	21.51	15.61	23.55
DNN+sMBR	Noisy	40.36	28.97	24.27	17.07	27.67
	MMSE	35.05	29.87	24.70	17.59	26.80
	CEGM	29.69	25.92	20.37	14.70	22.67
BGRU	Noisy	34.96	29.51	23.54	17.86	26.47
	MMSE	39.45	36.07	28.85	20.34	31.17
	CEGM	30.10	27.72	22.27	16.27	24.09
CNN	Noisy	33.95	25.68	20.72	14.87	23.80
	MMSE	41.76	37.95	29.56	21.37	32.66
	CEGM	31.99	26.02	20.78	15.00	23.44

could reduce the mismatch between the training and testing and thus improve the recognition performance on the real test set. Moreover, the performance gap between MMSE and CEGM was still remarkable using a larger training set, e.g., yielding a relative WER reduction of more than 10% from MMSE to CEGM in the "TS3" case. In the following experiments, 'only 'TS3" setting is adopted.

2) Acoustic Model Settings: CEGM provides an ASR-level optimization criterion for DNN-based speech enhancement by incorporating the acoustic model of the target ASR system. So it is possible to customize a speech enhancement front-end for a particular ASR back-end. We explore the effectiveness of DNN-based speech enhancement optimized by the CEGM criterion with various acoustic models using different input features (FBANK and fMLLR), optimization criteria (cross entropy and sMBR), and deep structures (feed-forward DNN, BGRU and deep CNN). The four acoustic models ("DNN," "DNN+sMBR," "BGRU" and "CNN") used in Section V-B1 were adopted here. The WER comparison between MMSE and CEGM for optimizing enhancement models to improve noise robustness of ASR systems with different acoustic models on the real test set is shown in Table II. Note that all the enhancement models employed a BGRU structure with 3 hidden layers and 512 units per layer. First, the CEGM criterion

TABLE III
AVERAGE WER (%) COMPARISON BETWEEN MMSE AND CEGM FOR
OPTIMIZING DIFFERENT ENHANCEMENT MODELS USING TWO ACOUSTIC
MODELS ON THE REAL TEST SET ACROSS THE FOUR ENVIRONMENTS

| Acoustic Model | Noisy | Speech Enhancement Model | | | | | |
| | | BGRU | | U-net | | CRNN | |
		MMSE	CEGM	MMSE	CEGM	MMSE	CEGM
BGRU	26.47	31.17	24.09	28.25	20.87	27.85	20.51
CNN	23.80	32.66	23.44	29.02	20.96	28.05	20.23

achieved consistent and significant WER reductions compared to the MMSE criterion for all noise types and acoustic models, yielding relative WER reductions of 11.9%, 15.4%, 22.7% and 28.2%, repectively, for the four acoustic models on average of all noise types, respectively. And the performance gains became larger with more powerful acoustic models of ASR system. Second, in comparison to the unprocessed system ("Noisy"), the enhancement models optimized using the CEGM criterion achieved relative WER reductions of 22.1%, 18.1%, 9.0% and 1.5%, on average for the four acoustic models, respectively. This implied that it was more challenging for enhancement models to improve recognition performance if the back-end acoustic model became more powerful. Third, the enhancement models optimized with the MMSE criterion could not guarantee good ASR results, e.g., leading to performance degradation for the acoustic models "BGRU" and "CNN". In contrast, the CEGM-optimized enhancement models usually achieved lower WERs.

3) Speech Enhancement Model Settings: It is clearly observed in Table II that for the powerful "CNN"-based acoustic model, the BGRU enhancement model optimized using the CEGM criterion can only reduce the average WER from 23.80% to 23.44% compared with unprocessed speech. This motivates us to design advanced deep structures for the front-end enhancement model. Intuitively, the enhancement model should be designed as powerful as the back-end acoustic model. Here, we select the acoustic models "BGRU" and "CNN" to explore the design of deep structures for the enhancement models and expect to further improve the ASR performance. Table III lists the WER comparison of the speech enhancement models on the real test set with different deep structures trained using the MMSE and CEGM criteria, respectively. In addition to the BGRU enhancement models, two more powerful deep structures denoted as "U-net" and "CRNN" were adopted. "U-net" was a deep fully convolutional network [61] consisting of a downsampling path and an upsampling path. The downsampling path had 5 convolutional blocks. Each block consisted of two convolutional layers with a filter size of 3×3 and stride of 1 in both directions, followed by a batch normalization and rectifier activation, which increased the number of feature maps from 1 to 512. For downsampling, max pooling with stride 2×2 was applied to the end of each block except the last block. The upsampling path had 4 upsampling, each block started with a deconvolutional layer with a filter size of 2×2 and stride of 2×2, which doubled the size of feature maps in both directions, followed by two convolutional layers which reduced the number of feature maps for the concatenation of deconvolutional feature maps and the feature maps from the downsampling path. Finally, a 1×1 convolutional layer was used to reduce the number of

feature maps to one. No fully connected layer was invoked in the network. "CRNN" was a combination of convolutional and recurrent neural networks [62]. Specifically, it started with the convolutional layers followed by a BGRU fixed at 3 hidden layers with 512 units in each layer and a fully connected layer. The VGG16 network adopted in the deep acoustic model "CNN" was also used as the convolutional component of "CRNN".

Table III shows that further ASR improvements are achieved by the more powerful front-end enhancement models. For example, for the back-end with the acoustic model "BGRU", the top row shows that CEGM reduces the average WER from 24.09% for enhancement model "BGRU" to 20.87% for model "U-net" while model "CRNN" achieves the best WER of 20.51%. For the back-end with the acoustic model "CNN," similar results were observed. The average WERs of "BGRU"/"U-net"/"CRNN" enhancement models using the CEGM criterion were 23.44%/20.96%/20.23%, demonstrating that "U-net" and "CRNN" were more powerful than "BGRU" as speech enhancement models. Although more powerful enhancement models also resulted in lower WERs for the MMSE criterion, they all failed to improve ASR performance when compared to unprocessed noisy speech for both the acoustic models of "BGRU" and "CNN". In contrast, all CEGM-optimized enhancement models yielded consistent WER reductions in comparison to unprocessed noisy speech. Overall, the enhancement models of "BGRU" "U-net" and "CRNN" using the CEGM criterion achieved relative WER reductions of 9.0%, 21.2% and 22.5%, respectively, for the back-end acoustic model "BGRU" and relative WER reductions of 1.5%, 11.9% and 15.0%, respectively, for the back-end acoustic model "CNN" when compared to those of unprocessed noisy speech. More interestingly, the WER gap between "BGRU" and "CNN" acoustic models for unprocessed noisy speech (26.47% vs. 23.80%) were largely reduced after CEGM-based enhancement (20.51% vs. 20.23% for "CRNN" model). This indicates that it is more flexible to design back-end models with a more powerful CEGM-optimized front-end model.

4) Output Analysis of the Enhancement Models: In Fig. 4, we show an example comparison between the outputs of the "CRNN" enhancement model optimized using the MMSE and CEGM criteria for the acoustic model "BGRU" using an utterance from the real test set of the CHiME-4 Challenge. Speech from Channel 0 (recorded by the close-talking microphone) was used as reference clean speech. Note that the FBANK features were further processed by utterance-level mean normalization. It was clearly observed that the enhanced FBANK features from the CEGM-enhanced output no longer had a complete spectral structure. In contrast, the enhanced FBANK features from the output of the MMSE-optimized enhancement model kept a similar spectral structure as that for clean speech. It is interesting to note that the CEGM-enhanced FBANK features were more discriminative and achieved lower WERs. This shows that, for speech enhancement used as the ASR front-end, restoring the perfect spectral structure of clean speech is not as important as preserving some acoustic cues needed for the ASR back-end.

5) Robustness Analysis: In previous experiments, although the training set for speech enhancement was simulated, the enhancement models optimized using the CEGM criterion
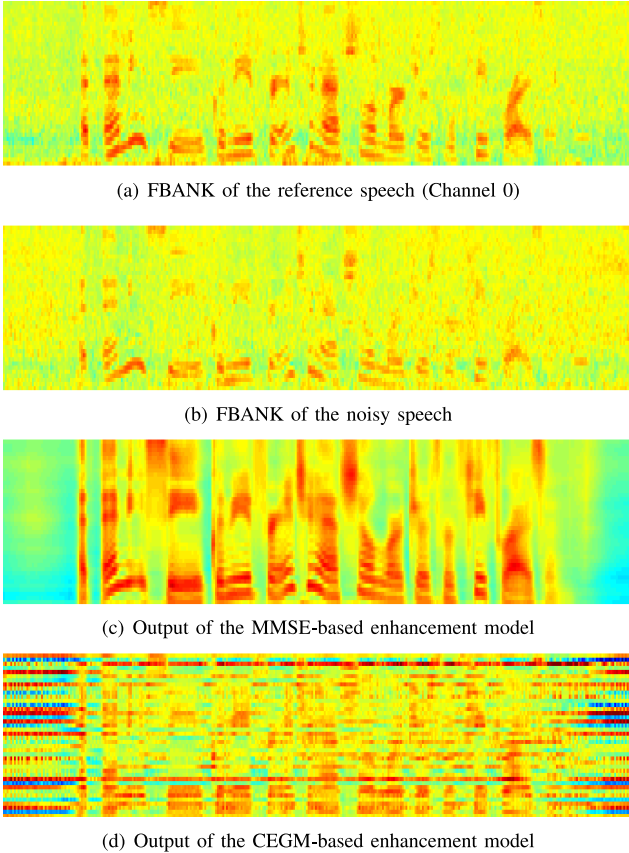
(a) FBANK of the reference speech (Channel 0)

(b) FBANK of the noisy speech

(c) Output of the MMSE-based enhancement model

(d) Output of the CEGM-based enhancement model

Fig. 4. An utterance comparison of the outputs of the "CRNN" enhancement models optimized using MMSE and CEGM criteria.

TABLE IV
AVERAGE RELATIVE WER REDUCTION (%) COMPARED TO UNPROCESSED NOISY SPEECH ON THE REAL TEST SET FOR THE ENHANCEMENT MODELS TRAINED ON THE THREE SIMULATED TRAINING SETS WITH MMSE AND CEGM CRITERIA RESPECTIVELY

	WSJ0+CHiME4noise	TIMIT+CHiME4noise	WSJ0+115noise
MMSE	-5.2	-27.0	-36.9
CEGM	22.5	2.9	-0.6

TABLE V
WER (%) COMPARISON BETWEEN MMSE AND CEGM FOR OPTIMIZING THE "CRNN" ENHANCEMENT MODEL USING THE "CNN" ACOUSTIC MODEL ON THE SIMULATED TEST SET AT DIFFERENT SNRs AVERAGED OVER THE 15 NOISE TYPES

	-6 dB	-3 dB	0 dB	3 dB	6 dB
Noisy	66.60	57.93	43.79	31.19	22.79
MMSE	75.42	67.90	53.16	40.74	31.12
CEGM	62.60	53.91	41.15	29.44	21.54

E. Evaluation of CEGM for Optimizing Enhancement Models on Other Datasets

We also evaluated and reconfirmed the effectiveness of CEGM for optimizing enhancement models with the LibriSpeech corpus [65]. The LibriSpeech is a relatively large corpus containing approximately 1000 hours of read English speech from audio books. We used the training subset composed of 100 hours to mix with 115 kinds of noise used in [63] at five levels of SNRs, i.e., −10 dB, −5 dB, 0 dB, 5 dB and 10 dB, to build a 100-hour multi-condition training data. For evaluation, the 2620 utterances in the "test-clean" set were used to mix with 15 kinds of noise used in [66] at the SNR level of −6 dB, −3 dB, 0 dB, 3 dB and 6 dB respectively. We selected "CNN" as the acoustic model structure and trained it using the 100-hour multi-condition training data. Then the CEGM was calculated with this acoustic model. "CRNN" was used as the enhancement model and trained with the MMSE and CEGM criteria on the 100-hour training data respectively. The WER comparison between MMSE and CEGM for optimizing the enhancement model to improve ASR noise robustness on the simulated test set at different SNRs averaged over the 15 noise types is shown in Table V. The same observarisons were made as in the CHiME-4 corpus. Specifically, the enhancement model optimized with the MMSE criterion could not guarantee good ASR results. Instead, the enhancement model optimized with the CEGM criterion achieved consistent and significant WER reductions compared to that optimized with the MMSE criterion and the unprocessed system ("Noisy") across all the SNR levels.

VI. CONCLUSION

In this article, we propose a CEGM which can be applied to assessing the ASR performance of the degraded speech using only a parallel corpus of degraded and clean speech utterances. Moreover, CEGM is differentiable and thus can be easily utilized to guide the optimization of DNN-based single-channel speech enhancement for improving ASR noise robustness. Results from a series of experiments demonstrate that the proposed CEGM yields consistently highest correlations with WER and achieves the most accurate assessment of ASR performance when compared to the commonly used perceptual evaluation measures

could achieve recognition performance improvements on the real test set. This was partially due to that both the clean speech and noise signals to simulate the training data for speech enhancement were from CHiME-4 challenge, denoted as "WSJ0+CHiME4noise". In this section, we further investigate the robustness of the CEGM-based speech enhancement on two more mismatched simulated training sets. One of the training sets consisted of 7138 × 5 utterances (about 60hours) constructed by mixing the clean speech data from the WSJ0 SI-84 training set with 115 kinds of noise used in [63] at the SNR level of −10 dB, −5 dB, 0 dB, 5 dB and 10 dB respectively. It is denoted as "WSJ0+115noise," which is a noise mismatched version compared with "WSJ0+CHiME4noise". The other one is also about 60 hours built by mixing the clean speech data from the TIMIT [64] training set with the four types of noise of the CHiME-4 challenge at the same five SNR levels. It is denoted as "TIMIT+CHiME4noise," which is a speech mismatched version compared with "WSJ0+CHiME4noise". "CRNN" was adopted as the front-end enhancement model while "BGRU" was used as the back-end acoustic model. As displayed in Table IV, the ASR performance decreased dramatically for both MMSE-based and CEGM-based enhancement as the training sets became more mismatched. Nevertheless, the CEGM-based enhancement could still maintain the comparable performance with the unprocessed noisy speech while MMSE-based enhancement led to severe degradation of recognition accuracy in the mismatched cases.

including STOI, PESQ and SDR. The conventional DNN-based enhancement models optimized using the MMSE criterion tend to lead to performance degradation for ASR systems with multi-condition training. The CEGM is differentiable and thus can be easily used to replace the conventional MMSE criterion to guide the DNN-based enhancement model optimization by automatic differentiation. Experiments show that the CEGM-based speech enhancement not only achieves considerable performance improvements for different multi-condition ASR systems with various acoustic model structures but also shows good generalization capabilities. The output analysis of the enhancement models reveals that restoring the perfect spectral structure of clean speech is not as important as preserving some acoustic characteristics that are crucial to the back-end ASR for speech enhancement in order to improve ASR noise robustness. CEGM currently aims to work with the conventional hybrid DNN-HMM ASR systems. In the future, we will explore evaluation measures for other types of acoustic models.

REFERENCES

[1] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[3] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 3, pp. 197–210, Jun. 1978.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[6] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[7] H.-J. Hsieh, B. Chen, and J.-W. Hung, "Employing median filtering to enhance the complex-valued acoustic spectrograms in modulation domain for noise-robust speech recognition," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process.*, 2016, pp. 1–5.

[8] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2008, pp. 4041–4044.

[9] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7398–7402.

[10] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 5210–5214.

[11] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 196–200.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[13] S.-J. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," *Proc. Interspeech*, pp. 1571–1575, 2018.

[14] T. Menne, R. Schlüter, and H. Ney, "Investigation into joint optimization of single channel speech enhancement and acoustic modeling for robust ASR," in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6660–6664.

[15] A. S. Subramanian, S.-J. Chen, and S. Watanabe, "Student-teacher learning for BLSTM mask-based speech enhancement," *Proc. Interspeech*, pp. 3249–3253, 2018.

[16] D. P. Ellis, "Evaluating speech separation systems," in *Proc. Speech Separation Humans Mach.*, Springer, 2005, pp. 295–304.

[17] U. M. B. Neto and E. R. Dougherty, *Error Estimation for Pattern Recognition*. New York, NY, USA: Wiley, 2015.

[18] A. K. Jain, R. C. Dubes, and C.-C. Chen, "Bootstrap techniques for error estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 5, pp. 628–633, Sep. 1987.

[19] C.-H. Huang, C. S. Lee, and H.-C. Wang, "A study on model-based error rate estimation for automatic speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 581–589, Nov. 2003.

[20] H. Sun, L. Shue, and J. Chen, "Investigations into the relationship between measurable speech quality and speech recognition rate for telephony speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2004, vol. 1, pp. 865–868.

[21] T. Yamada and N. Kitawaki, "A PESQ-based performance prediction method for noisy speech recognition," in *Proc. Int. Congr. Acoust.*, vol. 2, 2004.

[22] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance estimation of speech recognition system under noise conditions using objective quality measures and artificial voice," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2006–2013, Nov. 2006.

[23] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Process.*, vol. 88, no. 10, pp. 2578–2583, 2008.

[24] T. Fukumori, M. Nakayama, T. Nishiura, and Y. Yamashita, "Estimation of speech recognition performance in noisy and reverberant environments using PESQ score and acoustic parameters," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2013, pp. 1–4.

[25] I.-T. Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Rec. ITU-T P. 862, 2001.

[26] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

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

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[29] D. Thomsen and C. E. Andersen, *Speech Enhancement and Noise-Robust Automatic Speech Recognition*. Aalborg, Denmark: Aalborg Univ., 2015.

[30] A. H. Moore, P. P. Parada, and P. A. Naylor, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures," *Comput. Speech Lang.*, vol. 46, pp. 574–584, 2017.

[31] Y.-L. Shen, C.-Y. Huang, S.-S. Wang, Y. Tsao, H.-M. Wang, and T.-S. Chi, "Reinforcement learning based speech enhancement for robust speech recognition," in *Proc. ICASSP 2019-2019 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6750–6754.

[32] K. K. Paliwal, J. G. Lyons, S. So, A. P. Stark, and K. K. Wójcicki, "Comparative evaluation of speech enhancement methods for robust automatic speech recognition," in *Proc. 4th Int. Conf. Signal Process. Commun. Syst.*, 2010, pp. 1–5.

[33] C. H. You, B. Ma, and C. Ni, "Modification on LSA speech enhancement for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5475–5479.

[34] 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.

[35] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[36] 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.

[37] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[38] M. L. Seltzer, "Bridging the gap: Towards a unified framework for hands-free speech recognition using microphone arrays," in *Proc. Hands-Free Speech Commun. Microphone Arrays*, 2008, pp. 104–107.

[39] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4375–4379.

[40] 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.

[41] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.

[42] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: A survey," *J. Mach. Learn. Res.*, vol. 18, no. 153, 2018.

[43] L. Chai, J. Du, and C.-H. Lee, "A cross-entropy-guided (CEG) measure for speech enhancement front-end assessing performances of back-end automatic speech recognition," *Proc. Interspeech*, pp. 3431–3435, 2019.

[44] T. Yoshioka and M. J. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Comput. Speech Lang.*, vol. 31, no. 1, pp. 65–86, 2015.

[45] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop Autom. Speech Recognit. Understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[46] D. Yu and J. Li, "Recent progresses in deep learning based acoustic models," *IEEE/CAA J. Automatica Sinica*, vol. 4, no. 3, pp. 396–409, 2017.

[47] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (MOS) revisited: Methods and applications, limitations and alternatives," *Multimedia Syst.*, vol. 22, no. 2, pp. 213–227, 2016.

[48] P. Y. Chen, M. Smithson, and P. M. Popovich, *Correlation: Parametric and Nonparametric Measures*. no. 139, Newbury Park, CA, USA: Sage, 2002, no. 139.

[49] T. H. Falk *et al.*, "Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools," *IEEE Signal Process.Mag.*, vol. 32, no. 2, pp. 114–124, Mar. 2015.

[50] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "Student-teacher network learning with enhanced features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5275–5279.

[51] Z. Meng *et al.*, "Adversarial feature-mapping for speech enhancement," 2018, *arXiv:1809.02251*.

[52] L. Wu, H. Chen, L. Wang, P. Zhang, and Y. Yan, "Speaker-invariant feature-mapping for distant speech recognition via adversarial teacher-student learning," a∈A, vol. 1, p. 1, 2019.

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

[54] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[55] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, vol. 2013, 2013, pp. 2345–2349.

[56] J. Kang, W.-Q. Zhang, and J. Liu, "Gated recurrent units based hybrid acoustic models for robust speech recognition," in *Proc. 10th Int. Symp. Chin. Spoken Lang. Process.*, 2016, pp. 1–5.

[57] X. Liu, *Deep Convolutional and LSTM Neural Networks for Acoustic Modelling in Automatic Speech Recognition*. Hoboken, NJ, USA: Pearson Education Inc. 2017, pp. 1–9.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2015.

[59] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *Proc. Int. Conf. Artif. Neural Netw.*, 2010, pp. 92–101.

[60] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2017, pp. 006–012.

[61] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[62] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2401–2405.

[63] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, "Using generalized Gaussian distributions to improve regression error modeling for deep learning-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 1919–1931, Dec. 2019.

[64] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, 1990.

[65] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[66] L. Chai, J. Du, and C.-H. Lee, "Kl-divergence regularized deep neural network adaptation for low-resource speaker-dependent speech enhancement," in *Proc. INTERSPEECH*, 2019, pp. 1806–1810.

Li Chai received the B.S. degree from the Department of Electronic Science and Technology, Xidian University, Xi'an, China, in 2016. She is currently working toward the Ph.D. degree with the University of Science and Technology of China (USTC). Her current research mainly includes speech enhancement and robust speech recognition.

Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above period, he worked as an Intern twice for nine months with Microsoft Research Asia (MSRA), Beijing. In 2007, he was also a Research Assistant for six months with the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he was with iFlytek Research on speech recognition. From July 2010 to January 2013, he was with MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP) of USTC.

Qing-Feng Liu received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), in 1998 and 2003, respectively. He is the Founder and CEO & President with iFLYTEK, Director of the National Speech & Language Engineering Laboratory of China, Professor and Doctoral Advisor with the USTC, Director General of the Union of Speech Industry of China, and the Union of National University Student Innovation & Entrepreneurship.

Chin-Hui Lee (Fellow, IEEE) is a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had 20 years of industrial experience, ending with Bell Laboratories, Murray Hill, New Jersey, as a Distinguished Member of Technical Staff, and Director of the Dialogue Systems Research Department. Dr. Lee is a Fellow of the ISCA. He has authored or coauthored over 500 papers and 30 patents, and has been cited over 34 000 times for his original contributions with an h-index of 80 on Google Scholar. He was the recipient of the numerous awards, including the Bell Labs President's Gold Award in 1998. He also won SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition." In 2012, he was invited by ICASSP to give a plenary talk on the future of speech recognition. In the same year, he was awarded the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition.