End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks

Szu-Wei Fu, Yu Tsao*, Xugang Lu, and Hisashi Kawai

Abstract—Speech enhancement model is used to map a noisy speech to a clean speech. In the training stage, an objective function is often adopted to optimize the model parameters. However, in most studies, there is an inconsistency between the model optimization criterion and the evaluation criterion on the enhanced speech. For example, in measuring speech intelligibility, most of the evaluation metric is based on a short-time objective intelligibility (STOI) measure, while the frame based minimum mean square error (MMSE) between estimated and clean speech is widely used in optimizing the model. Due to the inconsistency, there is no guarantee that the trained model can provide optimal performance in applications. In this study, we propose an end-to-end utterance-based speech enhancement framework using fully convolutional neural networks (FCN) to reduce the gap between the model optimization and evaluation criterion (true targets). Because of the utterance-based optimization, temporal correlation information of long speech segments, or even at the entire utterance level, can be considered when perception-based objective functions are used for the direct optimization. As an example, we implement the proposed FCN enhancement framework to optimize the STOI measure. Experimental results show that the STOI of test speech is better than conventional MMSE-optimized speech due to the consistency between the training and evaluation target. Moreover, by integrating the STOI in model optimization, the performance of the automatic speech recognition (ASR) system on the enhanced speech is also substantially improved compared to those generated by the MMSE criterion.

Index Terms—automatic speech recognition, fully convolutional network, raw waveform, end-to-end speech enhancement, speech intelligibility

I. INTRODUCTION

Recently, deep learning based spectral mapping frameworks for speech enhancement have been proposed and extensively investigated [1-23]. Although they were demonstrated to perform better than state-of-the-art enhancement models, there are still some remaining unresolved research issues. One of the key problems is that the objective function used for optimization in the training stage, typically the minimum mean squared error (MMSE) [24] criterion, is different from the human perception-based evaluation measures. Formulating consistent training objectives that meet specific evaluation criteria has always been a challenging task for signal processing (generation) [25].

For human perception, the primary goal of speech enhancement is to improve the intelligibility and quality of noisy speech [26]. To evaluate these two metrics, perceptual evaluation of speech quality (PESQ) [27] and short-time objective intelligibility (STOI) [28] have been proposed and used as objective measures by many related studies [1-5, 10-22]. However, most of them did not use these two metrics as the objective function for optimizing their models. Instead, they simply minimized the mean square error (MMSE) between clean and enhanced features. Although some research [10, 11] introduced human perception into the objective function, they are still different from the final evaluation metrics. Optimizing a substitute objective function does not guarantee good results for the true targets. We will discuss this problem and give some examples in detail in Section III.

The reasons for not applying the evaluation metrics (true targets) as objective function may not only be due to the complicated computation, but also because the whole (clean and noisy/processed) utterances are needed to accomplish the evaluation. Usually, deep neural networks (DNNs) enhance noisy speech in a frame-wise manner due to restrictions of the model structures. In other words, during the training process, each noisy frame is individually optimized (or some may include context information), and there is no way to treat the utterance as a whole. Therefore, both PESQ and STOI cannot be employed as objective functions to optimize current speech enhancement models.

In this study, we employ a fully convolutional neural network (FCN) to solve the mismatch between evaluation metrics of speech enhancement and the employed objective function. FCN is very similar to a conventional convolutional neural network (CNN), except that the top fully connected layers are

Szu-Wei Fu is with Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan and Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taipei 11529, Taiwan (e-mail: jasonfu@cit.sinica.edu.tw).

Xugang Lu is with the National Institute of Information and Communications Technology, Tokyo 184-0015, Japan (e-mail: xugang.lu@nict.go.jp).

Hisashi Kawai is with the National Institute of Information and Communications Technology, Tokyo 184-0015, Japan (e-mail: hisashi.kawai@nict.go.jp).

Yu Tsao is with the Research Center for Information Technology Innovation (CITI) at Academia Sinica, Taipei 11529, Taiwan (e-mail: yu.tsao@citi.sinica.edu.tw).
removed [29]. Therefore, it only consists of convolutional layers, and hence the local feature structures can be effectively preserved with a relatively small number of weights. Through this property, waveform-based speech enhancement by FCN was proposed, and it achieved considerable improvements when compared to DNN-based models [30]. Here, we apply another property of FCN to achieve utterance-based enhancement, even though each utterance has a different length. The reason DNN and CNN can only process fixed-length inputs [31] is that the fully connected layer is indeed a matrix multiplication between the weight matrix and outputs of the previous layer. Because the shape of the weight matrix is fixed when the model structure (number of nodes) is decided, multiplication is infeasible if the input length is not fixed. However, the filters in convolution operations can accept inputs with variable lengths.

We mainly follow the framework established in [30] to construct an utterance-based enhancement model. Based on this processing structure, we further utilize STOI as our objective function. There are three reasons why we only focus on optimizing STOI. First, optimization of PESQ as an objective function is much more complicated. In fact, some functions (e.g., the asymmetry factor for modeling the asymmetrical disturbance) in PESQ computation are non-continuous, so gradient descent-based optimization cannot be directly used [32] (this problem can be solved by substituting a continuous approximation function for the non-continuous function or by reinforcement learning, as presented in [33]). Second, improving speech intelligibility is often more challenging than enhancing quality [34, 35]. Because the MMSE criterion used in most conventional enhancement-model learning algorithms are not designed to directly improve intelligibility, the STOI based optimization criterion is expected to perform better. Third, some research [36, 37] has shown that the correlation coefficient (CC) between the improvement in word error rate (WER) of ASR and the improvement in STOI is higher than other objective evaluation scores (e.g., PESQ). Since we demand high accuracy noise-robust ASR in real-world applications, an STOI-optimized speech enhancement front-end may achieve better ASR performance than MMSE-based alternatives under noisy conditions. Note that the proposed utterance-based FCN enhancement model can handle any kind of objective functions from a local time scale (frame) to a global time scale (utterance). More specifically, our model can directly optimize the final evaluation criterion, and the STOI optimization demonstrated in this paper is just one example.

Experimental results on speech enhancement show that incorporating STOI into the objective function can improve not only the corresponding objective metric, but also the robustness of ASR under noisy conditions. This is particularly important for real-world hands-free ASR applications, such as human-robot interactions [38].

The rest of the paper is organized as follows. Section II introduces the proposed FCN for utterance-based waveform speech enhancement. Section III details the optimization for STOI. The experimental results are evaluated in Section IV. Finally, Section V presents our discussion, and this paper is concluded in Section VI.

II. END-TO-END WAVEFORM BASED SPEECH ENHANCEMENT

In addition to frame-wise processing, the DNN-based models usually also cause two other drawbacks. First, they focus only on processing the magnitude spectrogram, such as log-power spectra (LPS) [1], and leave the phase in its original noisy form [1-6]. However, several recent studies have revealed the importance of phase to speech quality when speech is resynthesized back into time-domain waveforms [39-41]. Second, a great deal of pre-processing (e.g., framing, discrete Fourier transform (DFT) and post-processing (e.g., overlap-add method, inverse discrete Fourier transform) are necessary for mapping between the time and frequency domains, thus increasing the computation load.

Although some recent studies have taken the phase components into consideration using complex spectrograms [12-14], these methods still need to transform the waveform into the frequency domain. To solve the two issues listed above, waveform-based speech enhancement by FCN was proposed and achieved considerable improvements when compared to the LPS-based DNN models [30]. In fact, other waveform enhancement frameworks based on generative adversarial networks (GANs) [25] and WaveNet [42, 43] were also shown to outperform conventional models. Although most of these methods have already achieved remarkable performance, they still processed the noisy waveforms in a frame-based (or chunk-based) manner. In other words, the final evaluation metrics were still not applied as the objective functions to train their models.

A. FCN for Waveform Enhancement

FCN is very similar to a conventional CNN, except that all the fully connected layers are removed [29]. More specifically, FCN only consists of convolutional layers; hence, the local structures of features can be effectively preserved with a relatively small number of weights. Since the effect of convolving a time domain signal $x(t)$ with a filter $h(t)$ is equivalent to multiplying its frequency representation $X(f)$ with the frequency response $H(f)$ of the filter [44]. Therefore, it may be unnecessary to explicitly map the waveform to a spectrogram for speech enhancement.

The characteristics of a signal represented in the time domain are very different from those in the frequency domain. In the frequency domain, the value of a feature (frequency bin) represents the energy of the corresponding frequency component. However, in the time domain, a feature (sample point) alone does not carry much information; it is the relation with its neighbors that represents the concept of frequency. Fu et al. pointed out that this interdependency may make DNN laborious for modeling waveforms, because the relation between features is removed after fully connected layers [30]. On the other hand, because each output sample in FCN depends locally on the neighboring input regions [45], the relation between features can be well preserved. Hence, FCN was employed to solve this problem, and obtained considerable improvements compared to the DNN-based model.
B. Utterance-based Enhancement

In spite of the fact that the noisy waveform can be successfully denoised by FCN [30], it is still processed in a frame-wise manner (each frame contains 512 sample points). In addition to the problem of a greedy strategy [46], this also makes the convolution results inaccurate because of the zero-padding in the frame boundary. Therefore, in this study, we apply another property of FCN to achieve utterance-based enhancement, even though each utterance has a different length. The reason DNN and CNN can only process fixed-length inputs is that the fully connected layer is the result of matrix multiplication between the weight matrix and outputs of previous layer. Since the shape of the weight matrix is fixed when the model structure (number of nodes) is decided, multiplication is infeasible if the input length is not fixed. However, the filters in convolution operations can accept inputs with different lengths. Specifically, if the filter length is $l$ and the length of input signal is $L$, then the length of the filtered output is $L-l+1$. Because FCN only consists of convolutional layers, and all the fully connected layers are removed, it can process the whole utterance without frame processing.

The overall proposed FCN structure for utterance-based waveform enhancement is shown in Fig. 1, where $\text{Filter}_m$ represents the $n$th filter in layer $m$. Note that this is a complete end-to-end (noisy waveform utterance in and clean waveform utterance out) framework, and there is no pre- or post-processing needed.

III. Optimization for Speech Intelligibility

In addition to solving the frame boundary problem caused by zero-padding, another benefit of utterance-based optimization is the ability to design an objective function that is used for the whole utterance. In other words, each utterance is treated as a whole so that the global optimal solution (for the utterance) can be more easily obtained. Before introducing the objective function used for speech intelligibility optimization, we first show that minimizing the MSE between clean and enhanced features may not be the most suitable target due to the characteristics of human hearing.

A. Problems of Applying MSE as an Objective Function

One of the most intuitive objective functions used in speech enhancement is the MSE between the clean and enhanced speech. However, MSE simply compares the similarity between two signals, and does not consider human perception. For example, Loizou et al. pointed out that MSE pays no attention to positive or negative differences between the clean and estimated spectra [34, 35]. A positive difference would signify attenuation distortion, while a negative spectral difference would signify amplification distortion. The perceptual effect of these two distortions on speech intelligibility cannot be assumed to be equivalent. In other words, MSE is not a good performance indicator of speech, and hence it is not guaranteed that better-enhanced speech can be obtained by minimizing MSE. The upper row of Fig. 2 shows an example of this case in the frequency domain. Although the MSE (between clean LPS and enhanced LPS) of enhanced speech in Fig. 2 (b) is lower than that in Fig. 2 (c), its performance (in terms of STOI, PESQ, and human perception) is worse than the latter. This is because the spectrogram in Fig. 2 (b) is over-smoothing, and details of the speech components are missing. As pointed out in [25], the prediction results of MMSE usually bias towards an average of all the possible predictions. The two spectrograms are actually obtained from the same denoising model, but with a different training epoch. Fig. 2 (b) is from an optimal training epoch by early stopping [47] while Fig. 2 (c) comes from an “overfitting” model due to overtraining. Note that here we use double quotes to emphasize that this overfitting is relative to the MSE objective function, and not to our true targets of speech enhancement. The above discussion implies that minimizing the MSE may make the estimated speech looks like the clean one; however, sometimes a larger MSE in the optimization process can produce speech sounds more similar to the clean version.
An enhanced speech with lower MSE does not guarantee a better performance in evaluation. The upper row shows the case in the frequency domain, where the MSE is measured between a clean LPS and an enhanced LPS. The lower row shows the case in the time domain, where the MSE is measured between a clean waveform and an enhanced waveform.

Although the waveform-based FCN enhancement model in [30] is optimized with an MSE objective function, it is also not the best target for the time domain waveform, because the relation between the MSE value and human perception is still not a monotonic function. For example, as shown in Fig. 3, it is difficult for people to distinguish between a waveform, its negative version, and its shifted version by listening, although the MSE between them is very large.

This also verifies the argument made in section II-A that sample point itself does not carry much information; it is the relation with its neighbors that represent the concept of frequency. The lower row of Fig. 2 also shows a real example in the time domain in which an enhanced speech with a lower MSE (between the clean and enhanced waveforms) does not guarantee better performance.

B. Introduction of STOI

To overcome the aforementioned problem of MSE, here we introduce an objective function which considers human hearing perception. The STOI score is a prevalent measure used to predict the intelligibility of noisy or processed speech. The STOI score ranges from 0 to 1, and is expected to be monotonically related to the average intelligibility of various listening tests. Hence, a higher STOI value indicates better speech intelligibility.

STOI is a function of the clean and degraded speech, and the overall computational process is illustrated as in Fig. 4. The calculation of STOI includes 5 major steps, briefly described as follows:

1) Remove silent frames: Since silent regions do not contribute to speech intelligibility, they are removed before evaluation.

2) Short-time Fourier transform (STFT): Both signals are TF-decomposed in order to obtain a representation similar to the speech representation properties in the auditory system. This is obtained by segmenting both signals into 50% overlapping Hann-windowed frames, with a length of 256 samples, where each frame is zero-padded up to 512 samples.
3) One-third octave band analysis: This is performed by simply grouping DFT-bins. In total, 15 one-third octave bands are used, where the lowest center frequency is set to 150 Hz and the highest one-third octave band has a center-frequency of ~4.3 kHz. The following vector notation is used to denote the short-time temporal envelope of the clean speech:

\[ x_{j,m} = [X_f(m - N + 1), X_f(m - N + 2), \ldots, X_f(m)]^T \]

where \( X \in R^{15 \times M} \) is the obtained one-third octave band, \( M \) is the total number of frames in the utterance, \( m \) is the index of the frame, \( f \in \{1, 2, \ldots, 15\} \) is the index of the one-third octave band, and \( N = 30 \), which equals an analysis length of 384 ms. Similarly, \( \tilde{x}_{j,m} \) denotes the short-time temporal envelope of the degraded speech.

4) Normalization and clipping: The goal of the normalization procedure is to compensate for global level differences, which should not have a strong effect on speech intelligibility. The clipping procedure ensures that the sensitivity of the STOI evaluation towards one severely degraded TF-unit is upper bounded. The normalized and clipped temporal envelope of the degraded speech is denoted as \( \tilde{x}_{j,m} \).

5) Intelligibility measure: The intermediate intelligibility measure is defined as the correlation coefficient between the two temporal envelopes:

\[ d_{j,m} = \frac{(x_{j,m} - \mu_{x_{j,m}})(\tilde{x}_{j,m} - \mu_{\tilde{x}_{j,m}})}{\| x_{j,m} - \mu_{x_{j,m}} \|_2 \| \tilde{x}_{j,m} - \mu_{\tilde{x}_{j,m}} \|_2} \]

where \( \| \cdot \|_2 \) represents the L2-norm, and \( \mu_{(i)} \) is the sample mean of the corresponding vector. Finally, STOI is calculated as the average of the intermediate intelligibility measure over all bands and frames:

\[ \text{STOI} = \frac{1}{15M} \sum_{j,m} d_{j,m} \]

The calculation of STOI is based on the correlation coefficient between the temporal envelopes of the clean and the noisy/processed speech for short segments (e.g., 30 frames). Therefore, this measure cannot be optimized by a traditional frame-wise enhancement scheme. For a more detailed setting of each step, please refer to [28].

C. Maximize STOI for Speech Intelligibility

Although the calculation of STOI is somewhat complicated, most of the computation is differentiable, and thus it can be employed as the objective function for our utterance optimization. Therefore, the objective function that should be minimized during the training of FCN can be represented by the following equation.

\[ O = -\frac{1}{N} \sum \text{stoi}(w_n(t), \tilde{w}_n(t)) \]

where \( w_n(t) \) and \( \tilde{w}_n(t) \) are the clean and estimated utterance with index \( n \), respectively, and \( N \) is the total number of training utterance. \( \text{stoi}(\cdot) \) is the function that includes the five steps stated in previous section, which calculates the STOI value of the noisy/processed utterance given the clean one.

IV. EXPERIMENT

In this experiment, we prepare two data sets to evaluate the performance of different enhancement models. The first is the TIMIT corpus [48], so that the results presented here can also be compared to those shown in [30]. The other data set is the Mandarin version of the Hearing in Noise Test (MHINT) corpus [49], to verify that the improvement using the proposed method is dataset- and language-independent. Note that the results should not be compared across these two data sets, as their settings (e.g., SNR level, noise type, and size of training data) are different.

A. Experiment on TIMIT data set

The TIMIT corpus was used to prepare the training and test sets. For the training set, 600 utterances were randomly selected and corrupted with five noise types (Babble, Car, Jackhammer, Pink, and Street) at five SNR levels (-10 dB, -5 dB, 0 dB, 5 dB, and 10 dB). For the test set, we randomly selected another 100 utterances (different from those used in the training set). To make the experimental conditions more realistic, both the noise types and SNR levels of the training and test sets were
mismatched. Thus, we adopted three other noise signals: white Gaussian noise (WGN), which is a stationary noise, and an engine noise and a baby cry, which are non-stationary noises, using another five SNR levels (-12 dB, -6 dB, 0 dB, 6 dB, and 12 dB) to form the test set. All the results reported were averaged across the three noise types. For more detailed experiment settings and model structure, refer to [30].

To evaluate the performance of speech intelligibility, the STOI scores were used as a measure. We also present PESQ for speech quality to make a complete comparison with the results shown in [30]. (Although this metric is not optimized in this paper, we also report the results for completeness.) Table I presents the results of the average STOI and PESQ scores on the test set for the frame-based FCN [30] and the proposed utterance-based FCN with different objective functions, where obj represents the objective function used for training. Note that all three models have the same structure, and the only difference between them is the objective function or input unit (frame or utterance). From this table, we can see that the utterance-based FCN (with MSE objective function) can outperform frame-based FCN in both PESQ and STOI. This improvement mainly comes from solving the frame boundary problem in frame-based optimization. When employing the STOI as the objective function, it can considerably increase the STOI value (with an improvement of 0.04 on average), especially in low-SNR conditions. Although the average PESQ decreases, our goal in this study is mainly focused on maximizing STOI.

| SNR (dB) | STOI | PESQ | STOI | PESQ | STOI | PESQ | STOI | PESQ | STOI | PESQ |
|----------|------|------|------|------|------|------|------|------|------|------|
| 9        | 0.893| 1.733| 0.904| 2.048| 0.911| 2.361| 0.931| 2.115| 0.925| 2.278|
| 6        | 0.853| 1.557| 0.879| 1.873| 0.888| 2.183| 0.909| 1.935| 0.901| 2.109|
| 3        | 0.802| 1.395| 0.842| 1.682| 0.853| 1.975| 0.877| 1.732| 0.865| 1.933|
| 0        | 0.743| 1.259| 0.792| 1.496| 0.805| 1.735| 0.834| 1.510| 0.817| 1.729|
| -3       | 0.678| 1.124| 0.732| 1.317| 0.741| 1.444| 0.780| 1.286| 0.759| 1.468|
| -6       | 0.612| 0.942| 0.661| 1.152| 0.661| 1.134| 0.712| 1.062| 0.686| 1.183|
| Avg.     | 0.763| 1.347| 0.802| 1.595| 0.810| 1.805| 0.841| 1.607| 0.826| 1.783|

**B. Experiment on MHINT data set**

1) Experiment Setup

In this experiment, the MHINT corpus was used to prepare the training and test sets. This corpus includes 240 utterances, and we collected another 240 utterances from the same speaker to form the complete task in this study. Among these 480 utterances, 280 utterances were excerpted and corrupted with 100 noise types [50], at five SNR levels (-10 dB, -5 dB, 0 dB, 5 dB, and 10 dB). Another 160 utterances were mixed to form the test set. In this experiment, we still consider a realistic condition, where both noise types and SNR levels of the training and test sets were mismatched. Thus, we intentionally adopted three other noise signals: street noise, engine noise, and two talkers, with another six SNR levels: -6 dB, -3 dB, 0 dB, 3 dB, 6 dB, and 9 dB (different from the training conditions) to form the test set. All the results reported were averaged across the three noise types.

The FCN in this experiment had 6 convolutional layers with padding to preserve the same size as the input. Except for only 1 filter used in the last layer, each of the previous layers consisted of 15 filters with a filter size of 23.

Here, we also try to incorporate both the MSE and STOI into the objective function so that it can be represented by the following equation.

$$O = \frac{1}{L_n} \sum_{t=1}^{L_n} \left[ \| w_n(t) - \hat{w}_n(t) \|_2^2 - stoi(w_n(t), \hat{w}_n(t)) \right].$$  (5)

where $L_n$ is the length of $w_n(t)$ (note that each utterance has a different length), and $\alpha$ is the weighting factor of the two targets. Here, $\alpha$ is simply set to 100 to balance the scale of the two targets. Since the first term can be seen as related to maximizing the SNR of enhanced speech, and the second term is to maximize the STOI, the two targets in (5) can also be treated as a kind of multi-metrics learning [14] for speech enhancement.

2) Experiment Results of STOI scores

The raw waveform-based (frame-wise) FCN has shown its superior enhancement performance compared to the LPS-based DNN model in [30]. Therefore, like the experiments for TIMIT data set, here we also treat the frame-wise FCN as a baseline and compared it with the proposed utterance-based models, as shown in Table II. From this table, we can observe similar results as the TIMIT data set.
First, the utterance-based FCN (with MSE objective function) can outperform the frame-based FCN both in PESQ and STOI. Second, when changing the objective function from MSE to STOI, the STOI value of the enhanced speech can also be considerably improved with a decreased PESQ score. In addition, the last two columns in Table II show that the results of combining MSE and STOI as the objective function reach a balance between speech intelligibility and quality.

3) Spectrogram Comparison

Next, we present the spectrograms of a clean MHINT utterance, the same utterance corrupted by engine noise at 9 dB, and FCN-enhanced speeches with different input units and objective functions in Fig. 5. When comparing Fig. 5(c) and (d), we can clearly observe that the utterance-based FCN can remove more noise than the frame-based one. Though the spectrogram enhanced by STOI-optimized FCN (Fig. 5(e)) is noisy and the high frequency components are missing, the speech components in the low- to mid-frequency regions can be well preserved when compared to Fig. 5(d)). The noisy output is mainly due to the calculation of intermediate intelligibility in (2), during the estimation of STOI based on the correlation coefficient. The correlation coefficient is a measure with scale- and shift-invariance (i.e., when a vector is shifted or scaled by a constant, then the correlation coefficient with another vector keeps unchanged). Therefore, during the optimization of STOI, removing the stationary artifacts (as shown in the dashed black box in Fig. 5(e)) gives limited improvement to STOI. The missing high frequency components are also attributed to the calculation of STOI. As the highest one-third octave band has a center-frequency equal to ~4.3 kHz [28], the frequency components above this value do not affect the estimation of STOI (i.e., whether this region is very noisy or empty, the STOI value is not decreased). Therefore, FCN learns not to make any effort on this high-frequency region, and just removes most of the components. These two characteristics are the main reasons for decreased PESQ compared to the MSE-optimized counterpart. The two aforementioned phenomena of the STOI-optimized spectrogram can be mitigated by also incorporating MSE into the objective function, as shown in Fig. 5 (f).
From this, we can, this implies that the ASR as a normal hearing human. Therefore, we take a trained ASR (Google Speech Recognition) [54] to accomplish this task.

The same MHINT test sentences used in the objective evaluations were also adopted in the ASR experiment, and the results reported were averaged across the three noise types. The WER of ASR for noisy speech, enhanced speech by LPS-based DNN method, and waveform-based FCN enhancement models with different objective functions are shown in Fig. 7. This figure provides the following four observations: 1) the conventional DNN-based LPS enhancement method can provide similar WER improvement to the proposed FCN model in low-SNR conditions. However, under high-SNR conditions, its performance decreases, and is even worse than the noisy speech at 9 dB case. 2) All the FCN enhanced speech samples can obtain lower WER compared to the noisy ones, and the improvement at around 0 dB is most obvious. 3) STOI-optimized speech can only provide marginal WER improvement compared to the MSE-optimized speech. This may be due to the spectrogram of STOI-optimized speech remaining too noisy for ASR (compare Fig. 5 (d) and (e)). Furthermore, PESQ is decreased by changing the objective function from MSE to STOI (compare the 5th to 8th columns in Table II). Although not as highly correlated as the STOI case, the decrease of PESQ may also degrade the ASR performance (the correlation coefficient between improvement in WER and the improvements in PESQ is 0.55 [36]). Therefore, most of the WER reduction from increasing STOI may be canceled out by the decreasing PESQ. 4) When incorporating both MSE and STOI into the objective function of FCN, the WER (except -6 dB) can be considerably reduced compared to the MSE-optimized model. This verifies that bringing STOI into objective function of speech enhancement can also help ASR to identify the speech content under noisy conditions.

Although this ASR experiment was tested on a trained system, this is indeed more practical in many real-world applications where an ASR engine is supplied by a third-party. The experimental results show that there is no need to retrain the ASR, and our proposed FCN enhancement model can simply be treated as pre-processing to obtain a noise-robust ASR.

In summary, although optimizing STOI alone can only provide marginal WER improvements, incorporating STOI into an objective function with MSE can provide considerable benefits. Since both the STOI and PESQ can strike a balance in this case (please compare the last six columns in Table II), this implies that to efficiently improve ASR, the speech quality (PESQ) cannot be totally neglected. From these discussions, we can conclude that despite the high correlation between ASR and human perception, they still have some difference.

V. DISCUSSION

Since the basic unit in the evaluation of STOI is a short segment (30 frames), it cannot be optimized by conventional frame-based enhancement model. Moreover, because these short segments overlap each other (shift 1 frame per step, in-
stead of 30 frames), segment-based enhancement still cannot be used for maximizing STOI.

It may also seem that STOI can be optimized by applying 2D FCN on the spectrogram of whole utterance (starting from the 3rd step in the STOI evaluation). However, the enhanced magnitude spectrogram $E$ may not correspond to the STFT of any time-domain signal anymore \[40, 55\]. As a consequence, the resynthesized signal may have a spectrogram $\hat{E}$, which is different from the desired spectrogram $\hat{E}$. This implies that even if the estimated magnitude $\hat{E}$ is optimal with respect to STOI, the magnitude spectrogram of the synthesized time-domain signal is not. Hence, to maintain optimality, the phase should also be considered in the optimization process or using waveforms directly.

From the above discussion, we observe the flexibility of the proposed utterance-based waveform enhancement FCN model. It can handle any kind of objective function, from the local time scale (frame or short segment) to the global time scale (long segment or utterance), and from measures in the time domain to the frequency domain. The STOI optimization demonstrated in this paper is just one example. Therefore, even if a better objective evaluation metric is proposed in the future, our model can still be directly optimized with that evaluation criterion, as long as every step in the evaluation is differentiable (otherwise, a continuous approximation function is needed).

Another advantage of utterance-based enhancement is that it can essentially solve the discontinuity of speech produced by frame-based enhancement models. Conventionally, because the enhanced speech is independently generated in a frame-by-frame manner, it shows some level of discontinuity, even though context features are used as input \[17\]. However, in our proposed framework, the basic processing unit is the whole utterance, instead of a frame. Hence, the discontinuity problem does not exist in the utterance-based denoising model.

VI. CONCLUSION

This paper proposes an end-to-end utterance-based raw waveform speech enhancement framework based on FCN. Through this novel structure, several problems that exist in conventional DNN-based enhancement model can be solved simultaneously.

1) There is no need to map the time domain waveform to the frequency domain for enhancing the magnitude spectrogram.

2) Therefore, all the related pre- and post-processing can be avoided.

3) Because the proposed model directly denoises the noisy waveform, the phase information is not ignored.

4) The discontinuity of enhanced speech observed in conventional frame-based processing is solved by treating each utterance as a whole.

5) The mismatch between the true targets of speech enhancement and the employed objective function can be solved by utterance-based waveform optimization. Experimental results show that incorporating STOI into an objective function can not only improve the intelligibility, but also increase the robustness of ASR under noisy conditions. In the future, we will try to investigate whether applying a dilated convolution and skip connection in FCN can further improve performance. In addition, we will also directly optimize PESQ through an approximation function, to substitute the non-continuous function encountered during the computation. Through combining the PESQ objective function with the proposed STOI optimization, an enhancement model that can simultaneously consider two targets (quality and intelligibility) can hence be trained by multi-metrics learning \[14\].

REFERENCES

\[1\] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “An experimental study on speech enhancement based on deep neural networks,” IEEE Signal Processing Letters, vol. 21, pp. 65-68, 2014.

\[2\] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “Global variance equalization for improving deep neural network based speech enhancement,” in IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), 2014, pp. 71-75.

\[3\] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “Dynamic noise aware training for speech enhancement based on deep neural networks,” in INTERSPEECH, 2014, pp. 2670-2674.

\[4\] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing vol. 23, pp. 7-19, 2015.

\[5\] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, “SNR-based progressive learning of deep neural network for speech enhancement,” in INTERSPEECH, 2016, pp. 3713-3717.

\[6\] S.-W. Fu, Y. Tsao, and X. Lu, “SNR-aware convolutional neural network modeling for speech enhancement,” in INTERSPEECH, 2016, pp. 3768-3772.

\[7\] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in INTERSPEECH, 2013, pp. 436-440.

\[8\] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, “Integration of speech enhancement and recognition using long-short term memory recurrent neural network,” in INTERSPEECH, 2015.

\[9\] F. Weninger, H. Erdogan, S. Watanabe, E. Vigna, J. Le Roux, J. R. Hershey, et al., “Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR,” in International Conference on Latent Variable Analysis and Signal Separation, 2015, pp. 91-99.

\[10\] B. Xia and C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," Speech Communication, vol. 60, pp. 13-29, 2014.

\[11\] P. G. Shivakumar and P. Georgiou, "Perception optimized deep denoising autoencoders for speech enhancement," in INTERSPEECH, 2016, pp. 3743-3747.

\[12\] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5220-5224.

\[13\] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, pp. 483-492, 2016.

\[14\] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in MSLP, 2017.

\[15\] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in INTERSPEECH, 2015.

\[16\] A. Ogawa, S. Seki, K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, et al., "Robust example search using bottleneck features for example-based speech enhancement," in INTERSPEECH, 2016, pp. 3733-3737.

\[17\] S.-S. Wang, H.-T. Hwang, Y.-H. Lai, Y. Tsao, X. Lu, H.-M. Wang, et al., "Improving denoising auto-encoder based speech enhancement with the speech parameter generation algorithm," in APSIPA, 2015, pp. 365-369.

\[18\] D. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," arXiv preprint arXiv:1708.07524, 2017.

\[19\] S. Wang, K. Li, Z. Huang, S. M. Siniscalchi, and C.-H. Lee, “A transfer learning and progressive stacking approach to reducing deep model sizes with an application to speech enhancement,” in
McInnes, and M. Jack, “Lateral inhibition net and software.”

A. Zhang. (2017). Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html.

A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, et al., “Wavenet: a generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

J. W. Lyons, “DARPA TIMIT acoustic-phonetic continuous speech corpus,” National Institute of Standards and Technology, 1993.

L. L. Wong, S. D. Soli, S. Liu, N. Han, and M.-W. Huang, “Development of the Mandarin hearing in noise test (MHINT),” Ear and hearing, vol. 28, pp. 708-745, 2007.

G. Hu. 100 nonspeech environmental sounds, 2004 [Online]. Available: http://www.cse.ohio-state.edu/pnl/corpus/HuCorpus.html.

L. Prechelt, “Early stopping—when?” in Neural Networks: Tricks of the trade, ed: Springer, 1998, pp. 55-69.

T. Gerkmann, M. Krawczyk and A. Hekstra, “Phase sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 708-712.

K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, “Speech enhancement using Bayesian Wavenet,” in INTERSPEECH, 2017, pp. 2013-2017.

D. Rethage, J. Pons, and X. Serra, “A Wavenet for speech denoising,” arXiv preprint arXiv:1706.07162, 2017.

A. V. Oppenheim, Discrete-time signal processing; Pearson Education India, 1999.

H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 708-712.