A Speech Intelligibility Enhancement Model based on Canonical Correlation and Deep Learning for Hearing-Assistive Technologies

Tassadaq Hussain, Muhammad Diyan, Mandar Gogate, Kia Dashtipour, Ahsan Adeel, Yu Tsao, and Amir Hussain

Abstract—Current deep learning (DL) based approaches to speech intelligibility enhancement in noisy environments are generally trained to minimise the distance between clean and enhanced speech features. These often result in improved speech quality however they suffer from lack of generalisation and may not deliver the required speech intelligibility in everyday noisy situations. In an attempt to address these challenges, researchers have explored intelligibility-oriented (I-O) loss functions to train DL approaches for robust speech enhancement (SE). In this paper, we formulate a novel canonical-correlation based I-O loss function to more effectively train DL models trained with short-time objective intelligibility (CC-STOI) metric as a training cost function. To the best of our knowledge, this is the first work that exploits the integration of canonical correlation in an I-O based loss function for SE. Comparative experimental results demonstrate that our proposed CC-STOI based SE framework outperforms DL models trained with conventional STOI and distance-based loss functions, in terms of both standard objective and subjective evaluation measures when dealing with unseen speakers and noises.

I. INTRODUCTION

Speech enhancement (SE) systems are designed to improve the quality and intelligibility of speech signals in real-world situations where speech is often distorted by many competing additive or convolutive noises. Recently, non-linear denoising spectral mapping-based approaches have shown excellent performance for SE task. For example, a deep denoising autoencoder (DDAE) framework has demonstrated promising SE performance compared to traditional methods [1]. Subsequently, a deep neural network (DNN) was adopted to handle a wide range of additive noises for the SE task [2]. In addition to standard feed-forward neural networks, different structures of convolutional neural networks (CNNs) have been employed in an attempt to improve the generalisation performance for SE. In [3], a CNN was trained in an encoder-decoder style with an additional temporal convolutional module to provide real-time SE. In [4], a fully convolutional neural network (FCN) was exploited to effectively recover the enhanced speech waveform for SE in an end-to-end manner. Different from traditional DL-based approaches, authors in [5] adopted a novel strategy and trained a FCN using an objective evaluation-based cost function for enhanced speech perception.

Despite the excellent performance achieved by DL based SE models, the parameters of such approaches are often optimized using distance-based loss functions including mean squared error (MSE) and mean absolute error (MAE). However, these may not be optimal performance metrics for speech-related applications as they are not based on human auditory perception. We believe that optimizing human perception-based evaluation metrics directly may lead to more optimal results corresponding to the target task. In the context of SE, researchers usually employ a number of performance evaluation metrics that are inspired by human auditory perception. There are two widely used metrics, specifically, perceptual evaluation of speech quality (PESQ) [6] and short-time objective intelligibility (STOI) [7], which are used to approximate subjective speech quality and intelligibility, respectively. Apart from conventional MSE/MAE-based DL approaches, a number of intelligibility-oriented (I-O) STOI-metric based DL approaches have also been proposed and shown to be effective for SE. For example, in [5], authors utilized the STOI measure as an objective function to optimize a fully-convolutional network (FCNN) model for SE. The results demonstrated that the STOI-based SE framework can perform significantly better than a conventional MSE-based SE system due to increased consistency between the training and evaluation target. More recently, authors in [8] studied the influence of six different loss functions (including the STOI-based cost function) and evaluated them in a structured manner with end-to-end time-domain DL-based SE systems.

Inspired by the excellent performance achieved by STOI-based SE systems, we propose a novel canonical-correlation based short-time objective
where $M$ framework [11]. The MSE can be computed as follows:

$$L_{\text{MSE}} = \min \left( \frac{1}{M} \sum_{m=1}^{M} \| \hat{Y}_m - Y_m \|_2 \right)$$  \hspace{1cm} (1)

where $M$ is the total number of speech frames, $\hat{Y}_M$ is the estimated magnitude spectrum, $Y_M$ is the reference magnitude spectrum of the utterance, and $\| \cdot \|_2$ denotes L2-normalization.

In this paper, we adopt a deep FCN-based U-Net architecture that takes the noisy magnitude spectrum as input and exploits a CC-STOI loss function to optimally learn the spectral mapping and estimate the enhanced magnitude spectrum. Figure 1 shows the block diagram of our proposed CC-STOI framework for SE, and Figure 2 illustrates the FCN-based U-Net framework used to optimize our SE model. Our key focus is to explore how incorporating CCA into STOI loss function impacts the quality and intelligibility of the enhanced speech signal as evaluated with a range of standardized objective measures.

A. Short-time Objective Intelligibility (STOI)

To train frequency domain I-O SE models, we use a modified version of a well-known STOI proposed in [12] as an objective function. The STOI measure is a five step process where the STOI function takes the clean and estimated speech signals as input and computes the score by: i) removing the silent frames from clean and estimated speech signals, ii) applying the short-time Fourier transform (STFT), iii) Estimating the envelope of clean and noisy speech using one-third octave-band analysis of the STFT frames, iv) Normalizing and clipping to compensate for global level differences and stabilisation of the STOI evaluation, and v) measuring intelligibility. To optimise the correlation between the two spectral envelopes, we replaced the correlation coefficient of standard STOI with the CCA. The correlation coefficient between the two spectral envelopes is estimated using the equations below to optimise the CC-STOI function.

$$d_{i,j} = \frac{(y_{i,j} - \mu_{y_{i,j}})^T(y_{i,j} - \mu_{\hat{y}_{i,j}})}{\|y_{i,j} - \mu_{y_{i,j}}\|_2 \|\hat{y}_{i,j} - \mu_{\hat{y}_{i,j}}\|_2}$$ \hspace{1cm} (2)

where $y$ and $\hat{y}$ are the short-time spectral envelope of the reference clean and estimated speech signals, $\mu_{y_{i,j}}$ and $\mu_{\hat{y}_{i,j}}$ are the corresponding sample mean vectors, and $\| \cdot \|_2$ represents the L2-normalization. The final CC-STOI is the average of the intelligibility measure over all bands and frames.

$$d_{\text{CC-STOI}} = \frac{1}{I(M - N + 1)} \sum_{i=1}^{I} \sum_{j=1}^{J} d_{i,j}$$ \hspace{1cm} (3)

where $I = 15$ is the number of one-third octave band and $M - N + 1$ is the total number of short-time temporal envelope vector. For a more detailed setting of each step, please refer to [7]. The computation of CC-STOI is differentiable, thus, it can be used as the objective function directly to optimize the SE model.

$$L_{\text{CC-STOI}} = -\frac{1}{M} \sum_{m=1}^{M} d_{\text{CC-STOI}}(\hat{Y}_m, Y_m)$$ \hspace{1cm} (4)

where $d_{\text{CC-STOI}}(\hat{Y}_m, Y_m)$ measures the CC-STOI score between the estimated and clean magnitude spectra of audio utterances. Unlike MSE, where the goal is to reduce the distance, we want to maximize the CC-STOI score to enhance speech intelligibility.
B. Proposed SE Framework

This section presents the DL models used for our I-O SE framework as depicted in Fig. 2. Specifically the network architecture for feature extraction and speech resynthesis pipeline is outlined below.

1) Audio feature extraction: For SE, the speech feature extraction step employs a U-Net [11] style network incorporated with an audio SE-modified encoder and the decoder block. The magnitude of noisy speech STFT of dimension $F \times T$ where $F$ and $T$ are the frequency and time dimensions of the spectrogram, is fed into the network. moreover, the input is provided two convolutional layers consisting of 4 and 2 strides to reduce the time-frequency up till then it reaches to 64. The reduced features are then processed via three convolutional blocks, each with convolution layers consisting with three and one stride, afterwards, a frequency pooling layer is used to reduces the size of the frequency dimension by two. keep in mind that amidst executing of convolutional blocks, the spatial dimension is kept preserved. When the noisy spectrogram is provided as input, the suggested model estimates the clean spectrogram. Using an inverse STFT, the predicted magnitude is blended with the noisy phase to produce enhanced speech.

The decoder consists of 3 up convolutional blocks each consisting of two upsampling layers that upsample the time dimension by 2, followed by convolutional layers with a filter size of 3 and stride of 1. The audio features are then fed to two transposed convolutional layers with filter size of 4 and stride of 2 to upsample the time-frequency dimension, until the time-frequency dimension is equal to the input. Next we use a sigmoid layer to map the output in the range of 0 to 1. The predicted mask is then multiplied with the input spectrogram to generate the masked spectrogram as output.

The proposed model estimates the clean spectrogram when the noisy spectrogram is fed as input. The estimated magnitude is combined with the noisy phase to produce enhanced speech using an inverse STFT.

III. EXPERIMENTS AND RESULTS

A. Experimental Setup

We used a small vocabulary corpus to train our proposed I-O SE model to see how CC-STOI loss function influences entire SE efficiency. The framework’s performance was specifically evaluated using the benchmark GRID corpus [9]. The dataset included a video recordings of clean utterances consisting of 34 male and female speakers, each having 1000 utterances lasting around three-seconds. In this paper, we only utilized the audio data to assess the performance of our proposed framework. The audio data was initially recorded at 48 kHz sampling rate which subsequently resampled to 16 kHz for processing. In this paper, we randomly selected ten speakers for the training set, two speakers for validation, and three speakers for the test sets. To evaluate the performance of the proposed I-O SE framework, we employed three objective evaluation measures to measure the quality and intelligibility of the enhanced speech (i.e., perceptual evaluation of speech quality (PESQ) and STOI) and to measure the distortion in the estimated speech (i.e., speech distortion index (SDI)).

B. STOI vs CC-STOI

Similar to the modified version of STOI as proposed in [12] that accounted 16 kHz signals in the frequency domain while ignoring down sampling and silent frame removal steps, we assessed the correlation between CC-STOI and modified STOI scores. Figures 3(a) and 3(b) show scatter plots for CC-STOI and modified STOI scores computed between clean and noisy utterances, and clean and enhanced utterances. The scatter plots show that the CC-STOI score correlates well with the modified STOI score and can be used for training DL models.

C. Speech Sound vs Non-speech Sound Noises

We first assessed the performance of the proposed CC-STOI framework using a speech sound background noise scenario, where the clean utterances are contaminated with speech sound background noises between 0 to 20 dB SNR. We extended the experiments by considering more challenging background noises by contaminating the clean utterances with the four real-world background noises, namely Bus, Cafeteria, Pedestrian, and Street (termed Bus, Caf, Ped, and Str) provided by CHiME-3 corpus [13] between {-12, 9 dB} SNR at a step of 3 dB.

Table I and Table II list the performance comparison of DL SE frameworks trained using three different loss functions for speech sound and non-speech background noises. It can be seen from Table I and Table II that the DL SE frameworks when trained with different loss functions, enhanced the original noisy speech utterances with a reasonable margin in terms of all performance measures. In short, the frameworks optimized using the three loss functions proved to be effective for SE. We note that in contrast to SE frameworks trained with $L_{MSE}$ and $L_{STOI}$ functions, the DL framework trained with the novel $L_{CC-STOI}$ loss function demonstrated better performance in terms of PESQ, STOI and SDI scores, respectively. In terms of MOS with CHiME3 noises, Fig.4 shows the...
results of three distinct speech augmentation approaches. In comparison to MSE loss and STOI loss, the experimental results demonstrate that the CC-STOI performed better. Furthermore, as compared to MSE and STOI losses, the CC-STOI performed better in lower SNRs. The subjects were normal-hearing listeners and were asked to report the results in terms of mean opinion score (MOS). Fig.4 depicts the MOS score of the GRID corpus for non-speech CHiME-3 background noises at -9, 9 dB SNR at a step of 3 dB using three loss functions. The experimental results show that $L_{CC-STOI}$ achieved better speech perception performance in terms of noise suppression when compared with $L_{STOI}$ and $L_{MSE}$ especially under low SNR conditions.

### IV. Conclusion

In this paper, we proposed a novel canonical-correlation based I-O approach to enhance the training and generalisation performance of conventional DL based SE systems by exploiting a more efficient intelligibility-based evaluation metric as an alternative cost function. Specifically, we developed and utilised a modified canonical-correlation based version of the conventional STOI loss function to train DL SE models, that can more effectively account for signals in the frequency domain as opposed to conventional I-O frameworks that require down sampling signals in the time-domain. Comparative experimental findings show that incorporating canonical correlation as part of a modified STOI-based DL framework can estimate the output signal with enhanced speech quality and intelligibility. In summary, we found that the CC-STOI loss function achieves good general performance and produces better results for a variety of SE evaluation metrics, implying that the modified STOI is a promising choice to optimize frequency-domain SE applications. Ongoing work is aimed at extending and evaluating our I-O SE system to deal with more challenging real-world AV corpora and subjective listening tests for speech and hearing-aid applications.

### REFERENCES

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

[2] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2014.

[3] A. Pandey and D. Wang, “Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain,” in Proc. ICASSP, 2019, pp. 6875–6879.

[4] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, “Raw waveform-based speech enhancement by fully convolutional networks,” in Proc. APSIPA ASC, 2017, pp. 006–012.

[5] 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 Transactions on Audio, Speech, and Language Processing, vol. 26, no. 9, pp. 1570–1584, 2018.

[6] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” in Proc. ICASSP, vol. 2, 2001, pp. 749–752.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125–2136, 2011.

[8] M. Kolbak, Z.-H. Tan, S. H. Jensen, and J. Jensen, “On loss functions for supervised monaural time-domain speech enhancement,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 825–838, 2020.

[9] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421–2424, 2006.

[10] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “Sdr-half-baked or well done?” in Proc. ICASSP, 2019, pp. 626–630.

[11] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.

[12] T. Hussain, M. Gogate, K. Dashtipour, and A. Hussain, “Towards intelligibility-oriented audio-visual speech enhancement,” arXiv preprint arXiv:2111.09642, 2021.

[13] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘chime’ speech separation and recognition challenge: Dataset, task and baselines,” in Proc. ASRU, 2015, pp. 504–511.