

RESEARCH ARTICLE

Speech Enhancement Algorithm Based on a Convolutional Neural Network Reconstruction of the Temporal Envelope of Speech in Noisy Environments

RAHIM SOLEYMANPOUR^{1,2}, MOHAMMAD SOLEYMANPOUR³, (Member, IEEE),
ANTHONY J. BRAMMER¹, MICHAEL T. JOHNSON³, (Senior Member, IEEE),
AND INSOO KIM^{1,2}, (Member, IEEE)

¹Department of Medicine, University of Connecticut School of Medicine, Farmington, CT 06030, USA

²Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06269, USA

³Department of Electrical and Computer Engineering, University of Kentucky, Lexington, KY 40506, USA

Corresponding author: Insoo Kim (ikim@uchc.edu)

This work was supported by the National Institute for Occupational Safety and Health (NIOSH) under Grant R21OH011552.

ABSTRACT Temporal modulation processing is a promising technique for improving the intelligibility and quality of speech in noise. We propose a speech enhancement algorithm that constructs the temporal envelope (TEV) in the time-frequency domain by means of an embedded convolutional neural network (CNN). To accomplish this, the input speech signals are divided into sixteen parallel frequency bands (subbands) with bandwidths approximating 1.5 times that of auditory filters. The corrupted TEVs in each subband are extracted and then fed to the 1-dimensional CNN (1-D CNN) model to restore the TEVs distorted by noise. The method is evaluated using 2,700 words from nine different talkers, which are mixed with speech-spectrum shaped random noise (SSN), and babble noise, at different signal-to-noise ratios. The Short-time Objective Intelligibility (STOI) and Perceptual Evaluation of Speech Quality (PESQ) metrics are used to evaluate the performance of the 1-D CNN algorithm. Results suggest that the 1-D CNN model improves STOI scores on average by 27% and 34% for SSN and babble noise, respectively, and PESQ scores on average by 19% and 18%, respectively, compared to unprocessed speech. The 1-D CNN model is also shown to outperform a conventional TEV-based speech enhancement algorithm.

INDEX TERMS Speech enhancement, temporal envelope (TEV), convolution neural network (CNN).

I. INTRODUCTION

Many applications such as hearing protection devices (HPDs), automatic speech recognition (ASR), and hearing aid devices can benefit from improvements in speech communication in noisy environments. In particular, misunderstanding of communication sounds in noisy workplaces puts workers at risk of injuries. It has been reported that as many as 34% of workers exposed to noise are unwilling to wear an HPD for fear communication will be impeded [1], [2], [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Valentina E. Balas¹.

Thus, it is not surprising that noise-induced hearing loss is one of the most reported, non-fatal occupational injuries [4].

Over the last decades, several speech enhancement algorithms such as spectral subtraction [5], [6], Wiener filtering [7], and statistical models [8] have been developed to suppress noise while limiting distortion of the speech component. However, such speech enhancement algorithms are still limited in their ability to improve speech intelligibility because the speech and noise components are not appropriately distinguished in the presence of noise [9], [10].

Numerous studies in the past have considered temporal modulation processing to improve speech intelligibility in

noisy environments [11], [12]. In temporal modulation processing speech signals are considered to consist of a temporal fine structure (TFS, i.e., carrier signal) and a temporal envelope (TEV) within narrow frequency bands [13], [14]. Several studies have demonstrated the importance of low-frequency temporal amplitude modulation for speech intelligibility by psychoacoustic experiments. Authors in [15] and [16] investigated the relative contribution of TEV and TFS to speech perception and concluded that the TFS provides less important information to speech intelligibility when compared to the TEV. Shannon et al. showed the replacement of the TFS with random noise retains intelligibility, thus confirming that TEV plays a critical role in understanding speech [17].

Noise-induced destruction of the TEV results in the loss of linguistic information. Thus, manipulation of the TEV can be considered potentially an effective approach for enhancing speech quality as well as intelligibility [11], [18], [19], [20]. According to Lezzoum et al., using the TEV as time-varying gains can help improve speech quality in noisy environments [21]. While this method effectively reduces noise, it has limited capability to improve speech intelligibility in noisy conditions, because it cannot detect speech components or correctly reconstruct the TEV [22], [23]. We have previously demonstrated with psychoacoustic experiments that successful reconstruction of the TEV results in significant improvement in speech intelligibility [23].

In recent years, deep learning-based speech enhancement methods have been developed [24], [25], [26], [27]. They are first employed to train the model to achieve the best mapping between noisy speech and clean speech in the spectral domain [52]. The authors in [53] construct the CNN model based on an end-to-end waveform approach. It is shown in this research that using the Mean Square Error (MSE) and STOI as the objective functions will optimize the model to achieve remarkable results. Several studies have demonstrated that convolutional neural networks (CNNs) can be an effective denoising method to improve the audibility and quality of noisy speech [28], [29]. A CNN model based on a loss function to attenuate environmental noise in the time domain has been proposed [30]. As an alternative approach, Yu et al. [31] and Roy and Kumar [32], [49] introduced deep neural network (DNN) methods for optimizing the performance of Kalman filters, by estimating linear prediction coefficients (LPCs) for clean (i.e., noise-free) speech from noisy speech. Most deep learning algorithms for speech enhancement focus on binary masking for noise suppression [33], [34], [35], [36]. In these algorithms, DNNs are trained to identify when speech is dominant in either the time-frequency domain or the frequency domain to reduce background noise [27], [36], [51].

Lan and Tian [50] developed a CNN-based speech enhancement algorithm in the frequency-time domain to identify the appropriate weight for each time-frequency point, resulting in attenuating the high noise speech channels. Similar studies have shown that DNNs could be effective methods to deal with reverberant noise in the acoustical

TABLE 1. Category of main trends of DNNs used for denoising.

Category	Description	Research
Mapping between input and output signal	Using raw data to produce clean speech	[52],[28],[53]
Feature prediction	Predict speech characteristics like LPC to estimate clean speech	[31],[32],[48]
Noise component identifier	Identify the noisy speech component and then attenuate the noise-dominant components by scaling factor	[50],[37],[38]

environment [37], [38]. We summarize and categorize the main trends of machine learning algorithms used for speech denoising in Table 1.

In [23], we demonstrate that improving the TEV, by using the same speech-in-noise mixture with increased SNR and applying it to the original corrupted speech as a time-varying gain, dramatically enhances the speech intelligibility. However, prior knowledge of the corrupted speech with an improved SNR is not usually available in real-world situations. Hence estimates must be made of the TEV of noise-free speech. Here we hypothesize that improving the reconstruction of the TEV of speech degraded by additive noise by using a DNN will result in improved speech intelligibility compared to a typical TEV-based speech enhancement algorithm. To test the hypothesis, the study proposes a hybrid speech enhancement algorithm using a one-dimensional model (1-D CNN) for this purpose. The convolutional part of the model helps to extract the local and temporal information for mapping between the envelopes of clean and noisy speech [41]. Furthermore, since a 1-D CNN requires low computational resources compared to two-dimensional CNNs, it is suitable for wearable applications that require real-time but low-cost processing [39], [40], such as HPDs.

In this study, the lexicon of the Modified Rhyme Test (MRT) with nine different talkers is used as the voice input for training the model, and for objective speech intelligibility and quality evaluations [42]. SNRs of noisy speech ranging from -8 dB to 0 dB are used with two types of noise: speech-spectrum shaped random noise (SSN) and speech babble noise. The former is widely accepted as the most difficult noise for comprehending speech and so presents most challenge to a speech enhancement algorithm. The latter simulates the difficulty understanding the speech of one (target) talker from that of simultaneous (competing) talkers. The study demonstrates the performance of the proposed 1-D CNN-based algorithm using the Perceptual Evaluation

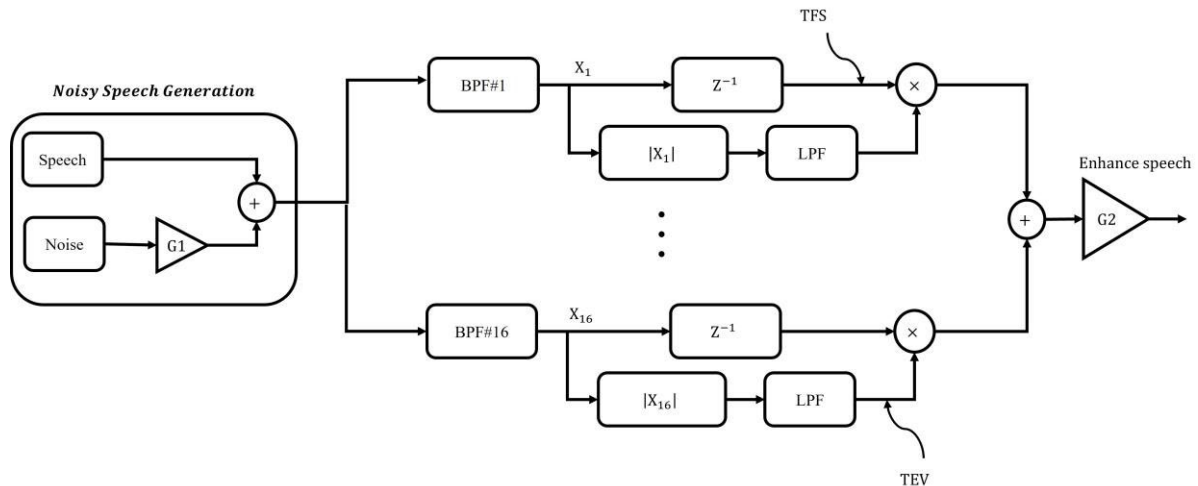


FIGURE 1. Overview of the TEV algorithm.

of Speech Quality (PESQ) [43], and the Short-time Objective Intelligibility (STOI) [44].

In this contribution, we first describe a basic time-frequency domain algorithm that forms the corrupted TEV of speech-in-noise as a means for speech enhancement. We then demonstrate how a 1-D CNN can be introduced into the algorithm to reconstruct a TEV that more closely resembles the envelope of noise-free speech. The errors in the TEVs constructed by the two algorithms are next compared and a qualitative comparison of the effects of the signal processing on speech in additive noise is shown by spectrograms to confirm the improvement. Quantitative comparison of the performance of the two algorithms for improving speech intelligibility and quality is obtained from the STOI and PESQ values, respectively.

II. SPEECH ENHANCEMENT ALGORITHMS

A. TEV-BASED SPEECH ENHANCEMENT ALGORITHM

Figure 1 illustrates the conceptual block diagram of a typical TEV-based speech enhancement algorithm suitable for wearable applications [23]. We first combine the speech and noise by tuning G1 to produce the desired SNRs to be fed to the algorithm. The input speech signals are divided into parallel, contiguous frequency bands (denoted as subbands) spanning the range from 200 to 6,000 Hz using sixteen, 512-order FIR (Finite Impulse Response) band-pass filters. The bandwidth of each subband approximates 1.5 ERB (Equivalent Rectangular Bandwidth of an auditory filter in the cochlea) [19], [45]. In each subband, the band-pass filtered signal is divided into two different paths as shown in Figure 1 – the TFS path that contains the fine structure of the speech and the TEV path in which the time-varying gain is constructed from the temporal envelope.

The TEV is obtained by rectifying and passing the signal through a low-pass filter (LPF). The LPF consists of a 512-order FIR filter with a cut-off frequency of 16 Hz to identify

the most relevant linguistic information [15]. Using an LPF to compute the modulation signal imposes a delay on this path, and so an equivalent delay of 256 samples is added to the fine structure path (denoted as Z^{-1} in Figure 1). The delayed signal is multiplied by the low-pass filtered envelope as a time-varying gain. The same procedure is performed in all subbands. The modified subband signals are then combined to reconstruct the speech. The signal is finally amplified with a gain value chosen to ensure it is audible.

The performance of this TEV algorithm in extremely noisy environments has previously been evaluated in listening tests [23]. The study concluded that the algorithm improves speech quality in a noisy environment but does not significantly improve speech intelligibility compared to unprocessed noisy speech. This implies that additive noise suppresses the speech features in the TEV, and the TEV cannot be reconstructed using simple filters. However, improved speech understanding can be obtained when the SNR of the TEV exceeds that of the TFS [17], [23], [45], [46].

B. 1-D CNN-BASED SPEECH ENHANCEMENT ALGORITHM

This model is used to map the temporal envelope of noisy speech to a “clean” speech in the time-frequency domain. Figure 2 shows the block diagram of the proposed 1-D CNN-based speech enhancement algorithm that reconstructs the TEV using a CNN model. We initially extract the TEV from the noisy speech in sixteen subbands, as in Figure 1, and then feed the subband TEVs to the 1-D CNN model. A sample-by-sample processing technique is used on this platform. The 1-D CNN model is applied to the extracted TEVs without considering the TFS. The output of the 1-D CNN model is sixteen time-varying gains, which are multiplied by their corresponding fine structure to produce the enhanced speech signal.

The 1-D CNN model is a hierarchical neural network consisting of different types of layers stacked together. Like a

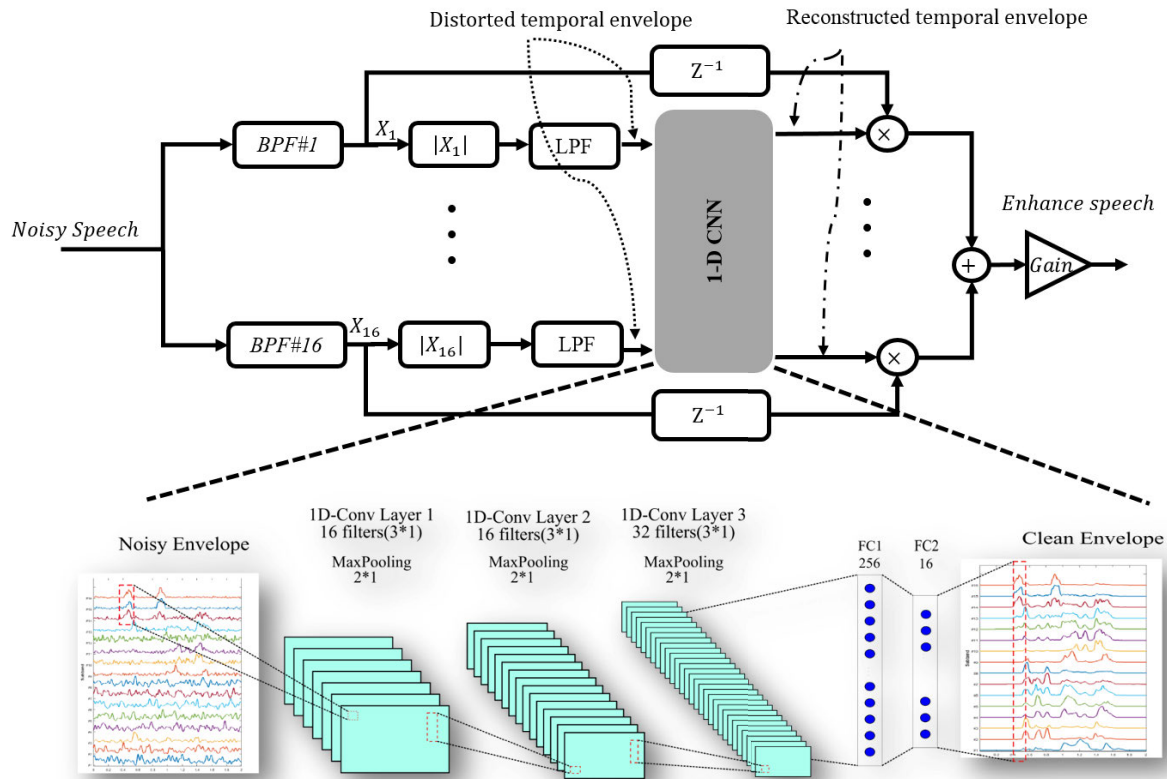


FIGURE 2. Block diagram of the 1-D CNN algorithm.

typical CNN architecture [47], the proposed 1-D CNN model includes convolution, pooling and dropout layers with an activation function. The convolution layers help the model to construct mapping between a noisy temporal envelope and a clean one. This model contains three such layers where each of them is followed by maxpooling and dropout layers. After the last layer of the 1-D CNN, two fully connected (FC) layers with a rectified linear unit (ReLU) activation function are used to reconstruct the TEV signal. The length of the input to the 1-D CNN is 16, indicating the number of frequency subbands used in this algorithm. At each time, these sixteen points corresponding to the modulation signals are convolved with 16 filters of kernel length 3 and stride 1. This layer is followed by the ReLU activation function and a maxpooling layer of size 2. After the maxpooling layer, a dropout layer is applied with a ratio of 0.2. Similarly, the second layer has the same structure and the same length and kernel size for convolution, maxpooling and dropout layers, respectively. However, in the third convolution layer, 32 filters are used with the same kernel size and stride followed by a ReLU activation function, and identical maxpooling and dropout layers. Finally, two FC layers are employed after the output of the convolution process. The first and second FC layers contain 256 and 16 neurons, respectively. The output of the second FC layer reconstructs the modulation signal for each frequency subband at each time.

During training, the TEVs of the noisy and clean speech are passed to the model as an input and desired output, respectively, as shown in Figure 3. The object function was the mean absolute error (MAE) of all subbands. The model was trained for 20 epochs with a batch size of 16. The Adam optimizer with a learning rate of 0.001 was used to train the model [54].

III. METHODS

A. SPEECH MATERIAL

The MRT contains a set of similar-sounding words that are easily misunderstood when presented in additive noise. The word lists were those standardized for American English as spoken by five males and four females [42]. The performance of the proposed 1-D CNN algorithm was established using 2700 utterances from the MRT lexicon of words embedded in a carrier sentence. Utterances of all speakers were recorded in a quiet environment [42]. Approximately 70 percent of the sentences were used to train the model, the remaining sentences being assigned to validate (15 percent) and test (15 percent) the model. The model was trained in a speaker-independent manner, with sentences and talkers used for the training phase excluded from the test phase.

The performance of the TEV and 1-D CNN algorithms was evaluated using two different noises, SSN and speech babble noise, and at different SNRs (0, -2, -4, -6, and

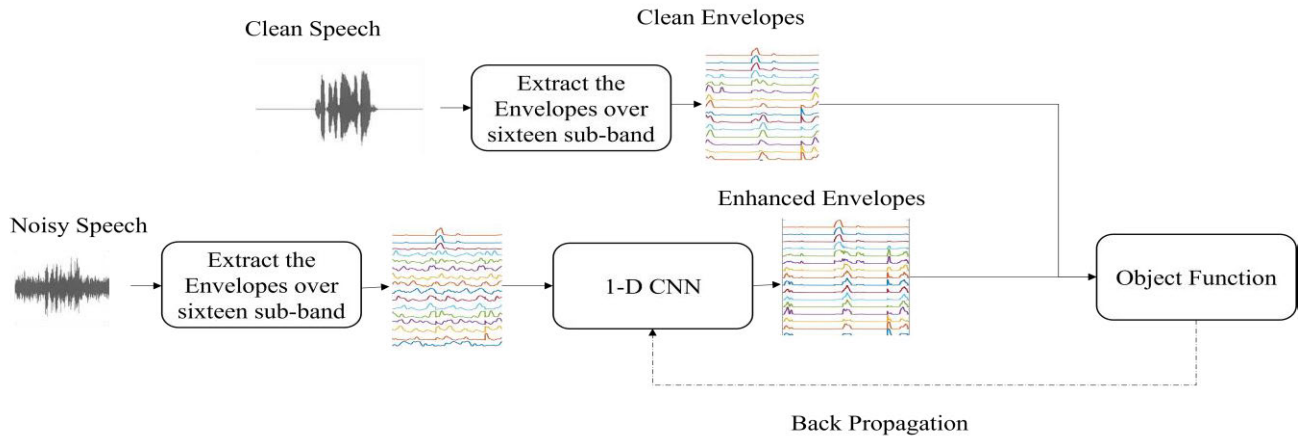


FIGURE 3. Block diagram of procedure for training the 1-D CNN model.

−8 dB). In our previous work [23], we found that SSNs with these SNRs provided a range of speech intelligibility scores from 80% to 35%. Thus, this experimental setup provides a framework for testing the hypothesis in this paper. In other words, these conditions can be expected to cause substantial distortion of the TEV in critical frequency bands, which is a necessary condition to evaluate the performance of the proposed algorithms to recover the TEV.

B. OBJECTIVE METRICS

The success of the proposed 1-D CNN algorithm depends on how similar the restored TEV is to the TEV of speech in quiet. Hence it is appropriate to determine the similarity between the noise-free temporal envelope and the envelope processed in noise. This study used the cosine similarity to assess the similarity between the clean and predicted TEVs. Cosine similarity is a non-dimensional value between 0 and 1 that indicates the similarity between two signals A_i and B_i and is defined as follow [48]:

$$\text{Similarity} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

where n is the total number of samples. We calculated the similarity metric for two different scenarios: 1) the similarity between the clean TEV and the predicted TEV produced by the 1-D CNN model, and 2) the similarity between the clean TEV and the TEV generated by the conventional TEV algorithm. A comparison between these cases informs whether the 1-D CNN model can generate a better temporal envelope potentially to enhance speech quality and intelligibility.

In order to assess the processed noisy speech, two, well-known objective metrics, PESQ and STOI, are used to evaluate the performance of the speech enhancement algorithms. PESQ and STOI evaluate speech quality and speech intelligibility, respectively. The purpose of the PESQ procedure is to model signal processing according to the auditory system, and it predicts speech quality scores on a scale of −0.5 to 4.5. To determine the STOI, a simple discrete Fourier transform is

used for a short time window (approximately 400 ms) in the time-frequency domain. The STOI score ranges from 0 to 1. The PESQ and STOI were determined for three scenarios: A) unprocessed noisy speech, B) noisy speech processed by the TEV algorithm described in Figure 1, and C) noisy speech processed by the 1-D CNN model described in Figure 2.

IV. RESULT AND DISCUSSION

Figure 4 compares the TEV of clean speech in the sixteen subbands with the TEVs reconstructed by the conventional TEV algorithm and the 1-D CNN algorithm from speech in noise. Figures 4(b) and 4(d) clearly demonstrate that the TEV signals are destroyed by additive noise at an SNR of −2 dB in most subbands. In contrast, Figures 4(c) and 4(e) demonstrate that the TEVs are successfully reconstructed regardless of noise type by the proposed 1-D CNN model. The figures illustrate the ability of the CNN model to reconstruct the TEV of speech when the speech is initially “buried” in noise. For example, looking closely at Figure 4, the speech TEV in subband #4 is destroyed by the SSN resulting in lost linguistic information (see Figure 4(b)). However, the speech TEV in subband #4 of figure 4(c) is successfully reconstructed by the 1-D CNN model.

Figures 5 (a) and (b) display the average similarity values as well as standard deviations (SD) for speech in SSN and babble noise, respectively, when processed by the two algorithms. These figures first demonstrate that increasing noise level (i.e., reduced SNR) results in progressively lower similarity between the TEVs of noisy and clean speech. In addition, both figures reveal that the 1-D CNN algorithm performs significantly better than the TEV algorithm regardless of the SNR and noise type (two-sided t-test, $p < 0.05$).

Figure 6 compares spectrograms of a sample utterance under four different conditions. Figure 6(a) shows the spectrogram of clean speech (“circle the bang again”). Figure 6(b) illustrates unprocessed noisy speech when the SNR of the noisy speech is −6 dB and reveals the speech is deeply

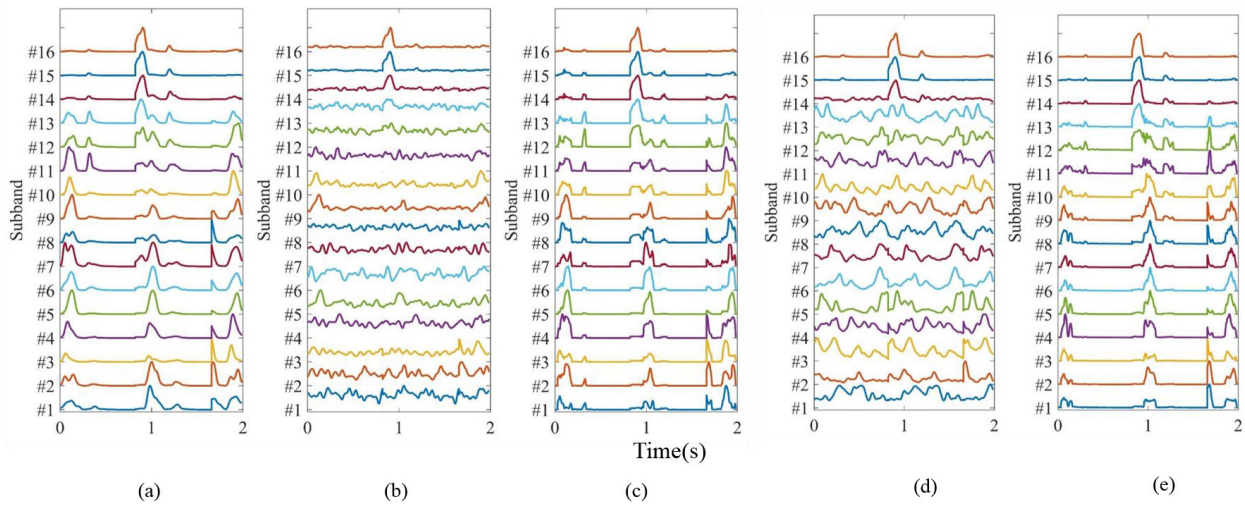


FIGURE 4. Temporal amplitude envelope of (a) clean speech, (b) speech with SSN processed by the TEV algorithm, (c) speech with SSN processed by the 1-D CNN-based model, (d) speech with babble noise processed by the TEV algorithm, and (e) speech with babble noise processed by the 1-D CNN-based model.

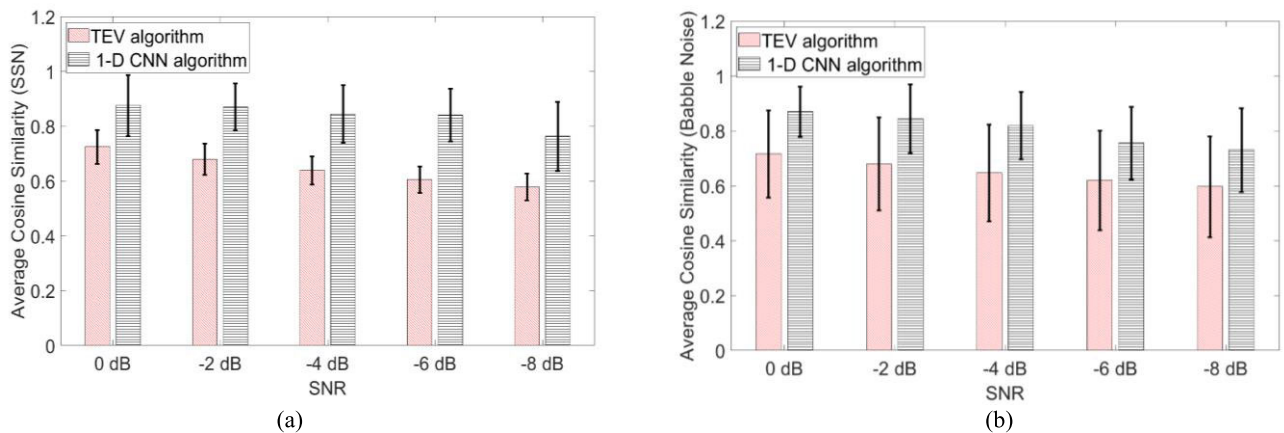


FIGURE 5. Average of cosine similarity measurement over sixteen frequency bands for speech in (a) SSN \pm SD and (b) babble noise \pm SDs compared to clean speech, for the TEV-based algorithm and 1-D CNN based model.

“buried” in SSN. Figures 6(c) and 6(d) show the spectrograms after processing the same words by the TEV and 1-D CNN algorithms, respectively. The 1-D CNN model provides substantial noise suppression compared to the unprocessed speech and the TEV algorithm and eliminates almost all the noise in the entire frequency range. For both speech in SSN and babble noise (not shown), there appears to be less corruption of the spectrograms produced by the 1-D CNN algorithm compared to those for the TEV algorithm.

Figures 7 and 8 illustrate the predicted benefits of the two algorithms on speech intelligibility for SSN and babble noise, respectively. According to the STOI, in the case of SSN (Figure 7), the TEV algorithm improves slightly the score for all SNR values used here (4% improvement on average compared to unprocessed speech). The psychoacoustic experiment performed by Soleymanpour et al., 2021, found the TEV algorithm described in Figure 1 reduced

the speech intelligibility by as much as 2%, hence confirming the prediction of this objective metric. In contrast, the 1-D CNN model, which we have shown can successfully recover the TEV, results in a substantial improvement in the STOI score of 26.7%, on average, for the SNRs used in this study.

In babble noise (Figure 8), the TEV algorithm reduces the intelligibility based on the STOI score (12.1% reduction on average). In contrast, the 1-D CNN algorithm increases the STOI score by 34%, on average.

In both cases, a two-sided t-test of these results confirmed that the 1-D CNN algorithm statistically improves STOI scores in a noisy environment ($p < 0.05$).

The PESQ score for unprocessed speech is compared with those obtained for the TEV and 1-D CNN algorithms in Table 2 (speech in SSN), and Table 3 (speech in babble noise). The TEV algorithm is predicted to improve the PESQ

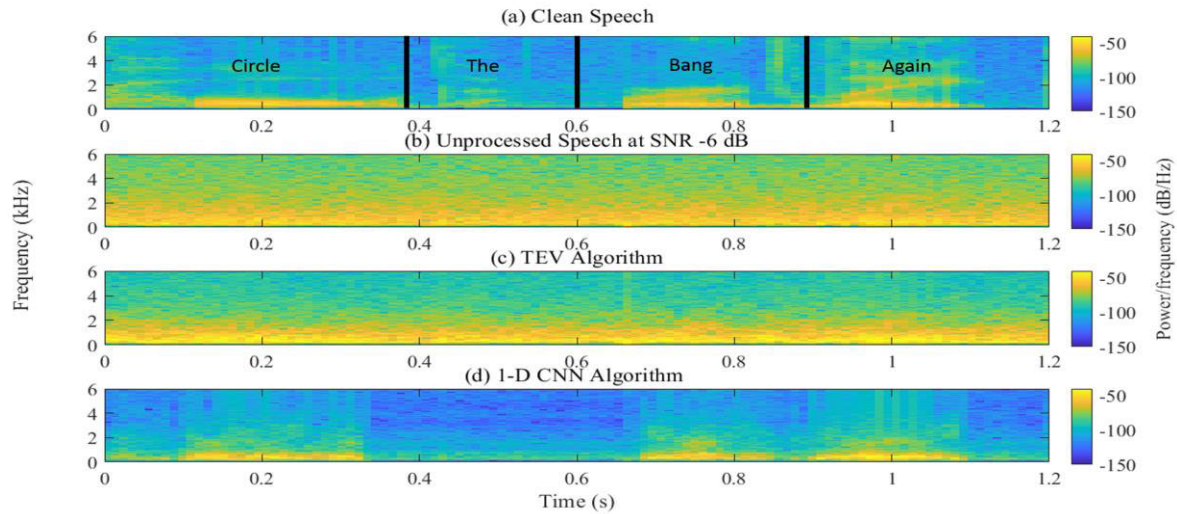


FIGURE 6. Spectrograms of: (a) clean speech, (b) unprocessed speech in SSN at an SNR of -6 dB, (c) noisy speech processed by the TEV-based algorithm, and (d) noisy speech processed by the 1-D CNN model.

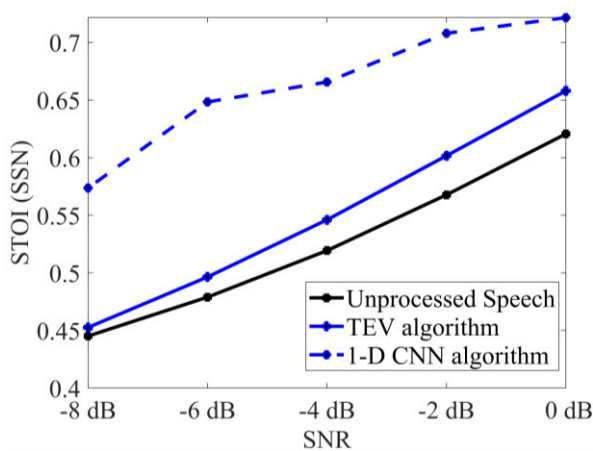


FIGURE 7. STOI scores in SSN at different SNRs.

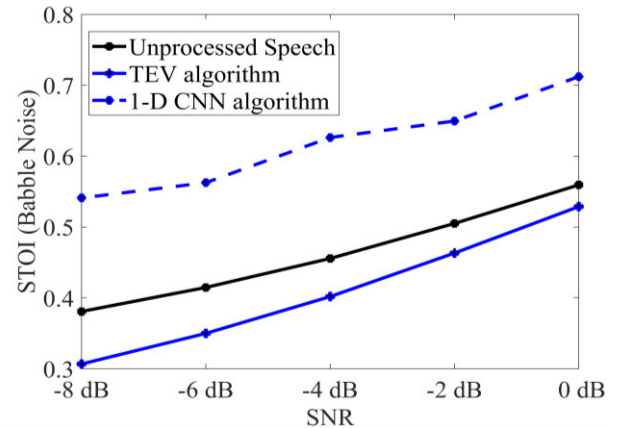


FIGURE 8. STOI scores in babble noise at different SNRs.

score in both cases, by as much as 15% and 6% for SSN and babble noise, respectively. The 1-D CNN model results in an average improvement of 19% and 18% for SSN and babble noise, respectively, which is better than the PESQ score for the TEV algorithm in both noises. The increases in scores are statistically significant (two-sided t-test, $p < 0.05$). As a result, the PESQ score predicts that reconstruction of the temporal envelope using the 1-D CNN algorithm is an effective solution for improving speech quality in noisy environments.

Hence, based on a comparison of PESQ and STOI scores, the 1-D CNN algorithm offers a significant improvement in performance compared to a conventional TEV algorithm in both SSN and babble noise.

The primary purpose of this study was to reconstruct the temporal envelope for speech enhancement applications. This concept can improve speech intelligibility and quality [23],

as has also been found here for our proposed 1-D CNN algorithm that processes the temporal envelope from the sixteen-frequency bands. The reconstruction of the TEV has been shown to be achievable using information in the time-frequency domain.

In other research, Lan and Tian [50] developed a denoising algorithm using a CNN model. They utilize information in the time-frequency domain as well as channel-wise information to construct a more accurate model. They obtained improvements in the STOI score of as much as 10% and 14% (on average) for babble noise and SSN, respectively, when the SNR was less than 0 dB. Comparing these results with those reported here reveals that reconstruction of the TEV by means of the CNN leads to a greater improvement in speech intelligibility score. On the other hand, Lan and Tian [50] showed greater improvement for PESQ (29% and 50% for babble noise and SSN, respectively). These observations are compatible with our conclusion in [23] that

TABLE 2. PESQ scores in SSN at different SNRs.

Noise	Methods	-8	-6	-4	-2	0
SSN	Unprocessed noisy speech	1.24	1.26	1.29	1.34	1.40
	TEV Algorithm	1.29	1.38	1.45	1.55	1.76
	1-D CNN Algorithm	1.40	1.54	1.61	1.64	1.87

TABLE 3. PESQ scores in Babble noise at different SNRs.

Noise	Methods	-8	-6	-4	-2	0
Babble noise	Unprocessed noisy speech	1.24	1.26	1.29	1.34	1.40
	TEV Algorithm	1.25	1.29	1.33	1.48	1.58
	1-D CNN Algorithm	1.35	1.38	1.54	1.62	1.78

enhancing TEV is a critical factor for speech intelligibility, but not for speech quality, in noisy environments. We have the same observations in the work presented by Saleem and Nasir [51] that utilize a Recurrent Neural Network (RNN) to construct spectral masking from the magnitude spectrograms in a noisy environment. This study found 12% and 11% STOI improvements, and 24% and 68% PESQ improvements for babble noise and SSN, respectively. It thus appears that this architecture provides better performance for speech quality as well, while using a 1-D CNN algorithm, which predicts the clean TEV, is more suitable for speech intelligibility purposes.

V. CONCLUSION

This work investigated a deep learning algorithm in the time-frequency domain for application to speech enhancement in noisy environments. The key idea is to train a 1-D CNN to predict the ideal temporal envelope signal in order to improve speech intelligibility and quality. The proposed algorithm has been evaluated in a talker-independent manner for speech in two different noises: SSN and babble noise. The 1-D CNN algorithm is predicted to improve STOI and PESQ scores by an average of 27% and 19%, respectively, for speech in SSN, while the envelope-based algorithm results in improvements of 4% for intelligibility and 15% for speech quality in this noise. For babble noise, STOI suggests that the 1-D CNN obtains an improvement in intelligibility, while the TEV algorithm negatively affects the intelligibility. According to the PESQ scores for babble noise, the 1-D CNN achieved a better result (17%) compared with the TEV algorithm (6%). Consequently, in all these cases, the STOI and PESQ scores indicate substantial improvements in performance are obtained by using the 1-D CNN model. In summary, the findings of this study indicate that the proposed 1-D CNN-based temporal envelope reconstruction model can retrieve the TEV perturbed by environmental noise. This conclusion leads us to consider validating the performance of the proposed algorithm through psychoacoustic experiments.

REFERENCES

- [1] T. C. Morata, A. C. Fiorini, F. M. Fischer, and E. F. Krieg, "Factors affecting the use of hearing protectors in a population of printing workers," *Noise Health*, vol. 4, no. 13, pp. 25–32, 2001.
- [2] R. L. McKinley, V. S. Bjorn, and J. A. Hall, "Improved hearing protection for aviation personnel," New Directions Improving Audio Effectiveness, RTO/NATO Neuilly sur, Seine, France, Tech. Rep. RTO-MP-HFM-123, 2005.
- [3] C. J. Smalt, P. T. Calamia, A. P. Dumas, J. P. Perricone, T. Patel, J. Bobrow, P. P. Collins, M. L. Markey, and T. F. Quatieri, "The effect of hearing-protection devices on auditory situational awareness and listening effort," *Ear Hearing*, vol. 41, no. 1, pp. 82–94, 2020.
- [4] E. A. Masterson, J. A. Deddens, C. L. Themann, S. Bertke, and G. M. Calvert, "Trends in worker hearing loss by industry sector, 1981–2010," *Amer. J. Ind. Med.*, vol. 58, no. 4, pp. 392–401, 2015.
- [5] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. ICASSP*, vol. 4. Princeton, NJ, USA: Citeseer, 2002, pp. 1–4.
- [6] L. Singh and S. Sridharan, "Speech enhancement using critical band spectral subtraction," in *Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP)*, Nov. 1998, pp. 1–4.
- [7] A. A. Montgomery and R. A. Edge, "Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults," *J. Speech, Lang., Hearing Res.*, vol. 31, no. 3, pp. 386–393, Sep. 1988.
- [8] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.
- [9] B. Edwards, "Hearing aids and hearing impairment," in *Speech Processing in the Auditory System*. New York, NY, USA: Springer, 2004, pp. 339–421.
- [10] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.
- [11] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. Adv. Signal Process.*, vol. 2003, no. 7, Dec. 2003, Art. no. 310290.
- [12] S. P. Bacon and D. W. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," *J. Acoust. Soc. Amer.*, vol. 85, no. 6, pp. 2575–2580, Jun. 1989.
- [13] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Amer.*, vol. 77, no. 3, pp. 1069–1077, Mar. 1985.
- [14] T. M. Elliott and F. E. Theunissen, "The modulation transfer function for speech intelligibility," *PLoS Comput. Biol.*, vol. 5, no. 3, Mar. 2009, Art. no. e1000302.
- [15] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, Feb. 1994.

- [16] R. Drullman, J. M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, pp. 2670–2680, May 1994.
- [17] R. V. Shannon, "Speech recognition with primarily temporal cues," *Science* vol. 270, no. 5234, pp. 303–304, 1995.
- [18] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *J. Neurophysiology*, vol. 85, no. 3, pp. 1220–1234, Mar. 2001.
- [19] B. C. J. Moore, *An Introduction to the Psychology of Hearing*. Leiden, The Netherlands: Brill, 2012.
- [20] B. Herrmann, M. J. Henry, S. Haegens, and J. Obleser, "Temporal expectations and neural amplitude fluctuations in auditory cortex interactively influence perception," *NeuroImage*, vol. 124, pp. 487–497, Jan. 2016.
- [21] N. Lezzoum, G. Gagnon, and J. Voix, "Noise reduction of speech signals using time-varying and multi-band adaptive gain control for smart digital hearing protectors," *Appl. Acoust.*, vol. 109, pp. 37–43, Aug. 2016.
- [22] S. Launer, J. A. Zakis, and B. C. Moore, "Hearing aid signal processing," *Hearing Aids*, 2016, pp. 93–130.
- [23] R. Soleymanpour, A. J. Brammer, H. Marquis, E. Heiney, and I. Kim, "Enhancement of speech in noise using multi-channel, time-varying gains derived from the temporal envelope," *Appl. Acoust.*, vol. 190, Mar. 2022, Art. no. 108634.
- [24] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Effect of spectrogram resolution on deep-neural-network-based speech enhancement," *Acoust. Sci. Technol.*, vol. 41, no. 5, pp. 769–775, Sep. 2020.
- [25] D. Hepsiba and J. Justin, "Role of deep neural network in speech enhancement: A review," in *Proc. Int. Conf. Sri Lanka Assoc. Artif. Intell.* Singapore: Springer, 2018, pp. 103–112.
- [26] N. Saleem, "Deep neural network for supervised single-channel speech enhancement," *Arch. Acoust.*, vol. 44, no. 1, 2019, pp. 3–12.
- [27] Y. Zhao, "A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions," *J. Acoust. Soc. Amer.* vol. 144, no. 3, pp. 1627–1637, 2018.
- [28] 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. (APSIPA ASC)*, Dec. 2017, pp. 006–012.
- [29] K. Oostemeijer, Q. Wang, and J. Du, "Frequency gating: Improved convolutional neural networks for speech enhancement in the time-frequency domain," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annual Summit Conf. (APSIPA ASC)*, 2020, pp. 465–470.
- [30] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 7, pp. 1179–1188, Jul. 2019.
- [31] H. Yu, Z. Ouyang, W.-P. Zhu, B. Champagne, and Y. Ji, "A deep neural network based Kalman filter for time domain speech enhancement," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2019, pp. 1–5.
- [32] S. K. Roy, A. Nicolson, and K. K. Paliwal, "A deep learning-based Kalman filter for speech enhancement," in *Proc. Interspeech*, 2020, pp. 1–5.
- [33] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Amer.*, vol. 123, no. 3, pp. 1673–1682, Mar. 2008.
- [34] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Amer.*, vol. 125, no. 4, pp. 2336–2347, Apr. 2009.
- [35] R. Koning, I. C. Bruce, S. Denys, and J. Wouters, "Perceptual and model-based evaluation of ideal time-frequency noise reduction in hearing-impaired listeners," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 3, pp. 687–697, Mar. 2018.
- [36] B. J. Borgström and S. Michael Brandstein, "Speech enhancement via attention masking network (SEAMNET): An end-to-end system for joint suppression of noise and reverberation," in *Proc. IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, Dec. 2020, pp. 515–526.
- [37] E. W. Healy, M. Delfarah, E. M. Johnson, and D. Wang, "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," *J. Acoust. Soc. Amer.*, vol. 145, no. 3, pp. 1378–1388, Mar. 2019.
- [38] N. Saleem, Khattak, and M. Irfan, "Deep neural networks for speech enhancement in complex-noisy environments," *Int. J. Int. J. Interact. Multimedia Artif. Intell.*, vol. 6, no. 1, pp. 84–90, 2020.
- [39] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.
- [40] O. Abdeljaber, O. Avci, M. S. Kiranyaz, B. Boashash, H. Sodano, and D. J. Inman, "1-D CNNs for structural damage detection: Verification on a structural health monitoring benchmark data," *Neurocomputing*, vol. 275, pp. 1308–1317, Jan. 2018.
- [41] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, pp. 611–629, Jun. 2018.
- [42] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Articulation-testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Amer.*, vol. 37, no. 1, pp. 158–166, Jan. 1965.
- [43] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4214–4217.
- [44] 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. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, May 2001, pp. 749–752.
- [45] Y. Oganian and E. F. Chang, "A speech envelope landmark for syllable encoding in human superior temporal gyrus," *Sci. Adv.*, vol. 5, no. 11, Nov. 2019.
- [46] O. Ghizta, "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. Amer.*, vol. 110, no. 3, pp. 1628–1640, Sep. 2001.
- [47] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.
- [48] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.
- [49] S. K. Roy, A. Nicolson, and K. K. Paliwal, "DeepLPC: A deep learning approach to augmented Kalman filter-based single-channel speech enhancement," *IEEE Access*, vol. 9, pp. 64524–64538, 2021.
- [50] T. Lan, Y. Lyu, W. Ye, G. Hui, Z. Xu, and Q. Liu, "Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement," *IEEE Access*, vol. 8, pp. 78979–78991, 2020.
- [51] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access*, vol. 8, pp. 160581–160595, 2020.
- [52] S. Rim Park and J. Lee, "A fully convolutional neural network for speech enhancement," 2016, *arXiv:1609.07132*.
- [53] 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, Language Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.
- [54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

...