Applied Acoustics 190 (2022) 108634

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Enhancement of speech in noise using multi-channel, time-varying gains derived from the temporal envelope

Rahim Soleymanpour [a,b], Anthony J. Brammer [a,b], Hillary Marquis [c], Erin Heiney [c], Insoo Kim [a,b,*]

[a] Department of Medicine, University of Connecticut School of Medicine, Farmington, CT 06030, USA
[b] Department of Biomedical Engineering, University of Connecticut, Storrs, CT 06269, USA
[c] Department of Surgery, Division of Otolaryngology, Head and Neck Surgery, University of Connecticut Health, 263 Farmington Avenue, Farmington, CT 06030, USA

ARTICLE INFO

Article history:
Received 30 August 2021
Received in revised form 2 December 2021
Accepted 8 January 2022
Available online 24 January 2022

Keywords:
Speech intelligibility
Temporal envelope
Noise
Modulation SNR
Wearable system

ABSTRACT

The envelope of speech in noise has been used as a control signal to amplitude-modulate corrupted speech in order to reduce noise and potentially enhance speech quality and intelligibility. In this study, separate speech and noise time histories were mixed at different SNRs to produce time-aligned speech in noise with a range of SNRs. The temporal envelope of the control signal was manipulated by selecting mixtures with the same, or different, SNRs as the signal being controlled. Signals from 200 Hz to 6 kHz were processed into sixteen, parallel contiguous subbands with bandwidths approximating 1.5 times that of auditory filters. The subband envelopes were formed from the absolute value of signals and low pass filtered at 16 Hz. Eleven subjects aged 29 ± 8 years (mean and range) with normal hearing underwent the Modified Rhyme Test to assess speech intelligibility. Speech quality was assessed objectively by PESQ. When the SNRs of the controlling signal and the speech in noise being controlled were the same, no improvement in speech intelligibility was obtained. When the SNR of the controlling signal exceeded that of the signal being controlled, statistically significant increases in mean word scores of up to 36% could be obtained, hence providing an estimate for the maximum improvement in speech intelligibility achievable by reducing noise in the control signal. There were also small improvements in speech quality.

1. Introduction

The presence of noise is a major cause of poor speech intelligibility in face-to-face communication. In the workplace, excessive noise can cause misunderstandings in oral communication that puts workers at risk of injuries. Numerous devices for applications as diverse as mobile phones, communication headsets, hearing protective devices (HPDs), and hearing aids would benefit from the capability to improve speech intelligibility in a noisy environment. All these wearable devices have limited on-board computational capacity: the solutions to reduced intelligibility of speech in noise explored in this contribution have been developed for such circumstances. Hence, deep neural networks, which have been recently employed to separate speech from noise and have demonstrated some improvement in intelligibility [14,17,31], are not considered here, as the computational cost would be unacceptable for non-networked wearable systems.

Over the past decades, numerous studies have demonstrated the importance of temporal modulation processing in the auditory system [36,3,34,10,2,29]; and [18]. Physiological and psychoacoustic experiments have been conducted to demonstrate the importance of amplitude modulation on speech intelligibility, especially at low modulation frequencies [21,20,4,11]. Smith et al. [38] showed that the temporal envelope of speech carried information related to speech perception, while the details of the time history, or "fine structure", provided mainly pitch perception and sound localization. They also showed that the contribution of the envelope to speech intelligibility increases as the number of frequency bands increases. Houtgast et al. [19] concluded that low-frequency modulation between 0.5 and 16 Hz carries the essential information in speech. This was confirmed in listening tests by Drullman et al. [13], who could reduce the temporal modulation to a frequency range from 0 to 16 Hz while maintaining speech intelligibility.

Studying the effectiveness of the temporal envelope on improving speech degraded by noise has been the subject of considerable research [39,6,30]. Several speech enhancement algorithms involving temporal envelopes have been developed in the

* Corresponding author at: Tel.: +1 860 679 6174
E-mail address: ikim@uchc.edu (I. Kim).

https://doi.org/10.1016/j.apacoust.2022.108634
0003-682X/

frequency-time domain to enhance speech quality and intelligibility [26,8,24]. The problem becomes more challenging when the noise level is high, causing the intelligibility of most components of speech to be compromised. Hence, equipping current electronic HPDs with an appropriate algorithm for improving face-to-face communication presents challenges [7,5,37,23].

Lezzoum et al. [27] have developed a method based on generating the instantaneous magnitude of intensity modulations in sixteen, parallel, contiguous frequency bands (or subbands) to improve speech quality when attempting to communicate face-to-face in a noisy environment. More importantly for the present work, they report this algorithm is suitable for application to an HPD. Other applications of time-varying subband gains to speech processing have been described primarily for hearing aids ([40]; [28,9]). The algorithms may reduce environmental noise but do not generally seem to improve speech intelligibility [25].

While past experience does not suggest that time-varying, envelope modulation-derived subband gains can substantially improve speech intelligibility, studies employing this processing have not demonstrated the potential that could be achieved by improving the signal-to-noise ratio (SNR) of the envelope. It is essential to establish the potential effectiveness of temporal envelope modulation before attempting to develop any related algorithms for speech enhancement. This is the purpose of the present contribution.

The study tests the hypothesis that increasing the SNR of the temporal envelope of speech in noise can produce time-varying subband gains that improve speech understanding. For this purpose, separate speech and noise signals are mixed in the time domain at different SNRs to produce time-aligned speech in noise with a range of SNRs. We manipulate the temporal envelope by selecting two of these signals, one to construct the temporal envelope to serve as a control signal and the other to be the signal being controlled (referred to here as the "data" signal). The signals contain the same speech, but the SNR of the control signal can be increased compared to that of the data signal. This process of changing SNRs in the control path separately from the data path allows time-varying gains to be constructed that better reflect the speech modulations in the mixture.

A preliminary investigation is conducted using identical signals in the data and control signal paths to form a temporal envelope modulation (TEM) algorithm and confirm the results of previous work. The results of experiments, which include listening tests, are described after an explanation of the signal processing.

2. Methods

2.1. Temporal envelop modulation algorithm

Algorithm Description: A block diagram of the essential elements of the algorithm is shown in Fig. 1, which was simulated using MATLAB (R2018b MathWorks, MA, USA). The platform was designed to impose modest computational requirements on any host, which is a most desirable feature for wearable applications such as HPDs. The signals are simulated using a sample-based processing method (sampling frequency 12 kHz).

The input noisy speech is initially divided into sixteen, parallel, contiguous subbands that span the frequency range from 200 Hz to 6000 Hz. The bandwidth of each subband is approximately 1.5ERB, where ERB is the equivalent rectangular bandwidth of an auditory filter in the cochlea [29]. A 512-tab finite impulse response (FIR) band-pass filter is designed for each of the sixteen subbands. The subband filters are designed with 1 dB passband ripple, 60 dB attenuation, and the transition from passband to stop band occurring at a rate of 48 dB/octave. The filters have the same group delay regardless of their cut-off frequencies.

In each subband, the filtered signal is divided into two different paths as shown in Fig. 1 – a data path (DP) that contains the speech (and noise if present) to be processed, and a control path (CP) in which the time-varying gain is to be constructed from the temporal envelope. For each subband, the signal in the control path amplitude modulates the gain of the data path to process the speech. In the control path, extraction of the envelope of the input signal is followed by low-pass filtering (LPF) to produce an instantaneous gain adjustment. The LPF consists of a 512-tab FIR filter with a cut-off frequency of 16 Hz, and with nominally a 1 dB pass band ripple, 48 dB attenuation, and transition from pass band to stop band occurring at a rate of 12 dB/octave. Owing to the delay of the LPF in the control path, the signal in the data path must pass through a matching time delay to synchronize with the control signal, after which they are combined to produce a modified subband signal. Next, the modified signals from all subbands are recombined to form the processed output, shown by the Σ in Fig. 1, and the reconstructed signal is adjusted in gain as necessary (G2). The complete algorithm is then repeated, leading to a time-varying gain being applied to each subband, and time-varying reconstruction of the original speech (and noise if present) that amplifies the processed sounds identified. Finally, the reconstructed sounds are saved as .wav files and can be subsequently replayed through headphones or an audio system to a listener.

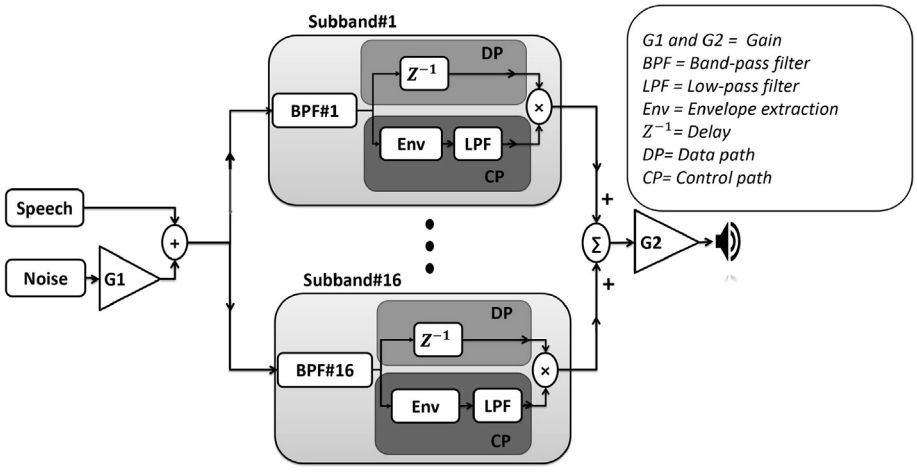

Fig. 1. Concept block diagram of temporal envelop modulation (TEM) algorithm.

Experiment: The first aim of the study was to analyze the effects of the speech enhancement algorithm described in Fig. 1 on speech intelligibility and quality in a noisy environment. Speech-spectrum shaped random noise (SSN) was applied to the input speech to produce SNRs of 0, −3, −6, and −9 dB. Our experience shows that these noise levels can significantly reduce speech understanding. To assess the performance of the algorithm, its output (processed speech) is compared with unprocessed speech. "Unprocessed speech" means here no processing is applied to the input speech and SSN except limiting the frequency range to between 200 Hz and 6 kHz to be compatible with the "processed speech". Nine audio files were prepared in this order: four for unprocessed speech, four for processed speech. An additional audio file that contains noise-free speech was presented to subjects before starting a listening test in noise to familiarize them with the procedure. It is also important to confirm that the manipulation of speech illustrated in Fig. 1 does not affect the intelligibility of speech-in-quiet. Therefore, a separate listening test was conducted without SSN.

2.2. SNR manipulation of temporal envelope

Algorithm Description: Fig. 2 shows how various SNRs are generated for the control path with a preselected fixed SNR for the data path. The two sources contain identical speech and noise time histories and are inputted separately. Signals constructed for and feeding the data path are shown by continuous lines, and signals constructed for the control path are shown by dashed lines. Selecting different values for G1 and G2, with G1 not equal to G2, enables us to provide independent SNRs in the control and data paths. The platform includes 32 band-pass filters (sixteen each for the control and data paths) arranged in parallel and sixteen low-pass filters. These create two sets of 16 subbands with matching characteristics that are identical to the subband filters in Section.2.1. G3 is used to adjust the output sound level.

The algorithm was also simulated using MATLAB (R2018b MathWorks, MA, USA). It should be noted that this platform was developed to address the potential of the temporal envelope to improve speech intelligibility in the presence of noise. It is not suggested to be a practical solution for face-to-face communication in environmental noise, as speech and noise will not be available separately.

Experiment: Table 1 shows the conditions used in this study, where different SNRs were applied to the control path while the SNR of the data path was kept constant. Thus, for example, when the SNR of the data path was −9 dB, the SNR of the control path, was either 0, −3, −6, or −9 dB. We believe using an SNR for the control path that is less than the data path is not relevant to our experiment, because the objective is to establish the increase in word scores from improving the temporal amplitude modulation of speech perturbed by noise. For example, when the SNR of the data path is −3 dB, setting the SNR of the control path to −6 dB or −9 dB would not be relevant.

Table 1
Measurement conditions manipulating the temporal envelope for listening tests.

		Control Signal SNR			
		0 dB	−3 dB	−6 dB	−9 dB
Data Signal SNR	SSN	✓	✓	✓	
	−9 dB	✓	✓	✓	✓
	−6 dB	✓	✓	✓	
	−3 dB	✓	✓		
	0 dB	✓			

SNRs of 0, −3. −6 and −9 dB were selected for the data path. The experiment also contained a set of tests in which unprocessed noisy speech in the data path was replaced by SSN. In addition to the main purpose of the experiment, we demonstrate the relative importance of fine structure and amplitude modulation to speech intelligibility. As in the previous experiment, an additional file containing noise-free speech was included for subject familiarization.

2.3. Subjective assessment of speech intelligibility

To evaluate the performance of the platforms on speech intelligibility, a series of controlled psychoacoustic tests was performed with human participants. Eleven adults (age: 29 ± 8 years) with normal hearing, who spoke English as their first language, were recruited for the study. Participants were comfortably seated in an audiometric booth and wore insert earphones (E·A·RTone type 3A). All participants gave their informed consent to participate in the study and were paid for their time. The study protocol was approved by the Institutional Review Board of UConn Health (Farmington, CT), where the measurements were conducted.

The Modified Rhyme Test (MRT) was used [1,22]. The speech was that of a male talker (Auditec, Inc., St.Louis). Each trial contained the carrier phrase ("Circle the ... [insert test word] ... again") and a test word chosen randomly from six alternate "rhyming" words (e.g., "kit", "bit", "fit", "hit", "wit", or "sit"). Participants circled their answer on a printed word list containing the six words for each trial. The test audio files were constructed

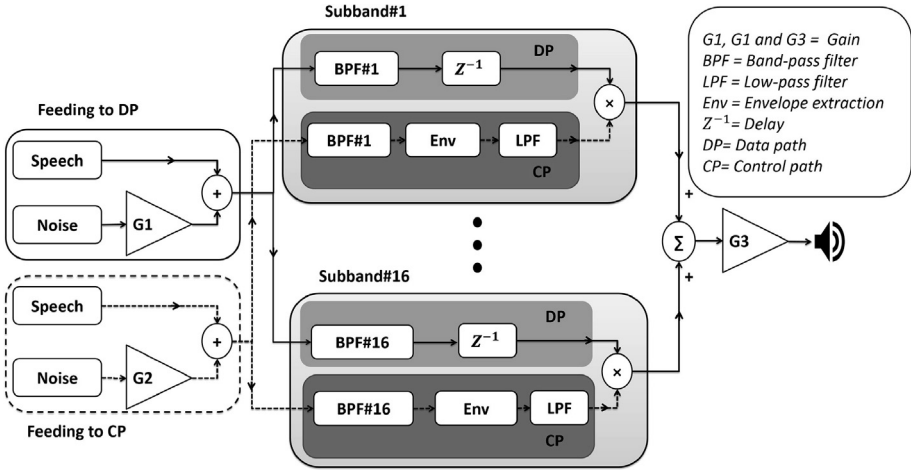

Fig. 2. Concept bock diagram for temporal modulation with different SNRs in control and data paths.

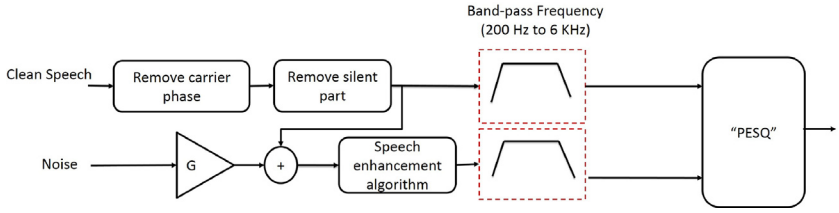

Fig. 3. Prediction of speech quality using PESQ.

from the MRT sentences and SSN. SSN was selected for the experiments as the similarity between the long-term average speech spectrum and that of the noise will present most challenge in a listening test. The signals were processed off-line by MATLAB.

Speech intelligibility was based on the number of correct answers from the six alternative-choice words presented to the listener. A series of 25 trials each consisting of the carrier sentence and a target word was presented to participants through an audio signal processor and amplifier (Tucker-Davis Technologies, Inc., FL, USA). A two-second pause between successive trials was provided for participants to respond. An adjustment needs to be made to the raw word score to account for a subject guessing the target word. To quantify this adjustment, ANSI standard S3.2, 1982 provides the following expression:

$$S_A = S - \frac{R}{n-1} \tag{1}$$

where represents the number of correct answers after adjusting for guessing, S and R are the raw score and the number of incorrect answers, respectively. The number of choices for each trial is represented by n. We will use as a measure of intelligibility for each test.

2.4. Objective estimate of speech quality and calculation of SNR

The quality of the processed speech was predicted using the Perceptual Evaluation of Speech Quality (PESQ) [32]. PESQ attempts to model the signal according to the auditory system and has been shown to possess an acceptable correlation with subjective assessment for telephony [32]. The numerical scores generated by PESQ can be approximately mapped into subjective assessments of speech quality according to the five-point Mean Opinion Score as: bad (∼1.7), poor (∼2.5), fair (∼3), good (∼3.5) and excellent (∼4.5) [33]. A numerical increase in PESQ is associated with an improvement in speech quality within each category.

Only the target words of the MRT were considered to compute the SNR for the speech in noise experiments, and for the speech quality predictions. To remove the carrier words, silent parts of the utterance were identified. The algorithm for this purpose employed a time window of 10 ms. If the maximum signal amplitude in a time window is less than a threshold, this window was considered to be "silent" (i.e. does not contain speech). The threshold value was based on the clarity and quality of the speech evaluated by three listeners after removing the silent windows. The experimental value for the long-term root mean square (RMS) amplitude of speech in this study was 0.01 of the maximum signal amplitude. After removing the silent windows, the isolated target speech signal was passed through 16 band-pass filters identical to those described in Section. II.A, and hence produced speech sounds with frequencies between 200 Hz and 6 kHz. Fig. 3 describes this preprocessing of the input signal before the objective assessment. The PESQ score of the MRT words in quiet calculated in this way was 4.5, so they are judged to be of excellent speech quality.

3. Results

3.1. Temporal envelop modulation algorithm

Examples of the time histories of speech in noise for processed and unprocessed speech when the data and control paths have the same SNR are depicted in Fig. 4. Panels to the left of Fig. 4 are for unprocessed speech in noise, and illustrate the "burying" of speech as the combined signal becomes more dominated by noise. The processed speech-in-noise waveforms are shown in the panels to the right of the diagram. Although temporal modulation can apparently distinguish the speech from the noise, and clearly improves the SNR, its effect on speech intelligibility is best considered from the results of listening tests.

Before considering the performance of the algorithm described in Fig. 1 for speech in noise, it is important to establish the performance of the algorithm for noise-free speech. Therefore, speech without noise was processed by the algorithm and presented to subjects. The mean word score and standard deviation (SD) for processed speech was 98.8 ± 2.2%, which is statistically indistinguishable from the mean word score for unprocessed speech without noise (100%) ($p = 0.18$, two-sided t-test). Hence, it may be inferred that manipulation of the speech signal by the algorithm introduces insufficient degradation of linguistic information to influence speech intelligibility in the absence of noise.

The mean correct word scores and SDs for unprocessed and processed measurement conditions are shown in Fig. 5 for the different SNRs. The unprocessed results display the magnitudes of the word scores at the selected SNRs (28 ∼ 87%) and allow us to study the influence of temporal envelope modulation over a range of speech intelligibility in noise. It is also evident from Fig. 5 that the TEM algorithm does not result in any improvement in speech understanding and even reduces intelligibility in this noise when the SNR is −6 dB. Analysis of the results obtained from the unprocessed and processed speech at various SNRs is performed using a two-way ANOVA (Analysis of Variance) with replication, which confirmed that the frequency-time modulation algorithm shown in Fig. 1 does not statistically improve speech understanding in SSN [$F(1,80) = 1.27$, $p = 0.26$]. These observations are not unlike the conclusions of [8], who found temporal envelope modulation produced little improvement of speech intelligibility in noise.

Although the main purpose of this study focuses on speech intelligibility, it is worthwhile to consider speech quality. Fig. 6 compares the quality of the unprocessed and processed speech in noise as predicted by PESQ. It is evident from these results that the increased PESQ scores indicate a positive effect of the signal processing on speech quality. However, all conditions would be judged subjectively to be of bad quality. For comparison, the PESQ score of the processed speech in the absence of noise was 2.9, which would be judged to be of fair quality.

In summary, for speech in SSN the speech enhancement algorithm described by Fig. 1 has been shown to improve the SNR and possibly speech quality, but not intelligibility.

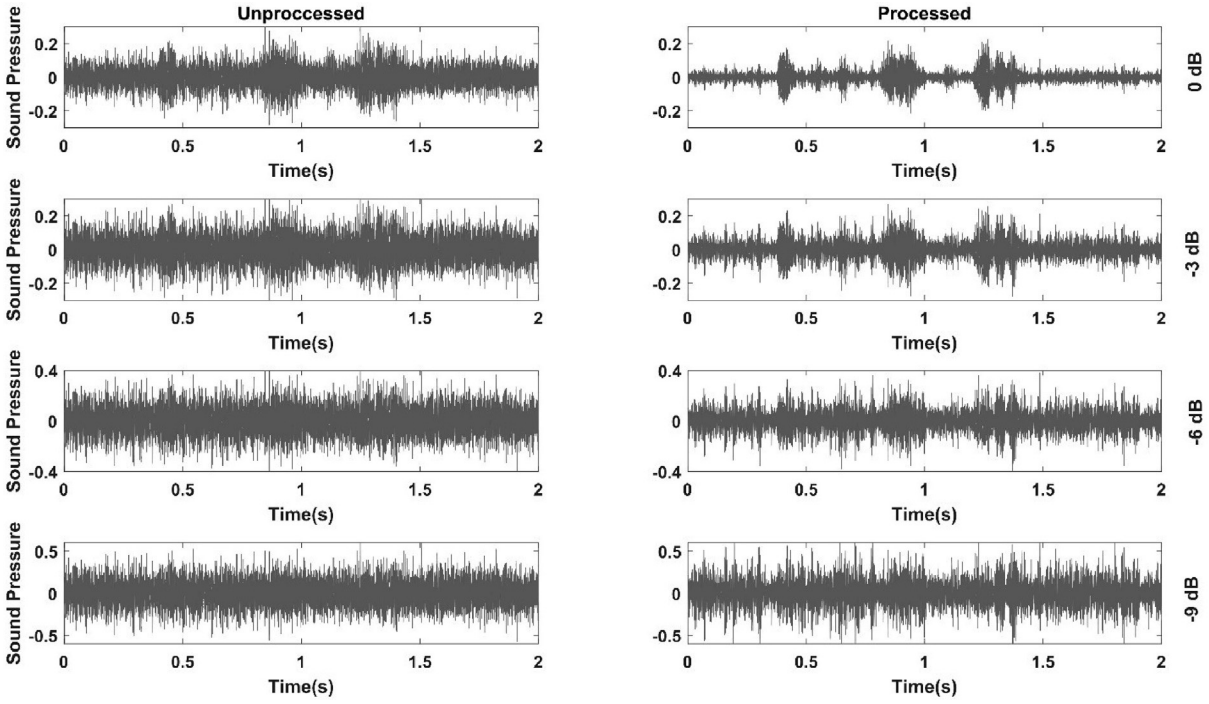

Fig. 4. Unprocessed and processed speech-in-noise waveforms at SNRs of 0, −3, −6, −9 dB when the data and control paths have the same SNR.

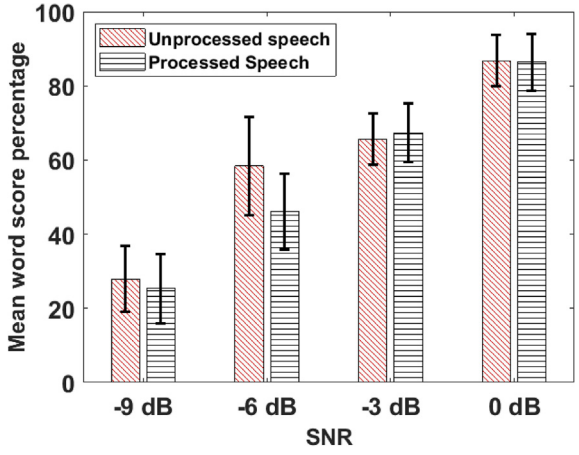

Fig. 5. Mean percentage correct (±SD) word scores of processed and unprocessed speech in SSN when the data and control path have the same SNR.

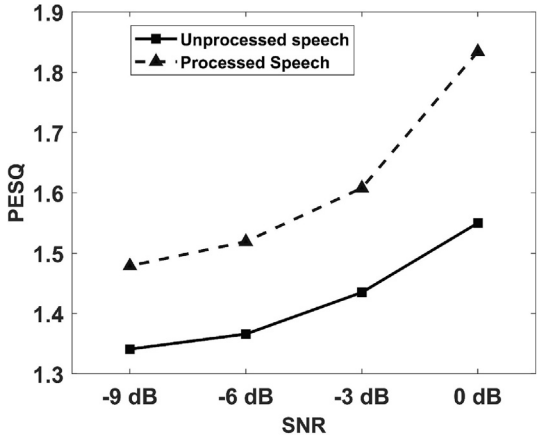

Fig. 6. PESQ assessment of unprocessed and processed speech quality when the data and control paths have the same SNR.

3.2. Manipulation of temporal envelope

The purpose of this experiment was to test the hypothesis that improving the temporal envelope of speech in noise can produce time-varying subband gains that increase speech understanding. As already described, increasing SNRs in the control path can be utilized to provide the improved gains.

Examples of a speech-in-quiet waveform and an unprocessed speech-in-noise waveform when the SNR is −3 dB are depicted in Fig. 7A and B, respectively. Fig. 7C and D show the processed speech in noise when the control path SNR is 0 and −3 dB respectively, while the SNR of the data path remains −3 dB in both cases. Apparently, having more appropriate gains for each time sample within each subband produces an output waveform that further reduces the noise.

As an example of the effectiveness of envelope-derived time-varying gain to enhance speech intelligibility, we first consider the performance of the method described in Fig. 2 when noise-free speech is fed into the control path and pure SNN is fed into the data path. In this experiment, there is no speech information in the fine structure (i.e., data signal) to contribute to speech intelligibility: the only potential source of information is the control signal containing the speech envelope. The mean word score obtained was 94.2 ± 6.3%, and should be compared with the score obtained when noise-free speech is fed into the data path and processed by the algorithm (98.8 ± 2.2% – see previous experiment). These word scores are not significantly different (two-sided *t*-test, $p > 0.05$). Previous studies using noise instead of speech for the fine structure have reported excellent word identification under similar conditions (e.g., [35]), which indicates the importance of the temporal envelope for preserving linguistic information. In other words, when the gain of the envelope of the time history is correct, speech understanding is preserved regardless of other factors, such as details of the fine structure.

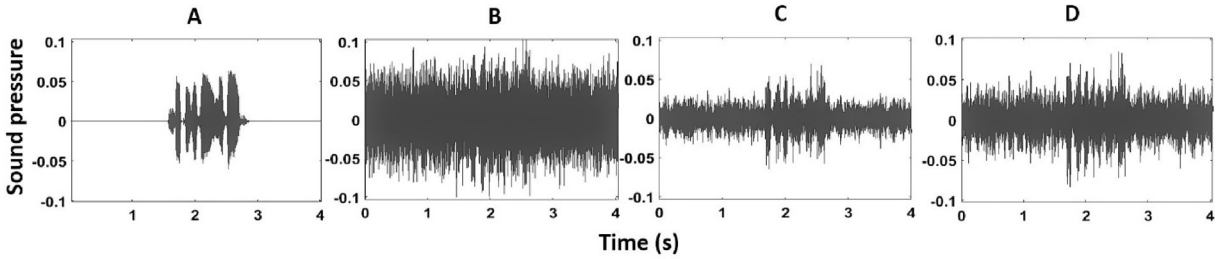

Fig. 7. Example of waveforms for speech, and speech in noise, unprocessed and processed. A — Speech in quiet; B — Unprocessed speech in noise, SNR = -3 dB; C — Processed speech, SNR of data path −3 dB, SNR of control path 0 dB; D — Processed speech, SNR of data path −3 dB, SNR of control path −3 dB.

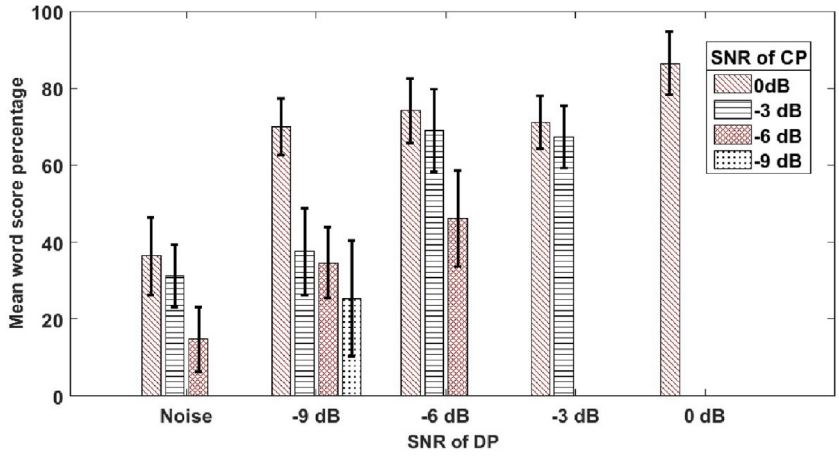

Fig. 8. Mean percentage correct word scores (±SD) for different control path SNRs applied to the data path. Results are shown when the data path (DP) contains speech in noise at selected SNRs that is time aligned with the control signal, and when the DP contains only noise.

When the fine structure consists of speech, or random noise, in SSN, the mean percentage correct word scores and SDs for all measurement conditions in Table 1 are shown in Fig. 8. The abscissa displays the results according to the SNR of the data path. For each condition listed on the abscissa, the mean percent correct score and SD are shown for the SNR of the envelope imposed by the control path. Without exception, providing better time-varying gain by improving the SNR of the control path increases the mean word score for all measurement conditions. It can be seen from these results that there is a relationship between improving the time-varying gain, by increasing the SNR of the temporal envelope, and word score regardless of the SNR or content of the data path. A one-way ANOVA conducted separately for each data path SNR confirms that the SNR of the temporal envelope statistically improves speech understanding in SSN for all categories, except for class #4 (Table 2).

Fig. 9 displays the quality of the processed speech in noise as predicted by PESQ, for the combinations of signal path and control path SNRs in Table 1. Seemingly, time-varying gain based on the temporal envelope can increase PESQ values especially when the SNR of the data path is high, for example, when the data path SNR is −3 dB. But the metric does not predict a significant effect of control path SNR on speech quality at very low SNRs. Thus, when there is an SNR of −9 dB in the data path, changing the SNR in the control path does not improve the quality of speech according to PESQ predictions until it reaches 0 dB.

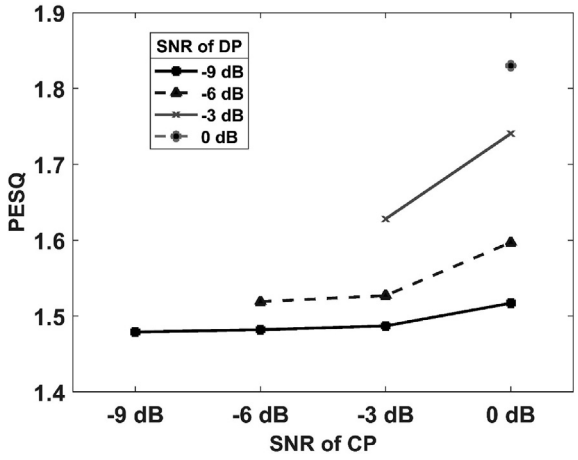

Fig. 9. PESQ assessment of speech quality when the data and control paths may have different SNRs.

Table 2
One-Way ANOVA for each class defined by the SNR of the control path.

Class	Data path	Control path	p-value
#1	SSN	{0,−3,−6} dB	7.99E06
#2	−9 dB	{0,−3,−6,−9} dB	3.26E−12
#3	−6 dB	{0,−3,−6} dB	1.06E−05
#4	−3 dB	{0,−3} dB	0.246

4. Discussion

In this research, we have considered the potential of using multi-channel time-varying gains based on the temporal envelope

to enhance speech communication in noise. We conducted two experiments to determine different aspects of the performance of the platforms. The first experiment, in which the SNR of the control signal was the same as that of the signal being controlled, found that temporal envelop modulation does not present an acceptable method for increasing the score of confusing words in noise, especially when the score for the unprocessed speech was about 60 percent. To the best of our knowledge, other studies in the literature have not reported significant improvement in intelligibility for similar algorithms ([8]; van Buuren et al., 1999; [12]). However, many speech enhancement algorithms as well as the platform in this study improve speech quality [27]. As a result, an alternative method is needed to improve speech comprehension for wearable devices used in an environment where speech is difficult to understand and, more importantly, it must be a practical solution for real-time application.

The results of the second experiment clearly demonstrate improved speech intelligibility can be obtained when appropriate gains are provided to time histories by manipulating the temporal envelope. In this study, we manipulated the SNR of the temporal envelope. Fig. 8 shows that the increase in mean word score can be as much as 36%. Improved speech understanding compared to an initial condition in which the SNRs of the data and control paths were identical was observed for all SNRs. The improvement was statistically significant for all initial conditions except for an SNR of −3 dB. The results also indicated that any degree of improvement in the gain derived from the SNR of the temporal envelope would have a positive effect on speech intelligibility in a noisy environment. When, for example, the SNR of the data path is −6 dB, increasing the SNR of the control path from −6 dB to −3 dB results in an improvement of more than 20% in word score. The current study also suggests that performing a multi-pass gain for each sample in each subband without having a perfect temporal envelope may enhance speech intelligibility for speech extensively disrupted by additive noise.

That temporal envelopes are critical to intelligibility is evident from the results of the second experiment when the data path contains only SSN. As with data paths containing speech in noise, increasing the SNR of the control signal substantially increases the mean word score, and the increases are statistically significant. Also, use of a control signal with an SNR of −6 dB can be seen in Fig. 8 to produce a mean word score of 14.7%, which is substantially greater than that obtained with chance responses to each trial (zero percent). Changing the SNR in the control path from −6 dB to 0 dB improves the mean score by approximately 18 percent. This is further evidence that improving the temporal envelope is crucial for increasing speech understanding in noise. The use of noise instead of the fine structure of speech in signal processing has found application in vocoders and cochlea implants [16,28,15].

Although this study is mainly focused on speech comprehension, it is useful to analyze the quality of speech produced by the algorithms. The first experiment shows that the TEM algorithm somewhat improves speech quality as assessed by PESQ and performs better for high SNR values. In the second experiment, speech quality was found to improve with increasing SNR of the temporal envelop. However, the improvement at low SNR values in the control path was negligibly small. For example, comparing the values of PESQ for SNRs of −3 dB and −9 dB in the control path shows that improving speech modulation in the control path when the SNR of the data path is low (e.g., −9 dB) does not have a meaningful effect on speech quality (see Fig. 9). Therefore, it seems that improving the SNR of the modulation of speech in noise would not be a promising method for enhancing speech quality at low SNRs. Hence, it will be necessary to investigate other methods to deal effectively with this issue. Finally, it can be concluded that the performance of temporal modulation as applied here to speech intelligibility is not the same as its effect on speech quality.

5. Conclusions

In this study, we have examined a potential low-latency approach for improving the intelligibility of speech in noise. We explored employing multi-channel time-varying gains constructed from the temporal envelope for speech enhancement. An initial experiment, in which the SNR of the control signal was the same as that of the signal being controlled, did not result in improved intelligibility, and confirmed work published elsewhere. A second experiment showed that manipulating the temporal envelope by applying an improved SNR to the control path compared to that of the data path increased word scores by up to 36%. Temporal processing was also predicted to improve speech quality, but there would be little improvement experienced subjectively. We expect these findings to enable us to develop robust speech enhancement algorithms for real-time applications in wearable devices.

CRediT authorship contribution statement

Rahim Soleymanpour: Methodology, Investigation, Formal analysis, Writing – original draft, Visualization. **Anthony J. Brammer:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition. **Hillary Marquis:** Investigation. **Erin Perez:** Investigation. **Insoo Kim:** Writing – review & editing, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Institute of Occupational Safety and Health under grant R21OH011552.

References

[1] ANSI S3.2-1960 (R1982), "Method for measuring the intelligibility of speech over communication systems," American National Standards Institute, New York.
[2] Atlas L, Shamma SA. Joint acoustic and modulation frequency. EURASIP J Appl Signal Process 2003:668–75.
[3] Bacon SP, Grantham DW. Modulation masking: Effects of modulation frequency, depth, and phase. J Acoust Soc Am 1989;85(6):2575–80.
[4] Bacon SP, Viemeister NF. Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners. Audiology 1985;24(2):117–34.
[5] Bernstein ER, Brammer AJ, Yu G. Improving speech understanding in communication headsets: Simulation of adaptive subband processing for speech in noise. Int J Ind Ergon 2013;43(6):526–35.
[6] Bosker HR, Cooke M. Talkers produce more pronounced amplitude modulations when speaking in noise. J Acoust Soc Am 2018;143(2):EL121–6.
[7] Casali JG. Powered electronic augmentations in hearing protection technology circa 2010 including active noise reduction, electronically-modulated sound transmission, and tactical communications devices: review of design, testing, and research. Int J Acoust Vib 2010;15(4). https://doi.org/10.20855/ijav10.20855/ijav.2010.15.410.20855/ijav.2010.15.4269.
[8] Clarkson PM, Bahgat SF. Envelope expansion methods for speech enhancement. J Acoust Soc Am 1991;89(3):1378–82.
[9] Corey, R. M., and Singer, A. C. (2020). "Modeling the effects of dynamic range compression on signals in noise," arXiv preprint arXiv:2012.03860.
[10] Depireux DA, Simon JZ, Klein DJ, Shamma SA. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 2001;85(3):1220–34.
[11] Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D. Temporal modulations in speech and music. Neurosci Biobehav Rev 2017;81:181–7.
[12] Dolan TG, O'Loughlin D. Amplified earmuffs: impact on speech intelligibility in industrial noise for listeners with hearing loss. Am J Audiol 2005;14(1):80–5.

[13] Drullman R, Festen JM, Plomp R. Effect of temporal envelope smearing on speech reception. J Acoust Soc Am 1994;95(2):1053–64.
[14] Fu, S. W., Tsao, Y., Lu, X., and Kawai, H. (2017, December). "Raw waveform-based speech enhancement by fully convolutional networks," In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 006-012). IEEE.
[15] Gaudrain E, Başkent D. Discrimination of voice pitch and vocal-tract length in cochlear implant users. Ear Hear 2018;39(2):226.
[16] Green T, Faulkner A, Rosen S. Spectral and temporal cues to pitch in noise-excited vocoder simulations of continuous-interleaved-sampling cochlear implants. J Acoust Soc Am 2002;112(5):2155–64.
[17] Hao, X., Su, X., Wang, Z., and Zhang, H. (2020). "UNetGAN: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition," arXiv preprint arXiv:2010.15521.
[18] Herrmann B, Henry MJ, Haegens S, Obleser J. Temporal expectations and neural amplitude fluctuations in auditory cortex interactively influence perception. Neuroimage 2016;124:487–97.
[19] Houtgast T, Steeneken HJM. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J Acoust Soc Am 1985;77(3):1069–77.
[20] Houtgast TAMMO, Steeneken HJ, Plomp R. Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. Acustica 1980;46(1):60–72.
[21] Houtgast, T., and Steeneken, H. J. (1973). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica, 28(1), 66-73.
[22] House AS, Williams CE, Hecker MHL, Kryter KD. Articulation-testing methods: consonantal differentiation with a closed-response set. J Acoust Soc Am 1965;37(1):158–66.
[23] Huang Y, Trinh MT, Le T. Critical Factors Affecting Intention of Use of Augmented Hearing Protection Technology in Construction. J Constr Eng Manage 2021;147(8):04021088. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002116.
[24] Koutsogiannaki M, Francois H, Choo K, Oh E. Real-time modulation enhancement of temporal envelopes for increasing speech intelligibility. Proc. Interspeech 2017:1973–7.
[25] Launer S, Zakis JA, Moore BC. Hearing aid signal processing. Hearing Aids 2016:93–130.
[26] Langhans T, Strube H. Speech enhancement by nonlinear multiband envelope filtering. Proc ICASSP 1982;7:156–9.
[27] Lezzoum N, Gagnon G, Voix J. Noise reduction of speech signals using time-varying and multi-band adaptive gain control for smart digital hearing protectors. Appl Acoust 2016;109:37–43.
[28] Loizou PC. Speech processing in vocoder-centric cochlear implants. Cochlear and brainstem implants. Adv Otorhinolaryngol 2006;64:109–43.
[29] Moore BC. An introduction to the psychology of hearing. 6th ed. Boston, MA): (Brill; 2013. p. Chap. 5..
[30] Nicolson A, Paliwal KK. Spectral distortion level resulting in a just-noticeable difference between an a priori signal-to-noise ratio estimate and its instantaneous case. J Acoust Soc Am 2020;148(4):1879–89.
[31] Oostermeijer, K., Wang, Q., and Du, J. (2020, December). "Frequency Gating: Improved Convolutional Neural Networks for Speech Enhancement in the Time-Frequency Domain," In 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 465-470). IEEE.
[32] Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Proc ICASSP 2001;2:749–52.
[33] Rix AW. "Comparison between subjective listening quality and P. 862 PESQ score", Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN'03). Prague, Czech Republic; 2003.
[34] Shamma SA. Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method. Network: Comput Neural Syst 1996;7(3):439–76.
[35] Shannon RV, Zeng F-G, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. Science 1995;270(5234):303–4.
[36] Schreiner CE, Urbas JV. Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). Hear Res 1986;21:227–41.
[37] Smalt CJ, Calamia PT, Dumas AP, Perricone JP, Patel T, Bobrow J, et al. The effect of hearing-protection devices on auditory situational awareness and listening effort. Ear Hear 2020;41(1):82–94.
[38] Smith ZM, Delgutte B, Oxenham AJ. Chimaeric sounds reveal dichotomies in auditory perception. Nature 2002;416(6876):87–90.
[39] Wiinberg A, Zaar J, Dau T. Effects of expanding envelope fluctuations on consonant perception in hearing-impaired listeners. Trends Hearing 2018;22:1–12.
[40] Kates JM. Digital hearing aids. Plural publishing. 2008.