Framewise Finite Impulse Response Filtering Based on Time-Frequency Mask for Low-Latency Speech Enhancement

Chiho Haruta*†, Nobutaka Ono* and Yuma Kinoshita*
* Tokyo Metropolitan University, Tokyo, Japan
E-mail: {haruta-chiho@ed., onono@, ykinoshita@}tmu.ac.jp
† RION Co., Ltd.

***Abstract*—We propose a low-latency speech enhancement method using framewise finite impulse response (FIR) filters based on time-frequency (T-F) mask. In many real-time audio applications, such as hearing aids, low-latency processing is highly required. T-F masking-based speech enhancement algorithms improve speech intelligibility for hearing impaired people in noisy environments, but an algorithmic delay due to frame analysis occurs. To shorten this delay, we replace time-frequency masking with framewise filtering in the time domain. The filters are designed on the basis of the signal and noise spectra in each frame, which are the same information used as that for designing a Wiener-filter-based T-F mask. The latency is shortened by designing a causal FIR filter and predicting the signal and noise spectra only from the information in the past frames. Evaluation experiments showed that causal framewise FIR filtering reduced the delay with little degradation of the performance compared with T-F masking.**

I. Introduction

People with hearing loss suffer not only from an elevated threshold for detecting sounds but also from understanding speech, especially in noisy environments [1]. One of the effective approaches to address hearing loss is the use of hearing aids. Hearing aids amplify the acoustic signals and send the amplified signals into the ear canal. The signal is amplified frequency-dependently and the amount of amplification at each frequency region is appropriately determined by the degree of the user's hearing loss in each frequency region [2]–[4]. This process significantly improves speech understanding in quiet environments [5]. However, when background noise is present, both the speech and the noise are amplified, and users with cochlear hearing loss cannot understand speech as effectively as prople with [1]. Therefore, nowadays, most commercially available hearing aids have speech enhancement systems, and there have been many studies on speech enhancement [6]–[9].

A deep neural network (DNN)-based speech enhancement algorithm with time-frequency (T-F) masking in the short-time Fourier transform (STFT) domain exhibited a significant performance gain for speech intelligibility in noisy environments for people with hearing loss [10]–[12]. On the other hand, sound processing in hearing aids requires low latency. If the time difference between the input and output of the hearing aid is large, discomfort due to the deviation between the movement of the speaker's lips and the voice and the difficulty of vocalization of the user may occur [13]–[15]. One study showed that a tolerable delay is approximately 6 ms at 1 kHz [13]. In the speech enhancement process using the T-F mask described above, a delay corresponding to the frame length is unavoidable owing to frame analysis.

Several methods for low-latency speech enhancement have been proposed [16]–[19]. In [16], higher frequency resolution was obtained together with low-delay processing by performing analytical resynthesis by combining a long analysis window and a short composition window. However, in this algorithm, a delay corresponding to the synthesis window is unavoidable. In [17] and [18], the time-frequency analysis and synthesis were replaced by time-variable FIR filtering, but the coefficients were estimated by autoregression and a DNN-based technique was not used. In [19], speech enhancement scheme implemented in the time domain, which separated signals in an intermediate feature space generated by a trained decoder, was proposed. This scheme replaced the STFT with an encoder-decoder architecture. It outperformed ideal time-frequency magnitude masks in a noncausal implementation and also achieved a high performance in a causal implementation.

In this paper, we propose a low-latency DNN-based speech enhancement scheme with framewise FIR filters. In our algorithm, T-F masking is replaced with framewise FIR filtering. By designing causal filters, a causal implementation is realized. The coefficients of the filters are derived from the power spectra of the input and estimated clean signals. The power spectra are derived by T-F masking using the DNN. This implementation is expected to show as high performance as T-F mask-based speech enhancement algorithm. Additionally, our proposed scheme can be used to compensate for the hearing threshold evaluation because amplifying the signal at a certain frequency is equivalent to setting a value larger than 1.0 for certain T-F components. Therefore, it is suitable for real-time implementation in hearing aids.

978-988-14768-9-0/21/$31.00 ©2021 APSIPA

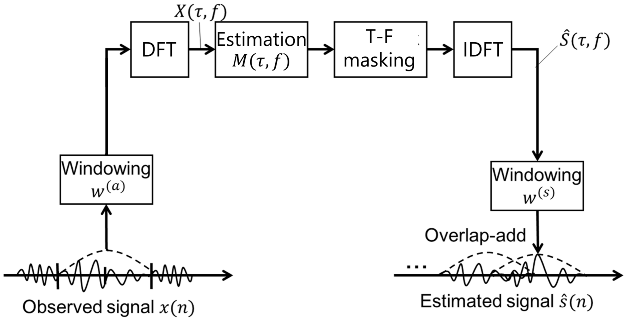


Fig. 1: Block diagram of conventional T-F masking algorithm.

II. Conventional Speech Enhancement Algorithms

A. T-F Mask-Based Speech Enhancement in the Frequency Domain

Let $x(n)$ be a mixture of clean speech $s(n)$ and a noise signal $v(n)$, where n is a discrete time index. Thus, $x(n) = s(n) + v(n)$. Let

$$X(\tau,f) = \sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1} w^{(a)}(m)x(\tau N_s + m)e^{\frac{-j2\pi fm}{N}} \tag{1}$$

be the STFT representation of $x(n)$, where τ and f denote the indices of the time frame and frequency, respectively, $w^{(a)}(m)$ denotes an analysis window function, N is the frame length, which is assumed to be an even number, and N_s is the frame shift.

In the T-F masking, the target signal is estimated by applying T-F mask $M(\tau,f)$ to $X(\tau,f)$ in the STFT domain such as

$$\hat{S}(\tau,f) = M(\tau,f)X(\tau,f). \tag{2}$$

Then, the estimated clean speech $\hat{s}^{(\mathrm{TF})}(n)$ in the time domain is obtained by the inverse discrete Fourier transform (IDFT) and the overlap-add as

$$\hat{s}^{(\mathrm{TF})}(n) = \sum_{\tau=-\infty}^{\infty} w^{(s)}(n-\tau N_s) \left[\frac{1}{N} \sum_{f=-\frac{N}{2}+1}^{\frac{N}{2}} \hat{S}(\tau,f)e^{\frac{j2\pi f(n-\tau N_s)}{N}} \right], \tag{3}$$

where $w^{(s)}(m)$ denotes a synthesis window function. For perfect reconstruction,

$$\sum_{\tau=-\infty}^{\infty} w^{(s)}(n-\tau N_s)w^{(a)}(n-\tau N_s) = 1, \tag{4}$$

must be satisfied.

Fig. 1 shows a block diagram of the T-F masking algorithm. There are various methods of designing the T-F mask. We can directly estimate the T-F mask or estimate the power spectrogram of the clean speech $|S(\tau,f)|^2$ and design the T-F mask as

$$M(\tau,f) = \frac{|S(\tau,f)|^2}{|X(\tau,f)|^2}. \tag{5}$$

For both methods, recent studies have found that a DNN works very well for designing the T-F mask [10].

Although the T-F mask is effective for speech enhancement, one problem is the latency. Fig. 2(a) shows the relationship between the time indices of the estimated sample and the observed samples used to estimate it. Because all the samples of one frame are necessary for processing in the STFT domain, one has to wait until the last sample of the observed signal in a frame is acquired before estimating the first sample in the frame. Thus, the delay of one frame is unavoidable in the T-F masking.

B. Low-Latency Speech Enhancement Algorithms

As mentioned in Sec. I, various methods to shorten the delay have been proposed such as using short synthesis window [16], adaptive T-F analysis [17], [18], and encoder-decoder architecture [19]. In this paper, we realize low-latency speech enhancement by framewise FIR filtering. The filters are designed by using DNN-based prediction of spectrogram and T-F mask. It can be a causal implementation by changing the number of noncausal components that remained when calculating the coefficients of the filter. This scheme is based on T-F masking in the STFT domain, thus, it is also suitable for hearing aids that amplify the signal frequency-dependently.

III. Time-Domain Implementation of T-F Masking

A. Framewise FIR Filtering

To shorten the delay due to the frame analysis mentioned in Sec. II-A, we propose the replacement of T-F masking with time-varying FIR filtering in this work. We here consider a framewise FIR filter $h_\tau(k)$ $(-N_2 \le k \le N_1)$ at the τth frame, where N_1 and N_2 denote positive integers or zero. When we replace T-F masking and the IDFT with the FIR filtering, the estimated clean speech $\hat{s}^{(\mathrm{FIR})}(n)$ is represented by

$$\hat{s}^{(\mathrm{FIR})}(n) = \sum_{\tau=-\infty}^{\infty} w^{(s)}(n-\tau N_s) \cdot \left[\sum_{m=-N_2}^{N_1} w^{(a)}(n-\tau N_s - m)x(n-\tau N_s - m)h_\tau(m) \right]. \tag{6}$$

For a perfect reconstruction, $w^{(s)}$ must satisfy

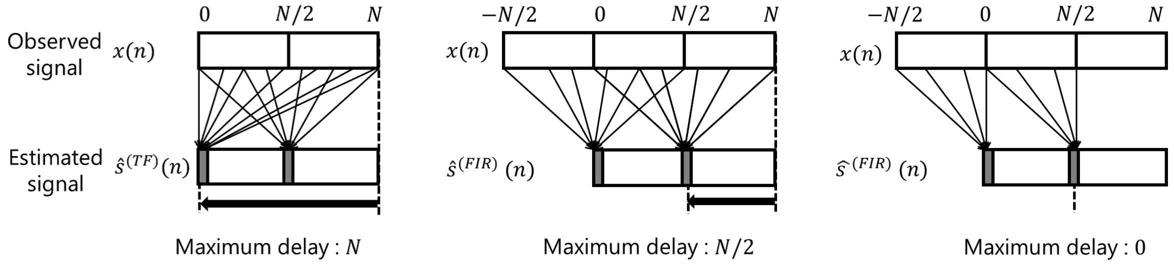
$$\sum_{\tau=-\infty}^{\infty} w^{(s)}(n-\tau N_s)w^{(a)}(n-\tau N_s) = 1. \tag{7}$$

Fig. 3 shows a block diagram of the proposed algorithm.

If the optimum FIR filter $\boldsymbol{h}_\tau$ in terms of the mean squared error (MSE) is calculated per frame, speech enhancement with T-F masking is replaced with FIR filtering.

B. Wiener Filter

Here, we would like to review the derivation of the Wiener filter [20] that minimizes MSE between the estimated and clean speech signals. In this section, we focus on one frame and derive an optimum FIR filter so that the frame index τ is omitted.



(a) T-F masking (conv.) (b) Noncausal filtering (prop.) (c) Causal filtering (prop.)

Fig. 2: Required observation for estimating each sample.

For integers n_1 and n_2, where $n_1 < n_2$, when the observed signal $\boldsymbol{x}(n) = [x(n-N_1), x(n-N_1+1), \cdots, x(n_2)]^T$ is given, by defining FIR filter $\boldsymbol{h} = [h(-N_2), h(-N_2+1), \cdots, h(N_1)]^T$, MSE is calculated using the following cost function of $\boldsymbol{h}$:

$$J(\boldsymbol{h}) = E\left[|\boldsymbol{h}^T\boldsymbol{x}(n) - s(n)|^2\right], \qquad (8)$$

where $E[\cdot]$ is the expectation operator. When $\boldsymbol{h}$ minimizes $J(\boldsymbol{h})$, it must satisfy

$$\frac{\partial J(\boldsymbol{h})}{\partial \boldsymbol{h}} = 2E[\boldsymbol{x}(n)\boldsymbol{x}(n)^T]\boldsymbol{h} - 2E[\boldsymbol{x}(n)s(n)] = \boldsymbol{0}. \qquad (9)$$

Let $\phi_{xx}(m)$ and $\phi_{ss}(m)$ be the autocorrelations of $x(n)$ and $s(n)$, respectively, corresponding to time lag m. Then, let us define $\boldsymbol{\Phi}_{xx}$ as an $N \times N$ matrix whose (k,l)th element is represented as

$$\boldsymbol{\Phi}_{xx}[k,l] = \phi_{xx}(k-l) \qquad (10)$$

and $\boldsymbol{\phi}_{ss}$ as an N-dimensional vector whose kth element is represented as

$$\boldsymbol{\phi}_{ss}[k] = \phi_{ss}(k - N_1 + 1). \qquad (11)$$

We define $\boldsymbol{\Phi}_{xx}^{(N_1+N_2+1)}$ as the top-left $(N_1+N_2+1)\times(N_1+N_2+1)$ submatrix of $\boldsymbol{\Phi}_{xx}$ and $\boldsymbol{\phi}_{ss}^{(N_1+N_2+1)}$ as the subvector of $\boldsymbol{\phi}_{ss}$ with entries indexed by 1 to N_1+N_2+1. Assuming that $x(n)$ is a stationary signal and that $x(n)$ and $s(n)$ are uncorrelated, we obtain

$$\tilde{\boldsymbol{h}} = (\boldsymbol{\Phi}_{xx}^{(N_1+N_2+1)})^{-1}\boldsymbol{\phi}_{ss}^{(N_1+N_2+1)}. \qquad (12)$$

Eqs. (10)-(12) indicate that the optimum FIR filter in terms of MSE can be calculated from the autocorrelations of $x(n)$ and $s(n)$.

C. Approximation of Autocorrelation Based on Frame Analysis

As shown in Sec. III-B, the Wiener filter can be designed from the autocorrelations of $x(n)$ and $s(n)$. To design the Wiener filter in each frame, we here consider how to estimate the autocorrelations of $x(n)$ and $s(n)$ in each frame. According to the Wiener–Khinchin theorem [20], an autocorrelation function is equal to the inverse Fourier transform of power spectrum density. Therefore, assume that we have the estimates of the power spectra of $s(n)$ and $x(n)$ in each frame. We will describe how to estimate them in relation to T-F masking in the next section. In such a discrete and finite-length case, the N-point autocorrelation function is obtained by the IDFT of the N-point power spectrum. In this case, the autocorrelation function becomes a periodic function with period N. As it is also an even function, we have

$$\phi_{xx}(m) = \phi_{xx}(N-m) \quad (m > \frac{N}{2}). \qquad (13)$$

Then, we consider approximating the autocorrelation functions of $s(n)$ and $x(n)$ by using Eq. (13). This involves replacing $\phi_{ss}(m)$ and $\phi_{xx}(m)$ with $\phi_{ss}(N-m)$ and $\phi_{xx}(N-m)$, respectively, when $m > N/2$. Note that they are obtained in an N-point frame. Let $\boldsymbol{\Psi}_{xx}$ be the matrix obtained by such a replacement of $\boldsymbol{\Phi}_{xx}$. Defining $\boldsymbol{\Psi}_{xx}^{(N_1+N_2+1)}$ as the top-left $(N_1+N_2+1)\times(N_1+N_2+1)$ submatrix of $\boldsymbol{\Psi}_{xx}$, in the same way as defining $\boldsymbol{\Phi}_{xx}^{(N_1+N_2+1)}$, Eq. (12) is rewritten as

$$\boldsymbol{h} = (\boldsymbol{\Psi}_{xx}^{(N_1+N_2+1)})^{-1}(\boldsymbol{\phi}_{ss}^{(N_1+N_2+1)}). \qquad (14)$$

D. Wiener Filtering and T-F Masking

Let us consider the special case that $N_1 = N/2 - 1$ and $N_2 = N/2$ and investigate the relationship between Wiener filtering and T-F masking. Then, $\boldsymbol{\Psi}_{xx}$, which is a circulant matrix, is diagonalized by using discrete Fourier transform (DFT) matrix of the same size. Defining $\boldsymbol{F}$ as an N-dimensional DFT matrix whose (k,l)th element is $e^{-i2\pi(k-1)(l-1)/N}$, we obtain

$$\boldsymbol{\Psi}_{xx}{}^{-1} = \boldsymbol{F}^{-1}\ \mathrm{diag}(\boldsymbol{p}_{xx}(n))\boldsymbol{F}, \qquad (15)$$

$$\boldsymbol{\phi}_{ss} = \boldsymbol{F}^{-1}\boldsymbol{p}_{ss}(n), \qquad (16)$$

where $\boldsymbol{p}_{xx}(n)$ and $\boldsymbol{p}_{ss}(n)$ denote the power spectra of $\boldsymbol{x}(n)$ and $[s(n), s(n+1), \cdots, s(n+N-1)]$, respectively, and diag$(\cdot)$ denotes a diagonal matrix with the vector in brackets on its main diagonal. Then, Eq. (14) is rewritten as

$$\boldsymbol{h} = \boldsymbol{F}^{-1}\boldsymbol{p}_{ss}(n) \oslash \boldsymbol{p}_{xx}(n), \qquad (17)$$

where $\oslash$ denotes elementwise division. Eq. (17) indicates that the Wiener filter is consistent with the IDFT of the ideal T-F

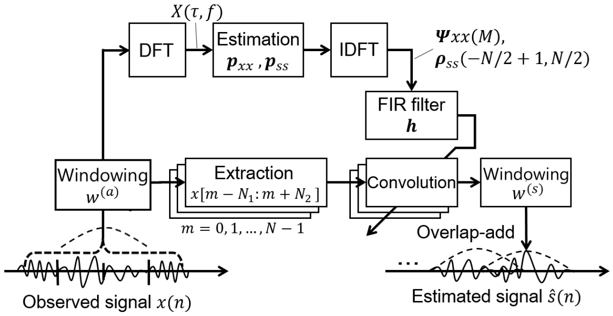


Fig. 3: Block diagram of the proposed algorithm.

mask when $N_1 = \frac{N}{2} - 1$ and $N_2 = \frac{N}{2}$. Therefore, to obtain the whole estimated signal in Eq. (6), framewise FIR filter $\boldsymbol{h}$ is calculated using framewise autocorrelation $\boldsymbol{\Psi}_{xx,\tau}$ and $\boldsymbol{\phi}_{ss,\tau}$ based on framewise power spectra $\boldsymbol{p}_{xx,\tau}$ and $\boldsymbol{p}_{ss,\tau}$.

Since the elements of $\boldsymbol{h}_\tau$ with positive and negative time indices correspond to causal and noncausal components, respectively, the algorithmic delay is equal to N_2 samples. The filter length is $N_1 + N_2 + 1$ and the frame length for estimating power spectra is equal to circulation period N. In the case of $N_1 = \frac{N}{2} - 1$ and $N_2 = \frac{N}{2}$, FIR filtering is equivalent to conventional T-F masking and the algorithmic delay corresponds to $\frac{N}{2}$ samples.

E. Design of Framewise FIR Filter

The causality of the FIR filter depends on N_2. To make it easier to compare the proposed algorithm with the T-F masking algorithm, we set $N_1 = \frac{N}{2} - 1$ below. As mentioned in Sec. III-C, noncausal FIR filter $\boldsymbol{h}_\tau^{\text{noncausal}}$ that is equivalent to T-F masking is derived as

$$\begin{aligned} \boldsymbol{h}_\tau^{\text{noncausal}} &= \big(\boldsymbol{\Psi}_{xx}^{(N)}\big)^{-1} \boldsymbol{\phi}_{ss}^{(N)} \\ &= \boldsymbol{\Psi}_{xx}{}^{-1} \boldsymbol{\phi}_{ss}, \end{aligned} \tag{18}$$

where $N_1 = \frac{N}{2} - 1$ and $N_2 = \frac{N}{2}$ and the delay corresponds to $\frac{N}{2}$ samples. Fig. 2(b) shows the relationship between the time indices of the observed and estimated signals in this case. Causal filter $\boldsymbol{h}_\tau^{\text{causal}}$ is calculated as

$$\boldsymbol{h}_\tau^{\text{causal}} = \big(\boldsymbol{\Psi}_{xx}^{(\frac{N}{2})}\big)^{-1} \boldsymbol{\phi}_{ss}^{(\frac{N}{2})}, \tag{19}$$

where $N_1 = \frac{N}{2} - 1$ and $N_2 = 0$ and the delay is zero. Additionally, if a slight delay of N_d samples is allowed, we can set $N_1 = \frac{N}{2} - 1$ and $N_2 = N_d$, then we obtain a filter with N_d noncausal components. We can set an appropriate N_d in accordance with the tolerable delay or its speech enhancement performance. Fig. 2(c) shows the relationship between the time indices of the observed and estimated signals in this case.

IV. Spectrum Prediction for Realizing Low-Latency Processing

For the continuous processing of FIR filtering, the optimum FIR filter at a certain frame should be determined at the start of the frame. This means that in Eq. (6), at the start of the τth frame, in other words, at time $n = \tau N_s$, power spectra $\boldsymbol{p}_{xx,\tau}$ and $\boldsymbol{p}_{ss,\tau}$ must be predicted.

We apply neural networks to predict them. We apply two networks to estimate $\boldsymbol{p}_{xx,\tau+N/N_s}$ from past power spectra $\boldsymbol{p}_{xx,\tau}$, $\boldsymbol{p}_{xx,\tau-1}$, $\boldsymbol{p}_{xx,\tau-2}, \cdots$ and estimate $p_{ss,\tau+N/N_s}$ from the same information as the former one. The former estimation is equivalent to predicting a future power spectrum from past power spectra. For the latter estimation, instead of directly estimating $\boldsymbol{p}_{ss,\tau+N/N_s}$, we estimate $\boldsymbol{m}_{\tau+N/N_s} = \boldsymbol{p}_{ss,\tau+N/N_s} \oslash \boldsymbol{p}_{xx,\tau+N/N_s}$ and have an estimate of $\boldsymbol{p}_{ss,\tau+N/N_s}$ by $\boldsymbol{m}_{\tau+N/N_s} \odot \boldsymbol{p}_{xx,\tau+N/N_s}$, where $\odot$ denots Hadamard product. The good performance of the power spectrogram prediction has already been reported in [21].

V. Experimental Evaluations

A. Dataset

We used clean speech from the Japanese Newspaper Article Sentences corpus [22] and noise from the TUT dataset [23]. To generate training data, 12 h of speech (50 male and 50 female speakers) of the corpus was combined with noise recorded in 15 types of environment at signal-to-noise ratios (SNRs) of 0, 5, and 10 dB. Evaluation data were obtained from 20 min of speech (five male and five female speakers different from the speakers used for the training data) and combined with five types of noise.

B. Setup

The sampling frequency was 16 kHz, the frame length was 1024 samples, the frame shift length was 512 samples, and a Hamming window was used for frame analysis as shown in Table I. We used two neural networks to estimate $\boldsymbol{p}_{xx,\tau+N/N_s}$ and $\boldsymbol{m}_{\tau+N/N_s}$. For the neural networks, we used the convolutional layers listed in Table II followed by fully-connected layers. The networks were designed to be causal and only require the information in the past frames for prediction. For training, the learning rate was set to 0.01 and Adam [24] was used as the optimization algorithm with decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For training to estimate $\boldsymbol{p}_{xx,\tau+N/N_s}$, ReLU [25] was used as the activation function for all layers. For training to estimate $\boldsymbol{m}_{\tau+N/N_s}$, ReLU was used as the activation function for all layers except for the output layer, which used a sigmoid function. Each model was trained for 500 epochs. To prevent the divergence of $\boldsymbol{p}_{xx,\tau+N/N_s}$ or $\boldsymbol{h}$ in Eq. (17), each element of predicted $\boldsymbol{p}_{xx,\tau+N/N_s}$ was bounded within the range from 0.01 to 2.0 times of each element of $\boldsymbol{p}_{xx,\tau}$. This process did not require the information of future samples.

We compared the performance of conventional T-F masking and the proposed FIR filtering algorithm. The proposed algorithm had three conditions: causal ($N_d = 0$) and noncausal ($N_d = 16$ and $N_d = 512$). Additionally, as a simple truncation condition in which the noncausal component of $\boldsymbol{h}_\tau^{\text{noncausal}}$ calculated in Eq. (18) is truncated for comparison. To evaluate the performance, the scale-invariant source-to-distortion ratio (SI-SDR) [26] is employed.

TABLE I: Experimental conditions.

Sampling frequency	16 kHz
Frame length	1024
Filter length	1024
Frame shift	512
Number of remaining noncausal components, N_d	0, 16, 512
Analysis window function	Hamming
Synthesis window function (only for prop.)	Hamming

TABLE II: Convolutional layer configuration. T width and F width are the sizes of the filter in time (frames) and frequency (bins), respectively. T and F dilations are the dilation factors in time and frequency, respectively.

Filters	T width	F width	T dilation	F dilation
32	1	7	1	1
32	5	5	1	1
32	5	5	2	1
32	5	5	4	1
32	5	5	8	1
32	5	5	16	1
8	1	1	1	1

TABLE III: Speech enhancement performance.

Method	Noncausal components	Design	SI-SDRi [dB]
T-F masking (conv.)	-	MMSE	8.17
Noncausal filtering (prop.)	512 (32 ms)	MMSE	8.18
Noncausal filtering (prop.)	16 (1 ms)	MMSE	5.78
Causal filtering (prop.)	0 (0 ms)	MMSE	5.74
Causal filtering	-	Truncation	-7.52

TABLE IV: Speech enhancement performance of each prediction by causal filtering.

$\boldsymbol{p}_{xx,\tau+N/N_s}$	$\boldsymbol{m}_{\tau+N/N_s}$	SI-SDRi [dB]
Ground truth	Predicted	6.12
Predicted	Ground truth	9.53
Predicted	Predicted	5.74

C. Results

Table III presents the SI-SDR improvement (SI-SDRi) for each condition. Compared with the conventional T-F masking algorithm with a delay of 64 ms, the degradation of the performance of the proposed algorithm with the causal filter was 2.43 dB. This degradation is considered to be caused by the introduction of the circular model in Sec. III-C.

To further investigate the performance, we compared the cross-correlations of the estimated signals and the target signal. Even though the causal implementation is realized in the proposed method, a group delay of the filter may cause a delay. Thus, we investigate the cross-correlations. Figs. 4 and Fig. 5, respectively, show the normalized cross-correlations between the target signal and the signals estimated by T-F masking and causal filtering. In both T-F masking and causal filtering, the peak appears at a lag of zero. This means that no delay occurs even in the case of causal filtering.

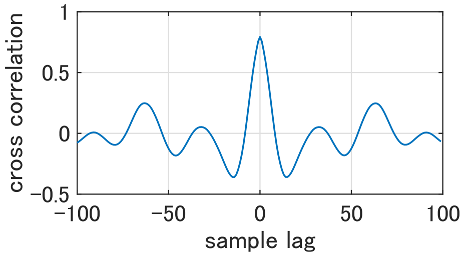


Fig. 4: Cross-correlation of processed signal by T-F masking and target signal.

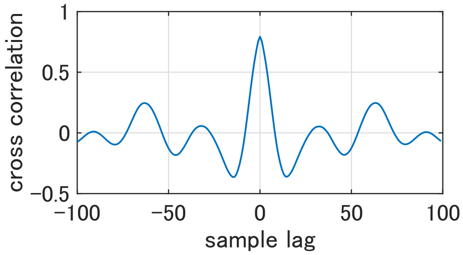


Fig. 5: Cross-correlation of processed signal by causal filtering and target signal.

To obtain better performance in the future, we investigate the speech enhancement performance of the proposed methods with the ground truth of $\boldsymbol{p}_{xx,\tau+N/N_s}$ and $\boldsymbol{m}_{\tau+N/N_s}$. Table IV shows the performance obtained when one of them is replaced with the ground truth data. It indicates that the performance could be improved by improving the accuracy of predicting both $\boldsymbol{p}_{xx,\tau+N/N_s}$ and $\boldsymbol{m}_{\tau+N/N_s}$. For example, both of them are trained independently in this study, but they may be combined and trained as multitask learning.

VI. Conclusions

We propose a low-latency speech enhancement method using framewise FIR filters based on T-F mask. By designing the filters to be causal, the implementation can be causal. The experiments showed that the proposed algorithms reduced the algorithmic delay from 64 to 0 ms with degradation of only 2.43 dB in SI-SDR, compared with the conventional T-F masking algorithm. Future tasks are to further investigate the tradeoff between the delay and the performance, combinate the speech enhancement and the compensation of the hearing threshold elevation together, reduce the computational load by element selection [27], and achieve better performance. performance.

Acknowledgment

This work was supported by JST CREST Grant Number JPMJCR19A3, Japan.

References

[1] M. C. J. Brian, *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*, 2$^{\text{nd}}$ ed., New Jersey: John Wiley & Sons Ltd., Hoboken, 2007, pp. 1–10.

[2] B. Caswell-Midwinter and M. W. William, "Discrimination of gain increments in speech," *Trends in hearing*, vol. 23, pp. 1–9, 2019.

[3] G. Keidser, H. Dillon, M. Flax, T. Ching, and S. Brewer, "The NAL-NL2 prescription procedure," *Audiology Research*, vol. 1, no. 1, 2011.

[4] R. Seewald, S. Moodie, S. Scollie, and M. Bagatto, "The DSL method for pediatric hearing instrument fitting: historical perspective and current issues," *Trends Amplif.*, vol. 9, no. 4, pp.145–157, 2005.

[5] Z. Zhang, S. W. Donald, and Y. Shen, "Impact of amplification on speech enhancement algorithms using an objective evaluation metric," *International Congress on Acoustics*, 2019.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[7] H. Luts, K. Eneman, J. Wouuters, M. Schulte, M. Vormann, M. Buechler, N. Dillier, R. Houben, W. A. Dreschler, M. Froehlich, H. Puder, G. Grimm, V. Hohmann, A. Leijon, A. Lombard, D. Mauler, and A. Spriet, "Multicenter evaluation of signal enhancement algorithms for hearing aids," in *Journal of the Acoustical Society of America*, vol. 127, no. 3, pp.1491–1505, 2010.

[8] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," in *Journal of the Acoustical Society of America*, vol. 122, no. 5, pp. 1777–1786, 2007.

[9] F. Dubbelboer and T. Houtgast, "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," in *Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3937–3946, 2008.

[10] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview," in *IEEE/ACM transactions on Audio, Speech, and Language Processing*, vol. 26, pp. 1702–1726, 2018.

[11] D. L. Wang, "Time-Frequency masking for speech separation and its potential for hearing aid design," in *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[12] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[13] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," in *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[14] M. A. Stone, B. C. J. Moore, K. Meisenbacher, and R. P. Derleth, "Tolerable hearing aid delays. V. Estimation of limits for open canal fittings," in *Ear and Hearing*, vol. 29, no. 4, pp. 601–617, 2008.

[15] R. Heurig and J. Chalupper, "Acceptable processing delay in digital hearing aids," *Hearing Review*, vol. 17, no. 1, pp. 28–31, 2010.

[16] D. Mauler and R. Martin, "A low delay, variable resolution, perfect reconstruction spectral analysis-synthesis system for speech enhancement," in *European Signal Processing Conference*, pp. 222–226, 2007.

[17] H. W. Lollmann and P. Varym "Low delay filter-banks for speech and audio processing", in *Speech and audio processing in adverse environments*, Berlin: Springer, 2008, pp. 13–61.

[18] K. T. Andersen and M. Moonen, "Adaptive time-frequency analysis for noise reduction in an audio filter bank with low delay," in *IEEE Transactions on audio, speech and language processing*, vol. 24, no. 4, pp. 784-795, 2016.

[19] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-frequency Magnitude Masking for Speech Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256-1266, 2019.

[20] S. Haykin, *Adaptive filter theory*, 3rd ed., New Jersey: Prentice-Hall, 1996, pp. 10–14.

[21] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in model for low-latency speech enhancement," *International Workshop on Acoustic Signal Enhancement*, pp. 366–370, 2018.

[22] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," vol. 20, no. 3, pp. 199–206, 1999.

[23] A. Mesaros, T. Heittola and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *Proc. European Signal Processing Conference*, pp. 1128–1132, 2016.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference for Learning Representations*, 2015.

[25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of International Conference on Artificial Intelligence and Statistics*, vol. 15, pp. 315–323, 2011.

[26] J. L. Roux, S. Wisdom, H. Erdogan and J. R. Hershey, "SDR half-baked or well done?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 626–630, 2019.

[27] C. Haruta and N. Ono, "A low-computational DNN-based speech enhancement for hearing aids based on element selection," in *European Signal Processing Conference*, pp. 1025–1029, 2021.