Multimedia Tools and Applications (2021) 80:10777–10796
https://doi.org/10.1007/s11042-020-10230-y

A frequency-domain nonlinear echo processing algorithm for high quality hands-free voice communication devices

Qingyun Wang [1] · Xin Chen [1] · Ruiyu Liang [1] · Haicheng Liu [2]

Received: 21 February 2020 / Revised: 14 August 2020 / Accepted: 9 December 2020/
Published online: 3 January 2021

Abstract
A frequency-domain nonlinear echo processing algorithm is proposed to improve the audio quality during double-talk periods for hands-free voice communication devices. To achieve acoustic echo cancellation (AEC), a real-time AEC algorithm based on variable step-size partitioned block frequency-domain adaptive filtering (VSS-PBFDAF) and frequency-domain nonlinear echo processing (FNLP) algorithm was employed in the DSP chip of the prototype device. To avoid divergence during double-talk periods, normalized variable step-sizes for each frequency were introduced to adjust the convergence speed. Then, the nonlinear suppression function of FNLP was applied to inhibit the residual nonlinear acoustic echo and ensure the good quality of the near-end voice. The results of the experiment with the prototype device show that the proposed algorithm achieved deeper and more stable convergence during double-talk periods compared to the NLMS, FNLMS and traditional PBFDAF algorithms. Less nonlinear acoustic echo in the output was also obtained due to the use of FNLP. A speech quality assessment based on ITU-T P.563 showed that the Sout of the proposed algorithm achieved higher scores than that of the WebRTC algorithm. In addition, the speech output of the proposed algorithm during the double-talk periods was clear and coherent.

Keywords Acoustic echo cancellation (AEC) · Double-talk (DT) · Frequency nonlinear processing (FNLP) · Variable step-size partitioned block frequency-domain adaptive filtering (VSS-PBFDAF) · Hands-free communication device

✉ Ruiyu Liang
liangry@njit.edu.cn

[1] School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing, People's Republic of China

[2] Information Science and Engineering, Southeast University, Nanjing, People's Republic of China

1 Introduction

Acoustic echo cancellation (AEC) is the most important part influencing the performance of hands-free voice communication systems, such as mobile phones, voice over IP (VoIP), digital hearing aids, intelligent loudspeakers, intelligent TVs, etc. [27]. Standard AEC schemes rely on the assumption that the echo path is modeled as a linear filter [10, 33]. However, because of the nonlinearity of the speaker, the microphone and the reflection coefficients of obstacles, there are often considerable residual nonlinear echoes that remain after linear AEC processing [25]. The situation gets worse when the room impulse response (RIR) is longer than the length of the adaptive estimated filter [23]. Some nonlinear models are introduced to reduce the nonlinear acoustic echoes, such as Volterra filters [2, 20], spectral feature-based artificial neural networks [29, 35], functional link adaptive filters [7, 8, 37] and kernel methods [5, 19].

For voice communication systems, acoustic echo cancellation is always studied along with the difficulties that need to be approached without harming the near-end voice that must be transmitted during double-talk (i.e., simultaneous far-end and near-end speech) [16]. The quality of the near-end voice is crucial for consumer experiences [12]. The high convergence rate of the adaptive filter is usually accompanied by a high divergence rate in the presence of double-talk [13]. Some approaches try to detect the existence of double-talk. When the system is considered to contain double-talk, the adaptive filter is frozen, and the system will not be updated to avoid divergence [6]. However, because of the environmental noise, interference and disturbing speech at the subscriber-side, double-talk is a changing and progressive process [1]. During the time that it takes to detect double-talk, the system may have changed. To solve the problem, some approaches estimate the probability of the presence of double-talk and then dynamically adjust the update speed of the adaptive filter by using soft decisions [18, 31].

How to inhibit nonlinear acoustic echoes in hands-free voice communication systems during double-talk periods has intrigued researchers over the past decades. Along with the divergence of the adaptive linear AEC, the nonlinear acoustic echoes occur during double-talk. As the near-end speech is mixed with the residual nonlinear echoes, the inhibition of the nonlinear echoes reduces the quality of the near-end speech, and it sometimes leads to truncated sentences. The method involving the use of nonlinear functions in the frequency domain has attracted considerable interest because of its good speech quality and low computational complexity [28]. Working in the frequency-domain, the algorithm calculates the inhibition coefficients in different frequencies to implement an accurate adjustment [4].

In this paper, a variable step-size partitioned block frequency-domain adaptive filtering (VSS-PBFDAF) scheme is proposed to eliminate the acoustic echoes for hands-free voice communication devices. To decrease the divergence during double-talk, a set of varying step-sizes for different frequencies is introduced to adjust the convergence speed according to the extent of the double-talk. Then, frequency-domain nonlinear echo processing (FNLP) is applied to inhibit the residual nonlinear echoes and ensure the quality of the near-end voice. According to the relevance and correlations between the received-side input (Rin, coming from far-end) and the subscriber-side input (Sin, consisting of the near-end speech and echoes) in the frequency-domain, the suppression function inhibits the residual nonlinear echoes by using a set of coefficients for each frequency. The synthesized output signals have good speech quality and satisfactory nonlinear echo inhibition. We applied the proposed algorithm in a prototype device based on a DSP platform. The simulation and experimental results have verified the effectiveness of the proposed algorithm. Compared with the NLMS, FNLMS and traditional PBFDAF algorithms [9], deeper and more stable convergence was obtained by the proposed algorithm. The speech quality assessment for the Sout based on ITU-T

P.563 showed that good scores were obtained by the proposed algorithm, and coherent speech was output during the double-talk periods.

2 Related works

In this section, we highlight the related works in adaptive acoustic echo cancellation algorithms and their developmental achievments.

2.1 Adaptive linear AEC

Mainly in the last decades, several methods have been developed to cope with the problem of acoustic feedback cancellation. LMS based algorithm [34] presented firstly by B. Widrow and M. E. Hoff has a complexity that is linear in the filter length, but it suffers from a rather slow convergence for signals with a colored spectrum such as speech [24]. Block-LMS decreases the complexity by segment the stream and update the filter every block iteratively [22]. The frequency-domain adaptive filter (FDAF) was introduced by J. Shynk, which is a direct translation of block-LMS to the frequency domain [30]. With the increase of the filter length of FDAF, the computational complexity and the time delay of the algorithm grow rapidly. Koen Eneman and Marc Moonen proposed iterated partitioned block frequency-domain adaptive filter (IPBFDAF) to decrease the latency while keeping the long length of the adaptive filter. The partitioned block frequency–domain row action projection (PBFDRAP), which is the fast version of PBFDAF, leads to reduced algorithmic complexity and is widely used in commercial echo cancellers nowadays [9].

2.2 Nonlinear AEC

One of the limitations of linear adaptive echo cancellers is nonlinearities which are generated mainly in the loudspeaker and the nonlinear echo path. The traditional linear AEC algorithm is difficult to eliminate this nonlinearity, so some scholars began to study the AEC algorithm based on nonlinear model. Pao [14, 26] proposed a functional links method, which is one of the powerful methods of modeling nonlinearity. Guerin and Faucon [14] developed a nonlinear module based on polynomial Volterra filters. The algorithm presents a very promising way of modeling a large range of nonlinearity. For real-time applications, the nonlinear acoustic echo suppression algorithm based on spectral correction has been widely concerned by researchers with the advantages of low computational complexity and fast convergence characteristics [11]. Working in the frequency-domain, the algorithm calculates the inhibition coefficients in different frequencies to implement an accurate adjustment of nonlinear echo cancellation.

2.3 AEC based on neural network (NN)

In recent years, with the development of deep learning technology, AEC algorithm based on NN has gradually become a research hotspot. Mehdi Bekrani [3] proposed a linear single-layer feedforward neural network to efficiently decorrelate the tap-input vectors, which can achieve

a high rate of misalignment convergence without significantly degrading the quality. Hao Zhang [36] trained a recurrent neural network with bidirectional long short-term memory (BLSTM) to separate and suppress the far-end signal, hence removing the echo. Qinhui Lei [21] proposed a deep NN-based regression approach that directly estimates the amplitude spectrum of the near-end target signal from features extracted from the mixtures of near-end and far-end signals. Halimeh and Huemmer [15] used the principle of transfer learning to train a neural network that approximates the nonlinear function. Based on a large number of training, the above algorithm can obtain better performance than the traditional algorithm. However, such algorithms still need further research. First, these algorithms often need a lot of training data. Therefore, the validity of the data is very important. Second, the high computational complexity of the neural network makes it difficult to be implemented on low-power acoustic devices.

3 Double-talk robust acoustic echo cancellation scheme

3.1 A. Modeling

Figure 1 shows the diagram of the acoustic echo cancellation scheme for hands-free voice communication devices based on variable step-size partitioned block frequency-domain adaptive filtering (VSS-PBFDAF) and frequency-domain nonlinear processing (FNLP) considering double-talk (DT).

1) $\mathbf{x}^{(n)}$ and $\mathbf{X}^{(n)}$: $\mathbf{x}^{(n)}$ is the referenced signal from the received-side (which is also represented as Rin). It is segmented into L-length blocks. Parameter n is the block time index. $\mathbf{X}^{(n)}$ is the FFT spectrum of each block $\mathbf{x}^{(n)}$.
2) $\mathbf{y}^{(n)}$ and $\mathbf{Y}^{(n)}$: $\mathbf{y}^{(n)}$ is the estimated linear echo signal of a block and $\mathbf{Y}^{(n)}$ is the FFT spectrum of $\mathbf{y}^{(n)}$.
3) $\mathbf{d}^{(n)}$ and $\mathbf{D}^{(n)}$: $\mathbf{d}^{(n)}$is the microphone signal from the subscriber-side (which is also represented as Sin) and $\mathbf{D}^{(n)}$ is the FFT spectrum of $\mathbf{d}^{(n)}$.
4) $\boldsymbol{\mu}^{(n)}$ is the variable step-size vector that is calculated for partitioned block frequency-domain adaptive filtering (PBFDAF) by using the double-talk soft decision.
5) $\mathbf{e}^{(n)}$ and $\mathbf{E}^{(n)}$:$\mathbf{e}^{(n)}$ is the residual signal that is obtained by subtracting $\mathbf{d}^{(n)}$ from $\mathbf{y}^{(n)}$, which is one of the inputs of frequency-domain nonlinear processing (FNLP).$\mathbf{E}^{(n)}$ is the FFT spectrum of $\mathbf{e}^{(n)}$.
6) $\mathbf{s}^{(n)}$: It is the signal after FNLP, which is the output of the system.

3.2 VSS-PBFDAF algorithm considering double-talk

For the partitioned block frequency-domain adaptive filter, suppose that the estimated time-domain filter coefficient of a block is

$$\widehat{\mathbf{w}}_p^{(n)} = \begin{bmatrix} \widehat{w}^{(n)}[pP] \\ \vdots \\ \widehat{w}^{(n)}[(p+1)P-1] \end{bmatrix}_{P\times 1}, p = 0,\ 1,\ \cdots,\ \frac{N}{P}-1 \tag{1}$$

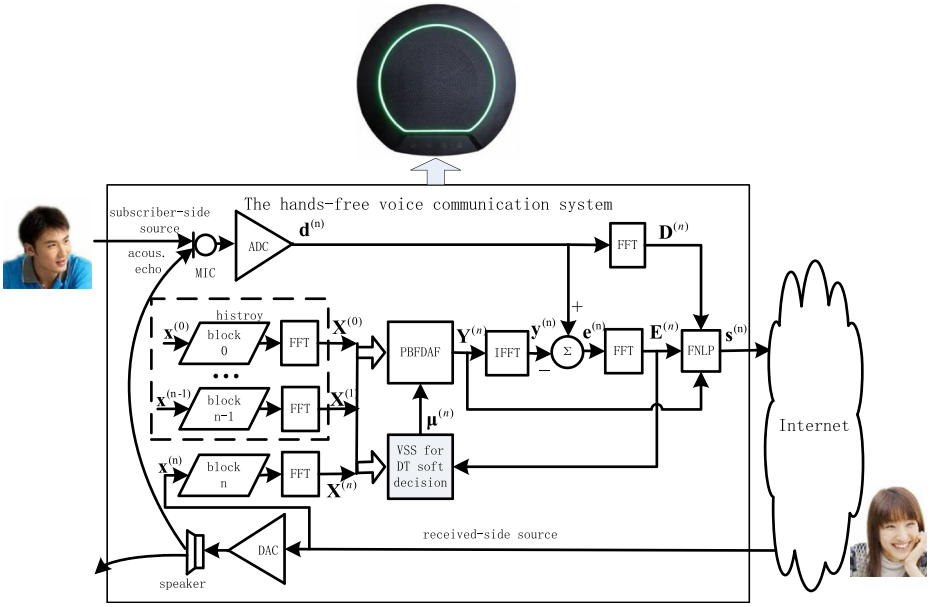


Fig. 1 Block diagram of the frequency-domain nonlinear echo processing algorithm

where n is the index of the blocks and depicts the n-th iteration of $\widehat{\mathbf{w}}_p$. N is the whole length of the filter. P is the length of a block $\widehat{\mathbf{w}}_p$. Then, $\frac{N}{P}$ is the number of blocks. To simplify the algorithm, it is assumed that P is set to be equal to L (the length of segments $\mathbf{x}^{(n)}$, $\mathbf{d}^{(n)}$ and $\mathbf{y}^{(n)}$).

By using (1), the frequency-domain filter coefficient of a block is given by

$$\widehat{\mathbf{W}}_p^{(n)} = FFT\begin{bmatrix} \widehat{\mathbf{w}}_p^{(n)} \\ \mathbf{0}_{M-P} \end{bmatrix}_{M\times 1} \tag{2}$$

Where M is the number of points of the FFT. Correspondingly, the frequency-domain referenced signal can be written as follows:

$$\mathbf{X}_p^{(n)} = diag\left\{ FFT\begin{bmatrix} x[(n+1)L-pP-M+1] \\ \vdots \\ x[(n+1)L-pP] \end{bmatrix} \right\}_{M\times M}, p = 0, \quad 1, \quad \cdots, \quad \frac{N}{P}-1 \tag{3}$$

The estimated echo signal $\widehat{\mathbf{y}}_p^{(n)}$ is calculated by the following steps: 1) summing the results of filtering $\mathbf{X}_p^{(n)}$ by $\widehat{\mathbf{W}}_p^{(n)}$, and 2) extracting the latest L points to avoid the effect of the overlap-save method.

$$\widehat{\mathbf{y}}^{(n)} = \begin{bmatrix} 0_{M-L} & 0 \\ 0 & \mathbf{I}_L \end{bmatrix} IFFT\left[\sum_{p=0}^{N/P-1} \mathbf{X}_p^{(n)} \widehat{\mathbf{W}}_p^{(n)} \right] \tag{4}$$

Then, the residual signal is

$$\mathbf{e}^{(n)} = \mathbf{d}^{(n)} - \widehat{\mathbf{y}}^{(n)} \tag{5}$$

$$\mathbf{E}^{(n)} = FFT\begin{bmatrix}\mathbf{0}_{M-P}\\ \mathbf{e}^{(n)}\end{bmatrix}_{M\times 1} \tag{6}$$

Then, a variable step size scheme is proposed to adjust the speed of the adaptive update process considering double-talk.

$$\widehat{\mathbf{W}}_p^{(n+1)} = \widehat{\mathbf{W}}_p^{(n)} + FFT\left[\Phi_p^{(n)}\right], p = 0, \quad 1, \quad \cdots, \quad \frac{N}{P}-1 \tag{7}$$

$$\Phi_p^{(n)} = \begin{bmatrix}\mathbf{I}_L & \mathbf{0}\\ \mathbf{0} & \mathbf{0}_{M-L}\end{bmatrix} IFFT\left[\prod_p^{(n)}\mathbf{X}_p^{(n)*}\mathbf{E}^{(n)}\right] \tag{8}$$

$$\prod_p^{(n)} = diag\left[\mu_{pi}^{(n)}\right] = \begin{bmatrix}\mu_{p0}^{(n)} & 0 & 0 & 0\\ 0 & \mu_{p1}^{(n)} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \mu_{p(M-1)}^{(n)}\end{bmatrix} \tag{9}$$

where '*' indicates the conjugate of a complex vector. $\mu_{pi}^{(n)}$, $i = 0, \quad 1, \quad \cdots, \quad M-1$ are variable step sizes that vary with the adjusted factor $\lambda_{pi}^{(n)}$.

Suppose that

$$\mu_{pi}^{(n)} = \frac{\mu_0}{\sum_{p=0}^{N/P-1}\mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*}}\lambda_{pi}^{(n)}, i = 0, \quad 1, \quad \cdots, \quad M-1 \tag{10}$$

where

$$\lambda_{pi}^{(n)} = \begin{cases}\dfrac{\Delta}{\left|\mathbf{E}_i^{(n)}\right|^2 \Big/ \mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*}}, & \left|\mathbf{E}_i^{(n)}\right|^2 \Big/ \mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*} > \Delta\\ 1, & \left|\mathbf{E}_i^{(n)}\right|^2 \Big/ \mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*} \le \Delta\end{cases} \tag{11}$$

In (10) and (11), $\mu_0 \in [0, 1]$ is a fixed step size, and $\frac{1}{\sum_{p=0}^{N/P-1}\mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*}} i = 0, \quad 1, \quad \cdots, \quad M-1$ are normalized factors of the i-th subband. $\left|\mathbf{E}_i^{(n)}\right|^2 \Big/ \mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*}$ depicts the relative strength of the near-end voice compared to the referenced signal. The larger that $\left|\mathbf{E}_i^{(n)}\right|^2 \Big/ \mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*}$ is, the higher the probability of double-talk. When $\left|\mathbf{E}_i^{(n)}\right|^2 \Big/ \mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*}$ is larger than the threshold Δ, the adjusted factor $\lambda_{pi}^{(n)}$ is defined as $\frac{\Delta}{\left|\mathbf{E}_i^{(n)}\right|^2 \Big/ \mathbf{X}_{pi}^{(n)}\mathbf{X}_{pi}^{(n)*}}$ and the step size of the update is diminished according to

the extent of double-talk to avoid divergence. If the signal from the far-end is silent, $\lambda_{pi}^{(n)} \to 0$ and $\widehat{\mathbf{W}}_p^{(n)}$ is frozen to avoid divergence.

3.3 Frequency domain nonlinear processing considering double talk

For the adaptive echo cancellation algorithmof hands-free voice communication devices, the acoustic echo path is modeled as a linear filter with the coefficient $\widehat{\mathbf{W}}$. However, the acoustoelectric transduction process of the microphone, the electroacoustic transformation process of the speaker, and the reflection coefficients of the obstacles in a room are all non-linear to some extent. Therefore, the residual echo signal is unavoidable after applying the adaptive linear echo cancellation algorithm. How to eliminate non-linear residual echoes and preserve near-end speech when double-talk exists is an issue of concern.

In this paper, a non-linear suppression function is constructed according to the relevance and correlations between the referenced signal $\mathbf{x}^{(n)}$, the microphone signal $\mathbf{d}^{(n)}$ and the estimated linear echo signal $\mathbf{y}^{(n)}$. Then, the suppression function inhibits the residual nonlinear echoes by using a set of coefficients for each frequency. The details are as follows.

Step 1: For the new block, calculate the frequency-domain signals $\mathbf{E}^{(n)}$, $\mathbf{X}^{(n)}$, $\mathbf{D}^{(n)}$ and $\mathbf{Y}^{(n)}$ by using the M-points FFT transform.

Step 2: Calculate the cross power spectra in the frequency domain as

$$Sxd_j^{(n)} = \mathbf{X}_j^{(n)}\mathbf{D}_j^{*(n)}, j = 0, 1, 2, \cdots, \mathrm{M}-1 \tag{12}$$

$$Syd_j^{(n)} = \mathbf{Y}_j^{(n)}\mathbf{D}_j^{*(n)}, j = 0, 1, 2, \cdots, \mathrm{M}-1 \tag{13}$$

$$Sde_j^{(n)} = \mathbf{D}_j^{(n)}\mathbf{E}_j^{*(n)}, j = 0, 1, 2, \cdots, \mathrm{M}-1 \tag{14}$$

Where '*' is the symbol of the conjugate and j is the frequency index.

Step 3: Calculate the cross correlations in the frequency domain.

$$Rxd_j^{(n)} = \frac{Sxd_j^{(n)} \cdot Sxd_j^{*(n)}}{\mathbf{X}_j^{(n)}\mathbf{X}_j^{*(n)}\mathbf{D}_j^{(n)}\mathbf{D}_j^{*(n)}}, j = 0, 1, 2, \cdots, \mathrm{M}-1 \tag{15}$$

$$Ryd_j^{(n)} = \frac{Syd_j^{(n)} \cdot Syd_j^{*(n)}}{\mathbf{Y}_j^{(n)}\mathbf{Y}_j^{*(n)}\mathbf{D}_j^{(n)}\mathbf{D}_j^{*(n)}}, j = 0, 1, 2, \cdots, \mathrm{M}-1 \tag{16}$$

$$Rde_j^{(n)} = \frac{Sde_j^{(n)} \cdot Sde_j^{*(n)}}{\mathbf{D}_j^{(n)}\mathbf{D}_j^{*(n)}\mathbf{E}_j^{(n)}\mathbf{E}_j^{*(n)}}, j = 0, 1, 2, \cdots, \mathrm{M}-1 \tag{17}$$

For a practical application system, $Rxd_j^{(n)}$, $Ryd_j^{(n)}$ and $Rde_j^{(n)}$ are usually smoothed by using historical values to avoid sudden changes.

Step 4: Calculate the average values of the above cross correlations to assess the degree of the residual echo and the near-end signal.

$$\overline{R}xd^{(n)} = \frac{\sum_{j=0}^{M-1} Rxd_j^{(n)}}{M} \tag{18}$$

$$\overline{R}yd^{(n)} = \frac{\sum_{j=0}^{M-1} Ryd_j^{(n)}}{M} \tag{19}$$

$$\overline{R}de^{(n)} = \frac{\sum_{j=0}^{M-1} Rde_j^{(n)}}{M} \tag{20}$$

Normally, the larger that $\overline{R}xd^{(n)}$ and $\overline{R}yd^{(n)}$ are, the heavier the residual echo. The larger that $\overline{R}de^{(n)}$ is, the higher the probability of the near-end source. When double-talk exists, $\overline{R}xd^{(n)}$ is smaller than that for single far-end talk but larger than that for single near-end talk. The varied $\overline{R}yd^{(n)}$ and $\overline{R}de^{(k)}$ also reflect the degree of double-talk.

Step 5: Define the non-linear echo suppression function as

$$F_j^{(n)} = {\varphi_j^{(n)}}^{\gamma_j^{(n)}} \tag{21}$$

where

$$\gamma_j^{(n)} = \frac{W_{xd} \cdot \left(1 - \overline{R}xd^{(n)}\right) + W_{yd} \cdot \left(1 - \overline{R}yd^{(n)}\right) + W_{de} \cdot \overline{R}de^{(n)}}{W_{xd} + W_{yd} + W_{de}} \tag{22}$$

W_{xd}, W_{yd} and W_{de} are the weights of $\overline{R}xd^{(n)}$, $\overline{R}yd^{(n)}$ and $\overline{R}de^{(n)}$ that satisfy

$$W_{xd} + W_{yd} + W_{de} = 1 \tag{23}$$

$\varphi_j^{(n)}$ is determined by the cross correlation values of each frequency j.

$$\varphi_j^{(n)} = \begin{cases} Rde_j^{(n)}, & \overline{R}xd^{(n)} < T_1 \quad and \quad \overline{R}de^{(n)} > T_2 \\ 1 - Rxd_j^{(n)} & \overline{R}xd^{(n)} \geq T_1 \quad or \quad \overline{R}de^{(n)} \leq T_2 \end{cases} \tag{24}$$

T_1 and T_2 are thresholds between 0 and 1. When $\overline{R}xd^{(n)} < T_1$ and $\overline{R}de^{(n)} > T_2$, the far-end referenced signal is very slight. The system is still silent or single near-end talk. When $\overline{R}xd^{(n)} \geq T_1$ or $\overline{R}de^{(n)} \leq T_2$, there is a far-end referenced signal, which means that echoes exist. Then, either$\varphi_j^{(n)} = Rde_j^{(k)}$ or $\varphi_j^{(n)} = 1\text{-}Rxd_j^{(k)}$ is between 0 and 1. When there is not a far-end signal, $\varphi_j^{(n)} = Rde_j^{(n)} = 1$. When there is single far-end talk, $\varphi_j^{(n)} = 1\text{-}Rxd_j^{(n)} \rightarrow 0$. When the system is experiencing double-talk, both the far-end signal and the near-end signal exist. In addition, $\varphi_j^{(n)}$, where $j = 0, 1, \cdots, M-1$, varies according to the cross correlation of $\overline{R}xd^{(n)}$, in which the extent of double-talk is implicit.

Step 6: Calculate the frequency-domain output signal by using the residual echo signal through the non-linear echo suppression function.

$$S_j^{(n)} = F_j^{(n)} E_j^{(n)} = \varphi_j^{(n)^{\gamma_j^{(n)}}} E_j^{(n)}, j = 0, 1, \cdots, M-1 \tag{25}$$

At last, the time-domain output signal $\mathbf{s}^{(n)}$ is calculated by using the IFFT transform of $\mathbf{S}^{(n)} = \left\{ S_j^{(n)}, j = 0, 1, \cdots, M \right\}$.

4 Real-time implementation and experiment by using a prototype device

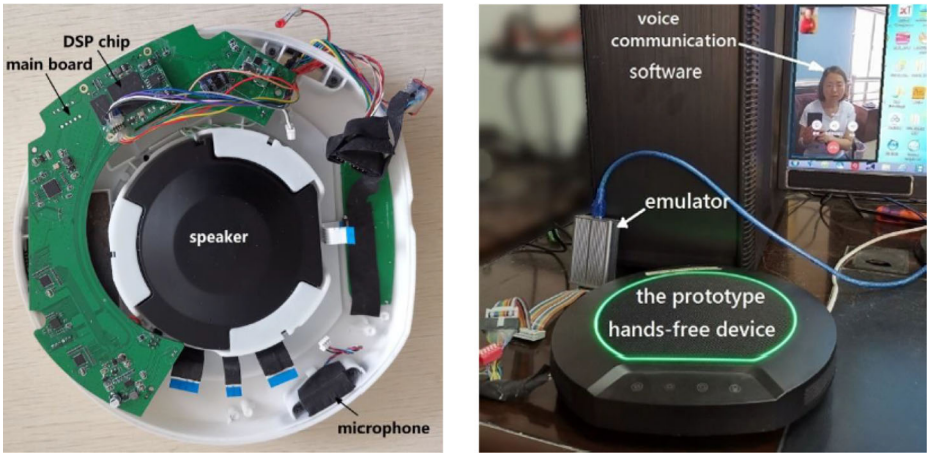
We implemented the proposed frequency-domain nonlinear echo processing algorithm in a prototype hands-free voice communication device and measured its performance in an actual room by using a popular voice communication software. The algorithm was programmed by using C language and implemented in a DSP chip. The prototype device consists of an electret condenser microphone, an audio codec chip (that includes AD and DA modules), a DSP mainboard, a power amplifier and a full-band loudspeaker, which is shown in Fig. 2.

4.1 Experimental environment

To compare the proposed algorithm with other acoustic echo cancellation algorithms, we measured the room impulse response(RIR) of the room (which is almost 2.5 m wide, 4 m long and 3 m high) using the M-series signals that were generated by a computer software, played over a loudspeaker and recorded by a measuring microphone. The sampling rate of the signals was $f_s = 16kHz$. The recorded signals were deconvoluted, and the measured room impulse response was truncated at the length of 2048 taps (corresponding to a 128 ms tail length). The measured room and its room impulse response (RIR) are shown in Fig. 3.

Additionally, we measured the actual received-side signal (Rin) and the synchronized subscriber-side signal (Sin) by placing the prototype hands-free voice communication device on a table in the room. The diagram of the prototype system is depicted in Fig. 4.

The ADC and DAC were implemented by using a codec with a sampling rate of $f_s = 16kHz$. The prototype device was connected to a computer through a USB cable and the data of the voice communication software were transferred through the USB-audio-class (UAC)



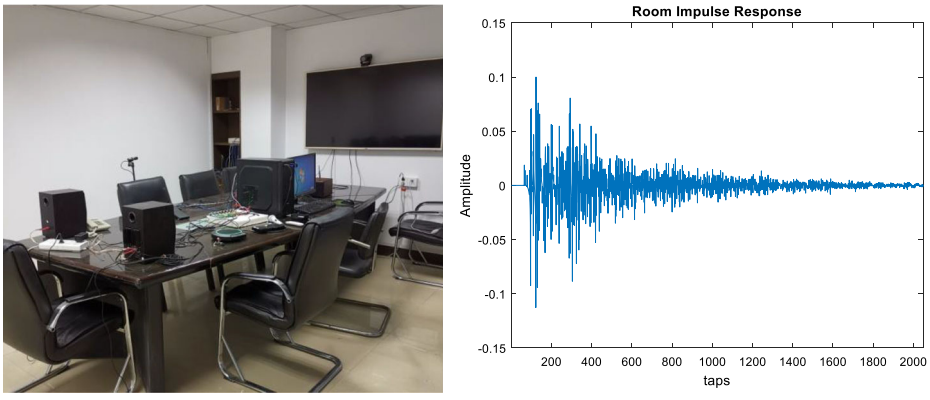
(a) (b)

Fig. 2 The implementation of the algorithm in the prototype device. **a** the hardware of the prototype device, **b** debugging the prototype device

architecture of the USB 2.0 protocol. When the prototype device was used for measuring input signals, the algorithms were bypassed. The computer outputs the referenced audio signal $\mathbf{x}^{(n)}$, which was played by the speaker. The subscriber-side signal $\mathbf{d}^{(n)}$ was recorded by the microphone and transferred to the computer.

4.2 Experiments on acoustic echo cancellation with single far-end talk

When the subscriber-side source was silent, the microphone received only the echoes from the speaker to the microphone through the room impulse response. The experiments with single far-end talk were compared to the performances of several classical AEC algorithms, such as



(a) (b)

Fig. 3 The measured room and its room impulse response with 2048 taps. **a** the measured room, **b** the RIR of the room

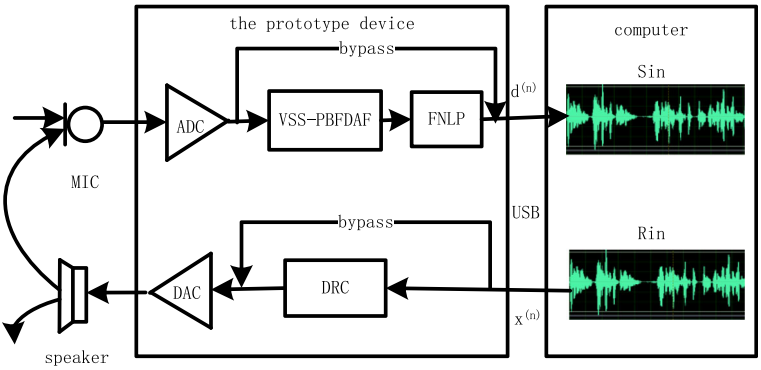


Fig. 4 The diagram of the prototype voice communication system

NLMS, FNLMS and PBFDAF, using the simulated signals that were generated by the RIR, as shown in Fig. 3.

Two types of Rin signals were simulated as the reference far-end signals. One type was a speech signal and the other was white noise. Figure 5 shows two groups of Rin signals and the simulated Sin signals. The sampling rate of these signals was 16 kHz. The misalignment of the AEC algorithm is defined as

$$misalignment = 10\lg\frac{\left\|\mathbf{w}-\widehat{\mathbf{w}}\right\|^2}{\|\mathbf{w}\|^2} \tag{26}$$

where $\mathbf{w}$ is the actual acoustic room impulse response that is shown in Fig. 3(b) and $\widehat{\mathbf{w}}$ is the estimated filter coefficients vector.

Three AEC algorithms NLMS, FNLMS and traditional PBFDAF were simulated. The filters of the NLMS and FNLMS were set to 2048 taps with $\mu = 0.5$. For FNLMS, Rin and Sin were segmented to blocks by using a filter length of 2048 and the coefficients of the filters were updated by each block in the frequency domain. PBFDAF partitioned the 2048-tap filter into *16* subblocks. The time delay of the buffer preparation of PBFDAF was then *1/16* that of FNLMS and the number of iterations in PBFDAF was *16* times that in FNLMS.

As shown in Fig. 6, the three algorithms achieved good misalignment performance. When a speech signal was used as the Rin, both NLMS and FNLMS achieved lower than −30 dB misalignment. PBFDAF achieved faster and deeper convergence than NLMS and FNLMS, but it fluctuated along with the speech amplitude. When white noise was used as the Rin, the three algorithms achieved misalignments lower than −70 dB. PBFDAF achieved the fastest convergence speed. Comparing the computational complexity and time delay of the three algorithms, FNLMS and PBFDAF achieve lower computational complexity and time delays than NLMS. Summarizing the above experiments, traditional PBFDAF has the advantages of fast convergence, low computational complexity and considerable convergence. Therefore, PBFDAF can be widely applied in real-time voice communication systems. However, the performance of PBFDAF is influenced by the fluctuation of the input waveform. When the system is experiencing double-talk, the traditional PBFDAF algorithm might diverge.

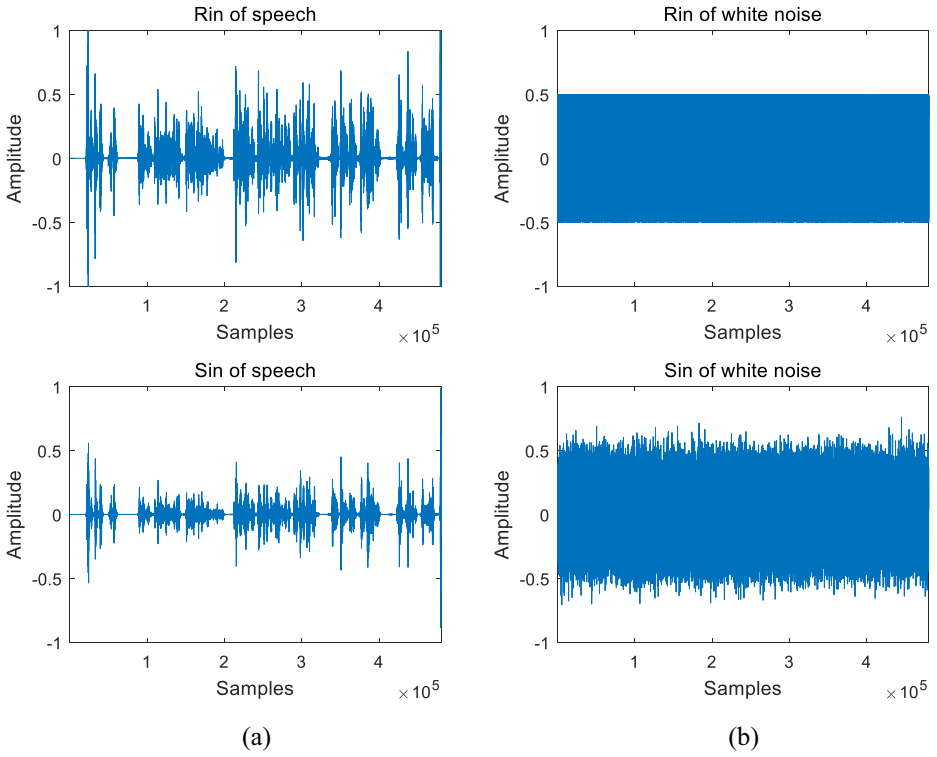


Fig. 5 The typical Rin and Sin waveforms of speech and white noise. **a** Rin and Sin of the speech input, **b** Rin and Sin of the white noise input

4.3 Experiments on acoustic echo cancellation during double-talk

Experiments on the acoustic echo cancellation of the proposed VSS-PBFDAF, the traditional PBFDAF, the NLMS and the FLMS algorithms when double-talk exists were conducted to

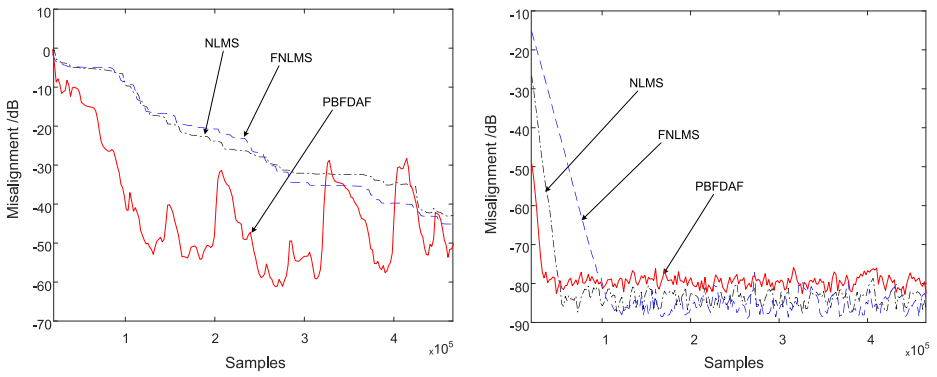


Fig. 6 Misalignments of NLMS, FNLMS and PBFDAF with two types of inputs

compare the performance of these algorithms. The two groups of tested Rin and Sin signals (one was the typical speech input and the other was white noise input) are shown in Fig. 7. The sampling rate of these signals was 16 kHz.

The output signal waveforms of NLMS, FNLMS and the proposed VSS-PBFDAF with $\Delta = 1$, 0.3 and 0.1 are shown in Fig. 8. Figures 8(a), (c), (e), (g), and (i) are the waveforms of the algorithms with speech as the input. On the double-talk segment, the output of NLMS has an obvious distortion, which means a large degree of divergence. In comparison, the output of VSS-PBFDAF with $\Delta = 0.1$ achieved the lowest distortion. These output signals were played and the subject assessed that Fig. 8(i) provided the clearest near-end speech during the double-talk period. Figures 8(b), (d), (f), (h), and (j) are the waveforms of the algorithms with white noise as the input. All the waveforms of the near-end signal during double-talk experienced low distortion and clear audition. However, the three VSS-PBFDAF algorithms (with $\Delta = 1$, 0.3 and 0.1) achieved faster convergence than NLMS and FNLMS.

The misalignments with speech and white noise as inputs are shown in Fig. 9. From Fig. 9(a), the proposed VSS-PBFDAF with $\Delta = 1$ achieved the deepest and fastest convergence compared to the other algorithms. However, the misalignment rapidly increases when double-talk exists. VSS-PBFDAF with $\Delta = 0.3$ and 0.1 achieved lower divergence than that of $\Delta = 1$. Overall, the misalignments of VSS-PBFDAF with $\Delta = 1$, 0.3 and 0.1 were lower than those of NLMS and FNLMS when speech was the input. Figure 9 (b) shows the misalignments of these algorithms when white noise was the

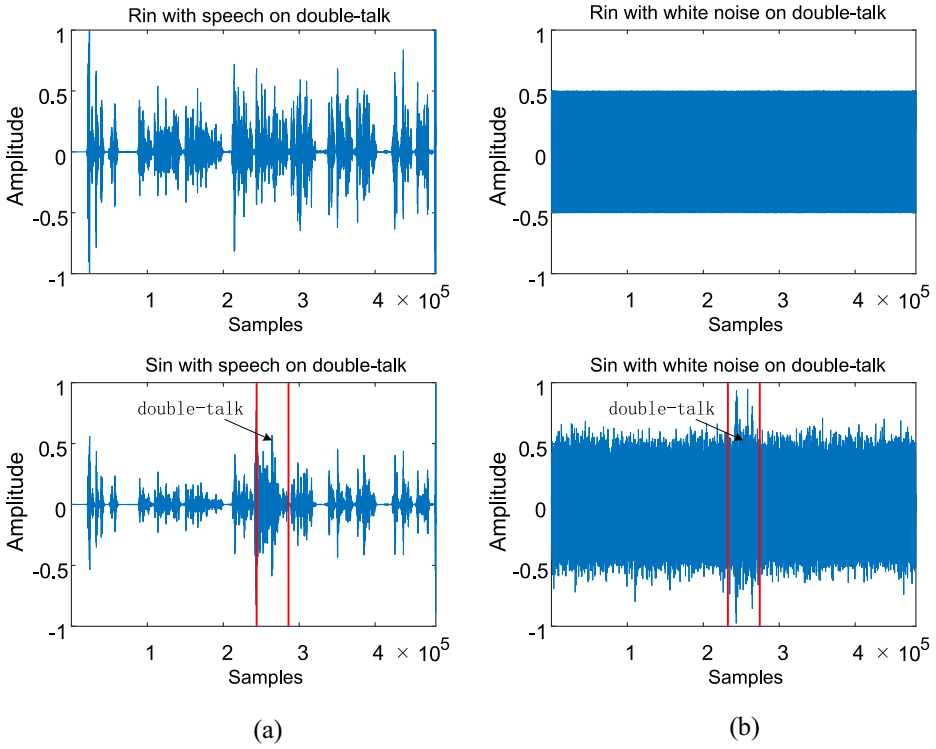


Fig. 7 Two groups of Rin and Sin during double-talk with speech and white noise input. **a** Rin and Sin during DT with speech, **b** Rin and Sin during DT with white noise

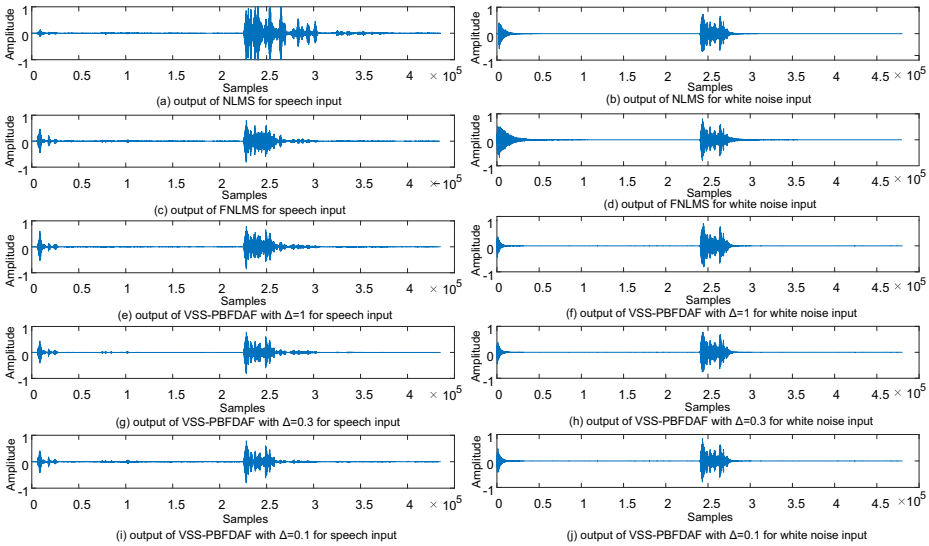


Fig. 8 Output waveforms of the algorithms with double talk for speech and white noise inputs

input. The three VSS-PBFDAF algorithms achieved faster and deeper convergence on single far-end and double-talk than NLMS and FNLMS.

4.4 Experiments on speech quality assessment with the prototype voice communication device

Nonlinear acoustic echo is a common problem for real-time voice communication systems. The inhibition of nonlinear acoustic echo usually degrades the speech quality. We implemented the proposed VSS-PBFDAF and FNLP algorithms in the prototype hands-free voice communication device and measured the output speech quality in the actual room. The diagram of the prototype device is shown in Fig. 10. Since the prototype device consists of complete functional modules, the subscriber-side signal (with mixed acoustic echo and near-end speech) was processed by HP (high pass), synchronization, VSS-PBFDAF, FNLP, CNG

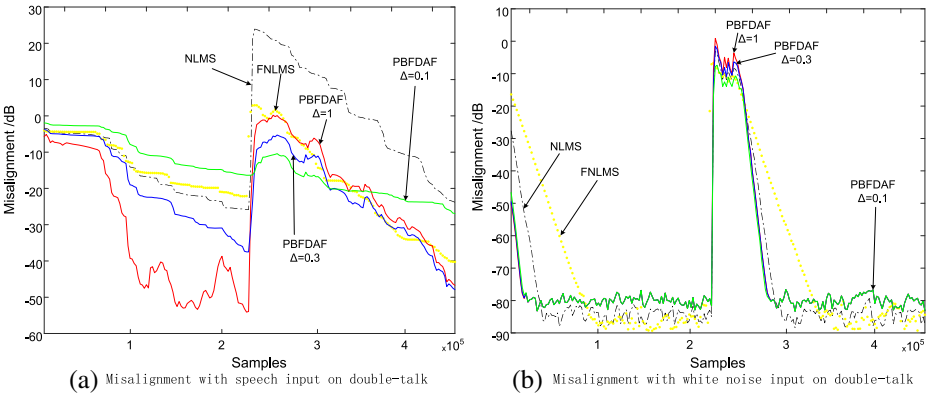


Fig. 9 Misalignment of algorithms with double-talk for speech and white noise inputs

(comfort noise generator), NS (noise suppression), and AGC (automatic gain control) modules, and then transferred to a PC through a USB interface. The received input from the far-end was processed by HP (high pass), EQ (equalizer), and DRC (dynamic range compression) modules, and then output to the speaker through a DAC and power amplifier. The Sin and Rout signals were synchronized and sent to the proposed VSS-PBFDAF and FNLP algorithms.

Figure 11 shows one typical measuring signal of the Sin and Sout of the system after the HP, VSS-PBFDAF, FNLP, CNG, NS and AGC modules. For the time period from the 4-th second to the 20-th second, the Rin was far-end speech and the near-end voice was silent, and then the Sin was only the echo from the speaker to the microphone. In this situation of single far-end talk, the average root mean square (RMS) power of the Sout was approximately −63 dB, while the average RMS power of the Sin was −37 dB. The echo energy of the Sout was 26 dB lower than that of the Sin. Because the echoes lower than −55 dB were almost imperceptible by the consumer at the far end, the convergence performance of the system was satisfied. For the time period from the 20-th second to the 32-th second, the far-end speech and the near-end source simultaneously existed and the system experienced double-talk. From the results of Fig. 11, we can see that during the double-talk period, the near-end speech passed uninterrupted through VSS-PBFDAF and FNLP. The output signal were played and the subject assessed that the near-end speech was fluent and undistorted.

The ITU-T P.563 standard is a single-ended method for objective speech quality assessment that was established by the International Telecommunication Union [32]. Compared to other speech quality assessment methods such as P.862 and P.863, it is applicable for speech quality predictions without a separate reference signal. Therefore, this method is recommended for nonintrusive real time speech quality assessments. The P.563 assessment is scored by using the MOS-LQO. We compared the speech quality scores of the Sout based on P.563 that were processed by the proposed frequency-domain nonlinear echo processing algorithm and WebRTC (Web Real-Time Communication, which is an open source audio and video communication framework) [17]. For the single near-end talk situation, the far-end source was silent. The near-end source was played 1 m and 3 m away from the microphone. For the double-talk situation, the computer played a speech file as the far-end signal. Additionally, the near-end source was played 1 m and 3 m away from the microphone. Table 1 shows the scores of these two algorithms that were implemented and tested by using the prototype hands-free voice communication device in the room that was depicted in Fig. 3.

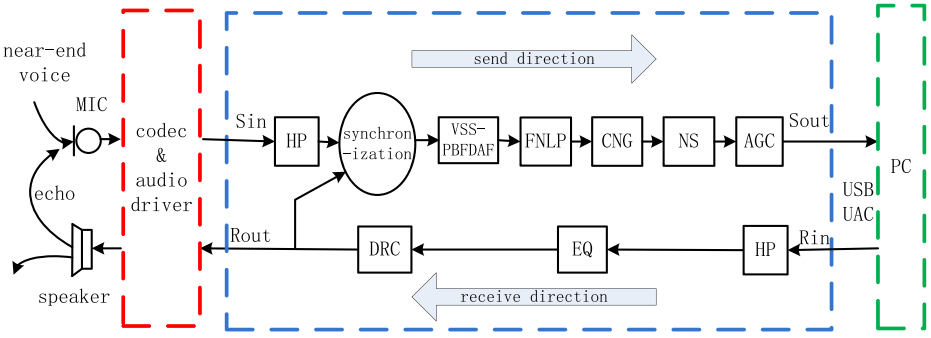


Fig. 10 The diagram of the prototype device with the complete functional modules

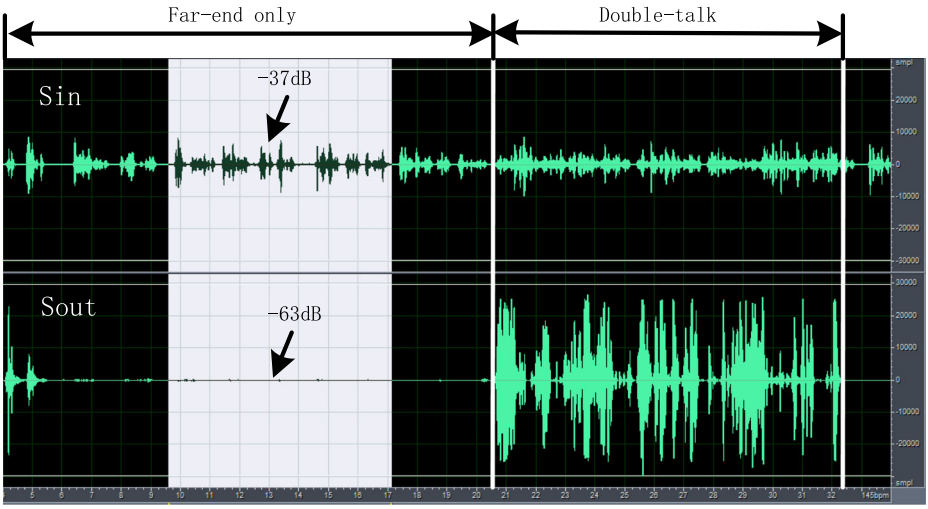


Fig. 11 One typical measuring signal of the Sin and Sout of the prototype device

From the results of Table 1, both the proposed algorithm and WebRTC achieved high scores for single near-end talk. As the distance between the near-end source and the microphone increased, the speech quality decreased. When double-talk existed, the proposed algorithm achieved a higher P.563 score than that of WebRTC. When the near-end source was played 1 m away from the microphone, the P.563 score of the Sout for the proposed algorithm was 3.0. By subjective assessment, it sounded clear and coherent. In the same situation, the score of the Sout by using WebRTC was 2.3, and the Sout sounded almost clear but was missing some syllables. When the distance between the near-end source and the microphone was 3 m long, the speech quality of the Sout by using the proposed algorithm decreased from 3.0 to 2.1. The signal-to-noise ratio (SNR) degraded, but the speech was almost fluent without any truncations. In the same situation, the speech quality of the Sout by using WebRTC decreased from 2.3 to 1.2. There were many truncations in the output speech, and it was difficult to understand the sentence.

4.5 Discussion

Overall, our studies established a prototype device to verify the proposed algorithm under the situation of real-time voice communication. In linear AEC, we achieved only 4 ms latency through a rather long adaptive filter (2048 taps) estimation, owing to the sub-block process. In nonlinear acoustic processing, residual acoustic echoes were eliminated efficiently in single far-end talk mode while the near-end speeches were preserved clearly in case of double-talk

Table 1 P.563 scores of the Sout by using the proposed algorithm and WebRTC

	Distance of the source	The proposed algorithm	WebRTC
Single near-end talk	1 m	4.4	4.3
	3 m	3.7	3.5
Double-talk	1 m	3.0	2.3
	3 m	2.1	1.2

mode. The results suggest that the non-linear echo suppression function varies according to the communication status and makes a good tradeoff between echo inhibition and speech quality. Furthermore, due to the spectral correction modulated by the correlations of the current microphone signal, the residual echo and the synchronized referenced signal per frame, the proposed nonlinear echo processing algorithm acquires low computation complexity and is easy to be implemented to the DSP platforms. Although there are some improvements achieved by these studies, there are also limitations. Nonlinear echo processing algorithms based on spectral correction only consider the direct sound and the early reflection of the echoes and ignore the late reflection, which leads to slight speech distortion on double-talk. Future work will be the study of low-complexity echo cancellation algorithms with high-quality subscribed-side speech while under the double-talk situation.

5 Conclusion

Following the studies focusing on acoustic echo cancellation when double-talk exists for hands-free voice communication devices, we evaluated a scheme based on the VSS-PBFDAF and FNLP algorithms, which was implemented by using a practical DSP prototype device. The experiments showed that compared with the traditional PBFDAF, NLMS and FLMS algorithms, the proposed algorithm was more robust, achieved faster convergence and maintained a clear near-end voice due to the variable step-sizes for each frequency by using frequency-domain nonlinear echo processing. When measured by using the prototype device based on the DSP platform, the proposed algorithm achieved higher speech quality than WebRTC, and the output when double-talk existed was clear and coherent.

Acknowledgements This work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC2004003 and Grant 2020YFC2004002. The authors would like to thank the reviewers for their valuable comments that helped in significant improvement of the quality of the paper. They would also like to thank Professor Zou Cairong for the suggestions of experimental analysis and discussions.

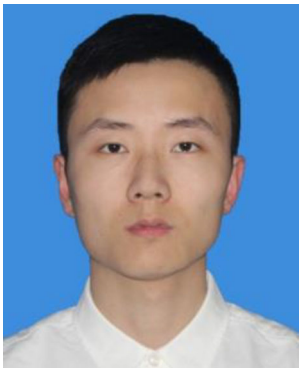
References

1. Ahgren P, Jakobsson A (2006) A study of doubletalk detection performance in the presence of acoustic echo path changes. IEEE Trans Consum Electron 52(2):515–522
2. Azpicueta-Ruiz LA, Zeller M, Figueiras-Vidal AR (2011) Adaptive combination of volterra kernels and its application to nonlinear acoustic echo cancellation. IEEE Transactions on Audio, Speech, and Language Processing 19(1):97–110
3. Bekrani M, Khong AWH, Lotfizad M (2011) A linear neural network-based approach to stereophonic Acoustic Echo cancellation. IEEE Trans Audio Speech Lang Process 19(6):1743–1753
4. Bernardi G, Waterschoot TV, Wouters J, Moonen M (2015) An all-frequency-domain adaptive filter with PEM-based decorrelation for acoustic feedback control. in 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA):1–5
5. Birkett AN, Goubran RA (1995) Nonlinear echo cancellation using a partial adaptive time delay neural network. in Neural Networks for Signal Processing:449–458
6. Cecchi S, Romoli L, Piazza F (2016) Multichannel double-talk detector based on fundamental frequency estimation. IEEE SIGNAL PROCESSING LETTERS 23(1):94–97
7. Comminiello D, Scarpiniti M, Azpicueta-Ruiz LA, Arenas-García J, Uncini A (2013) Functional link adaptive filters for nonlinear acoustic echo cancellation. IEEE Transactions on Audio Speech & Language Processing 21(7):1502–1512

8. Comminiello D, Scarpiniti M, Azpicueta-Ruiz LA, Arenas-Garcia J, Uncini A, Full proportionate functional link adaptive filters for nonlinear acoustic echo cancellation, in European Signal Processing Conference 2017. 1145–1149.
9. Eneman K, Moonen M (2003) Iterated partitioned block frequency-domain adaptive filtering for acoustic echo cancellation. IEEE Transactions on Speech & Audio Processing 11(2):143–158
10. Enhanced ITU-T G.168 echo cancellation. 2000, ITU. 128.
11. Faller C, Tournery C, Robust Acoustic ECHO Control using a simple ECHO path model, in IEEE international conference on acoustics, Speech & Signal Processing. 2006.
12. Fukui M, Shimauchi S, Hioka Y, Nakagawa A, Haneda Y (2014) Double-talk Robust Acoustic Echo cancellation for CD-quality hands-free videoconferencing system. IEEE Trans Consum Electron 60(3):468–475
13. Gansler T, Gay SL, Sondhi M, Benesty J (2000) Double-talk robust fast converging algorithms for network echo cancellation. Speech & Audio Processing IEEE Transactions on 8(6):656–663
14. Guerin A, Faucon G, Bouquin-Jeannes RL (2003) Nonlinear acoustic echo cancellation based on Volterra filters. IEEE Transactions on Speech and Audio Processing 11(6):672–683
15. Halimeh MM, Huemmer C, Kellermann W (2019) A neural network-based nonlinear Acoustic Echo canceller. IEEE Signal Processing Letters 26(12):1827–1831
16. Huang F, Zhang J, Zhang S (2018) Affine projection Versoria algorithm for Robust adaptive Echo cancellation in hands-free voice communications. IEEE Trans Veh Technol 67(12):11924–11935
17. Inc. G. WebRTC. https://webrtc.org/start/#2011
18. Jiang T, Liang R, Wang Q, Zou C, Li C (2019) An improved practical state-space FDAF with fast recovery of abrupt Echo-path changes. IEEE Access 7(1):61353–61362
19. Jose M. Gil-Cacho M S, Toon Vanwaterschoot, Marc Moonen. Nonlinear acoustic echo cancellation based on a sliding-window leaky kernel affine projection algorithm. IEEE Trans Audio Speech Lang Process, 2013, 21(9): 1867–1878
20. Lee GW, Lee JH, Moon JM, Kim HK (2019) Non-linear acoustic echo cancellation based on mel-frequency domain volterra filtering. 2019 IEEE International Conference on Consumer Electronics (ICCE):1–2
21. Lei Q, Chen H, Hou J, Chen L, Dai L (2019) Deep neural network based regression approach for acoustic echo cancellation, in 4th International Conference on Multimedia Systems and Signal Processing, ICMSSP 2019, May 10, 2019 - May 12, 2019. Association for Computing Machinery: Guangzhou, China. 94-98.
22. Li X, Jenkins WK (1996) The comparison of the constrained and unconstrained frequency-domain block-LMS adaptive algorithms. IEEE Trans Signal Process 44(7):1813–1816
23. Liu J (2004) Efficient and robust cancellation of echoes with long echo path delay. Communications IEEE Transactions on 52(8):1288–1291
24. Long G, Ling F, Proakis JG (1989) The LMS Algorithm with delayed coefficient adaptation. IEEE Trans Acoust Speech Signal Process 40(9):1397–1405
25. Panda B, Kar A, Chandra M (2014) Non-linear adaptive echo supression algorithms: A technical survey. in International Conference on Communications and Signal Processing:076–080
26. Pao YH (1989) Adaptive pattern recognition and neural networks. Addison-Wesley
27. Papp II, Šarić ZM, Teslic N (2011) Hands-free voice communication with TV. IEEE Trans Consum Electron 57(2):606–614
28. Park Y-J, Park H-M (2010) DTD-free nonlinear acoustic echo cancellation based on independent component analysis. Electron Lett 46(12):866–869
29. Schwarz A, Hofmann C, Kellermann W, (2014) Spectral feature-based nonlinear residual echo suppression, in 2013 IEEE Workshop on Applications of Signal Processing To Audio and Acoustics. New Paltz, NY. 1-4.
30. Shynk JJ (2002) Frequency-domain and multirate adaptive filtering. IEEE Signal Process Mag 9(1):14–37
31. Tashev IJ (2012) Coherence based double talk detector with soft decision. in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP):165–168
32. Union T I T (2004) ITU P.563 Single-ended method for objective speech quality assessment in narrow-band telephony applications.
33. Waterschoot TV, Moonen M (2011) Fifty years of acoustic feedback control: state of the art and future challenges. Proc IEEE 99(2):288–327
34. Widrow B (2005) Thinking about thinking: the discovery of the LMS algorithm. IEEE Signal Process Mag 22(1):100–106
35. Yu D, Li J (2017) Recent progresses in deep learning based acoustic models. IEEE/CAA Journal of Automatica Sinica 4(3):396–409
36. Zhang H, Wang D, (2018) Deep learning for acoustic echo cancellation in noisy and double-talk scenarios, in 19th Annual Conference of the International Speech Communication, INTERSPEECH 2018, September

2, 2018 - September 6, 2018. International Speech Communication Association: Hyderabad, India. 3239-3243.
37. Zhang S, Zheng WX (2017) Recursive adaptive sparse exponential functional link neural network for nonlinear AEC in impulsive noise environment. IEEE Transactions on Neural Networks & Learning Systems PP(99):1–10

Qingyun Wang received the M.S. degree in Computer Engineering and the Ph. D. degree in Information and Communication Engineering from Southeast University, Nanjing, China, in 2001 and 2011 respectively. Now she is a professor of Nanjing Institute of Technology, Nanjing, China. Her current research interests include acoustic signal processing, speech enhancement and microphone array signal processing.



Xing Chen received the B.S. degree in Electrical engineering and automation from Yancheng Institute of Technology, Yancheng, China, in 2016. He is currently working toward the M.S. degree in the field of electrical engineering at Nanjing Institute of Technology, Nanjing, China. His research interests include the area of adaptive filtering algorithms and active noise reduction.

Ruiyu Liang received the Ph.D degree from Southeast University, China, in 2012. He is currently an associate professor with Nanjing Institute of Technology, Nanjing, Jiangsu province, China. His research interests include speech signal processing and signal processing for hearing aids.

Haicheng Liu received the B.S. degree in Electronic information engineering from Southeast University Chengxian College, Nanjing, China, in 2016. He is currently working toward the M.S. degree in the field of signal processing at Southeast University, Nanjing, China. His research interests include the area of adaptive filtering algorithms and speech enhancement.