Applied Acoustics 141 (2018) 9–18

Contents lists available at ScienceDirect

Applied Acoustics

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

An integrated acoustic echo and noise cancellation system using cross-band adaptive filters and wavelet thresholding of multitaper spectrum

E.P. Jayakumar*, P.S. Sathidevi

Department of ECE, National Institute of Technology Calicut, NITC Campus PO, Calicut-673601, Kerala, India

ARTICLE INFO

Keywords:
Speech
Acoustic echo
Noise cancellation
Cross-band adaptive filters
Wavelet packets

ABSTRACT

An efficient integrated system for acoustic echo and noise cancellation in hands free devices is proposed in this paper. Adaptive filters are usually employed in such systems for the acoustic echo cancellation. In order to reduce the computational complexity, sub band adaptive filters are commonly used since the order of filter required will be less but at the cost of distortion due to aliasing effect. In this work, cross-band adaptive filters are developed in wavelet domain/wavelet packet domain in tune with psychoacoustic model, to reduce the above distortion and thereby effectively estimate the echo and cancel it. The system uses wavelet thresholding technique for smoothening the log multitaper spectrum which eliminates the annoying musical noise in the processed output. To achieve efficient echo and noise cancellation, a novel switching mechanism to switch between wavelet filter bank and perceptual wavelet packet filter bank with most matching wavelet basis, from a dictionary of wavelet bases, depending on the type of noise present in each frame of the noisy speech is designed and developed in this work. The performance of the proposed integrated system is evaluated in terms of Acoustic Interference Cancellation (AIC) and Noise Attenuation (NA) under different Signal to Noise Ratio (SNR) conditions. The system exhibits very good performance even at very low SNRs close to −10 dB when compared with other such competitive systems.

1. Introduction

Hands-free mode in mobile phones helps the user to make normal conversation without holding the phone while the hands are engaged in another task. The flexibility it offers makes it an essential and popular feature in mobile phones. The user hears the speech from the distant person (far-end speech) through the loudspeaker and the speech of the user (near-end speech) is picked up by the microphone. Hands-free mode suffers from two major problems namely acoustic echo and environment noise which will affect the quality of conversation. Acoustic coupling between the loudspeaker and the microphone results in acoustic echo. The effect of acoustic echo is more annoying if the delay between the speech and its echo is more than a few tens of milliseconds [1]. The main source of noise is the background noise that depends on the place at which the mobile phone is used such as car, train, street, restaurant, airport and so on. Background noise affects the intelligibility of the speech. These two problems have been the topics of active research for the past several years. However, combined systems which deal with removing acoustic echo and background noise are limited.

In the case of Acoustic Echo Cancellation (AEC), several methods were proposed in the past based on adaptive filtering techniques. The basic principle of these methods is to adaptively estimate the Room Impulse Response (RIR) which in turn is used to estimate the acoustic echo. The estimated acoustic echo is then deducted from the microphone input signal which includes near-end speech, background noise and acoustic echo. Most common algorithms for adaptive filtering are the Least Mean Square (LMS), Normalized Least Mean Square (NLMS) [2] and the Recursive Least Squares (RLS) [3] methods. Improved versions of these algorithms are proposed in [4–7] which show better performance in terms of convergence and steady state error. A channel shortening filter was introduced to reduce the length of the effective acoustic echo path which reduces the computational complexity [8]. Frequency domain methods for echo cancellation were proposed in [9–12].

Several methods have been proposed for removing noise from the speech signal. The basic method is the spectral subtraction introduced by [13] where an estimate of the noise spectrum found during silence period is subtracted from the noisy speech spectrum to get an estimate of the clean speech spectrum. After subtraction, the noise which is left over is called as residual noise or musical noise and it requires further processing to reduce this noise. There are several variants of this spectral subtraction method which improves the quality of speech

* Corresponding author.
E-mail addresses: jay@nitc.ac.in (E.P. Jayakumar), sathi@nitc.ac.in (P.S. Sathidevi).

https://doi.org/10.1016/j.apacoust.2018.05.029
Received 27 August 2017; Received in revised form 14 May 2018; Accepted 30 May 2018

enhancement. In [14], authors have proposed a spectral floor to mask the musical noise. Non linear spectral subtraction [15] method subtracts larger values at frequencies with low SNR levels and smaller values at frequencies with high SNR levels assuming that noise does not affect all spectral components equally. In Multiband spectral subtraction method [16], the spectral subtraction is done in individual bands of the speech spectrum. A Minimum Mean Square Error (MMSE) Short Time Spectral Amplitude (STSA) estimator is used for speech enhancement in [17]. In [18], a method based on the masking property of human auditory system is suggested. Speech enhancement is achieved through cross correlation based sub band wiener filtering and the harmonic distortions introduced were reduced by regenerating the missing harmonics [19]. Hu and Loizou [20] used multitaper spectrum estimators and wavelet thresholded the log multitaper spectra to get spectral estimates with low variance which results less musical noise. Unlike [20] which used only Discrete Wavelet Transform (DWT) for wavelet thresholding, in [21], based on the type of noise present in the noisy speech, either DWT or Perceptual Wavelet Packet Transform (PWPT) with a suitable wavelet basis is used which further reduced the musical noise.

Most of the systems available today, handle the elimination of acoustic echo and noise separately. Only few integrated systems, which cancel the acoustic echo and environment noise together are available in the literature.

Martin and Altenhoner presented adaptive algorithms for improved echo attenuation with reduced complexity of implementation [22]. This approach combines an Finite Impulse Response (FIR) echo canceller with an NLMS-adapted FIR filter to attenuate residual echoes. The one-microphone system improves the echo attenuation, while the two-microphone approach can attenuate acoustic echoes, near-end speech reverberation and ambient noise. The performance of echo cancellation degrades with low SNR values, and noise reduction is not effective as there is low frequency residual noise in the output.

Gustafsson et al., [23,24] proposed a combined system based on MMSE criterion. It consists of a low order time domain acoustic echo compensator followed by a frequency domain adaptive filter for residual echo and noise reduction. The system flexibly combines separate estimations of Power Spectral Density (PSD) of the background noise and residual echo so that residual echo is masked by the intentionally left background noise. Though the system aims at effective removal of leftover echo and noise, the computational complexity is on the higher side as it requires an extra post-filter for completing the task.

Gustafsson et al., [25] proposed a structure which consists of an acoustic echo canceller followed by an adaptive postfilter. The adaptive postfilter attenuates the echo left by the echo canceller and the background noise. The estimated noise and residual echo power densities are combined adaptively to control the attenuation and masking of residual echo by a low level of intentionally left background noise. A psychoacoustically motivated weighting rule is used for the combined attenuation of the background noise and residual acoustic echo. Though the method claims that the overall computational complexity is reduced by having a low order acoustic echo canceller and post filter, the performance of the system under various noisy environments is not evaluated.

Another scheme proposed by Rombouts and Moonen [26], where, acoustic echo cancellation and noise reduction are combined as a single optimization problem and is solved adaptively using a QRD-based least squares lattice (QRD-LSL) algorithm. The derivation is based upon QR-decomposition (QRD)-based unconstrained optimal filtering methods for acoustic noise cancelers. By this technique, the filter length is significantly reduced resulting in low computational cost when compared to the filter length in traditional echo cancellers, without incurring a major performance loss. It trades off noise/echo reduction versus near-end distortion. But the method still involves more computations resulting in delay and the performance under different types of background noises is unknown.

Reuven et al. [27] introduces an echo transfer-function generalized sidelobe canceller (ETF-GSC) which is obtained by using the transfer-function generalized sidelobe canceller (TF-GSC) in parallel with an echo cancellation module. The noise reduction is achieved by the primary TF-GSC. Echo cancellation is achieved with the help of the secondary modified TF-GSC consisting of a replica of the primary TF-GSC components. This technique involves the use of an array of microphones, and hence the system is highly demanding in terms of computing power.

Most recent works on the topic includes the following. An integrated acoustic echo and background noise suppression method is proposed by [28] considering the near-end speech uncertainty. The noise and echo were estimated separately and the combined disturbance spectrum was obtained by summing the two estimates. But, the performance is seen to deteriorate under low SNR conditions and do not work well under all noisy environments. An improved integrated acoustic echo and noise suppression was proposed [29] using modulation spectral manipulation. Though the performance is good under different SNR conditions, the performance in the low SNR region still need to be improved.

A novel integrated system for effective echo cancellation and noise removal even under low SNR conditions is proposed in this paper. The log multitaper spectral components are decomposed into various sub bands using DWT or PWPT with a suitable wavelet basis based on the type of noise present in the input noisy speech. Sub band adaptive filtering of the DWT/PWPT coefficients of the log multitaper spectrum estimate (LMTS) is done to cancel the acoustic echo. Noise is removed based on a Short Time Spectral Amplitude (STSA) estimator which results in reduced musical noise due to the low variance LMTS estimates. The proposed method is compared with that of the most recent related work [28,29] and found to perform better.

The paper is organised as follows. In Section 2 the basics of acoustic echo cancellation and multitaper spectrum estimator are discussed. In Section 3 the integrated system for echo and noise cancellation using cross-band adaptive filters and multitaper spectrum estimator is developed and described. Section 4 presents the performance evaluation of the proposed integrated system and Section 5 concludes the paper.

2. Background

2.1. Acoustic Echo Cancellation

Speech signal from the loudspeaker of mobile phone and teleconferencing system may get reflected from the wall and other objects present in the room before being picked up by the microphone of the system. This echo signal will be heard by the far-end speaker and is quiet annoying. Echo signal is treated as a filtered signal and it depends on the impulse response of the room. The basic principle in an Acoustic Echo Cancellation (AEC) system is to model the echo path and estimate the echo so that it can be subtracted from the microphone signal. To identify the echo path, adaptive filters are widely used. Adaptive filter is used in a particular configuration as shown in Fig. 1 to find the room impulse response.

Let $x(n)$ represents the speech signal from the far-end where n indicates the time index, $s(n)$ represents the speech signal in the near-end region, $d(n)$ represents the echo signal of the far-end speech reaching the microphone and $\eta(n)$ represents the background noise in the near-end region. The original echo path has a room impulse response of $g(n)$. The echo signal $d(n)$ is obtained as

$$d(n) = x(n) * g(n) \tag{1}$$

where * represents the convolution operation. The microphone signal $y(n)$ includes the near-end speech signal, background noise and echo which is given by

$$y(n) = s(n) + \eta(n) + d(n) \tag{2}$$