Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 70 (2015) 1–12

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speech enhancement by noise driven adaptation of perceptual scales and thresholds of continuous wavelet transform coefficients

Preety D. Swami [a,*], Rupali Sharma [b], Alok Jain [a], Dhirendra K. Swami [c]

[a] *Department of Electronics and Instrumentation Engineering, Samrat Ashok Technological Institute, Vidisha, MP, India*
[b] *Department of Electronics and Communication Engineering, Samrat Ashok Technological Institute, Vidisha, MP, India*
[c] *Department of Computer Science Engineering, VNS Group of Institutions, Bhopal, MP, India*

Received 7 May 2014; received in revised form 22 February 2015; accepted 23 February 2015
Available online 2 March 2015

Abstract

This paper focuses on employing adaptive scales for computation of perceptually scaled continuous wavelet transform coefficients (CWT) and adaptive thresholding of these coefficients for speech enhancement. The adaptive scales and thresholds both were decided on the basis of the noise level of the noisy speech signal. The CWT coefficients were scaled perceptually and the proposed algorithm suggests selection of number of scales required for analysis on the basis of noise level. The CWT coefficients were then thresholded and for this a novel method of generating adaptive thresholds that too depends on the noise level of the noisy signal has also been proposed. Speech signals were acquired from the TIMIT database and evaluation of the proposed method is done by corrupting these signals by white Gaussian noise (at −10, −5, 0, 5, 10, 15 and 20 dB SNRs) and four real world noises (each at 0 dB SNR); pink, babble, car interior and F16 cockpit noise from the NOISEX-92 database. Enhancement results are compared on the basis of signal to noise ratio (SNR), segmental SNR (SSNR), spectral distortion (SD) and perceptual evaluation of speech quality (PESQ).

Results of the proposed method are evaluated against Ephraim Malah filtering, Stein's unbiased risk estimate (SURE) thresholding of bionic wavelet transform (BWT) coefficients (BWT-SURE), Wiener filtering (WF), perceptually scaled wavelet packet transform (PWT), multi-model WF and multi-model sparse code shrinkage (MultiSCS) enhancement methods. For the white Gaussian noise case, at all noise levels, SNR and SSNR of the proposed method were better than all the methods under comparison. SD and PESQ results were lower than multiSCS method at 10 dB SNR but better at 15 dB and 20 dB SNRs. For the babble noise case, the obtained results were lower than Ephraim Malah but better than BWT-SURE. SNR and SSNR results for the cockpit noise were comparable with Ephraim Malah and BWT-SURE while for the pink noise case, the proposed method gives the best results.

Keywords: Speech enhancement; Continuous wavelet transform; Adaptive thresholding; Adaptive scaling; Bionic wavelet transform

1. Introduction

Enhancement of speech focuses on improving the quality of degraded speech by reversing the degradation using signal processing techniques. The quality of speech is judged by two aspects: perceptibility and intelligibility (Kondo, 2012). Perceptual quality solely depends on the listener in terms of its 'goodness'. Speech intelligibility, on the other hand, is a measure of accuracy that depends on the correct identification of syllables, words or sentences employed for testing. Thus, apart from perceptual quality, many efforts are laid in predicting intelligibility of the speech obtained after processing by noise reduction algorithms (Healy et al., 2013; Baykaner et al., 2013). Common degradations are background noise, reverberations and speech from competing speakers. Applications of speech enhancement include hands free communication,

[*] Corresponding author. Tel.: +91 9406522431.
E-mail address: preetydswami@yahoo.com (P.D. Swami).

http://dx.doi.org/10.1016/j.specom.2015.02.007
0167-6393/© 2015 Elsevier B.V. All rights reserved.

Voice over Internet Protocol (VoIP), hearing aids, local and long distance communications, to name a few (Benesty et al., 2005).

One way of classifying speech enhancement methods is to categorize them into spectral processing and temporal processing methods (Krishnamoorthy and Prasanna, 2009). Spectral processing methods rely on the basis that human spectral perception is not sensitive to short time phase (Loizou, 2007). Thus in these methods only the spectral magnitude associated with the original signal is estimated. Spectral processing methods can be broadly classified into two categories: nonparametric methods and statistical model based methods. Nonparametric methods include subtractive methods (Boll, 1979; Kamath and Loizou, 2002; Lockwood and Boudy, 1992; You et al., 2007) and wavelet denoising (Fu and Wan, 2003; Johnson et al., 2007; Tasmaz and Erçelebi, 2008) that work by first computing an average of the degradation and then removing it. Statistical model based methods (Chang et al., 2007; Chen and Loizou, 2007; Gerkmann and Krawczyk, 2013; Jancovic et al., 2012; Soon et al., 1998) work by assuming certain models for the distribution of the spectral component of either speech or noise. A possible assumption is that the Fourier expansion coefficients of speech and noise can be modeled as independent, zero mean Gaussian variables. Minimum-mean square error short-time spectral amplitude (MMSE-STSA) estimator (Ephraim and Malah, 1984) is an example of statistical model based method that minimizes the mean square error between the short time spectral magnitudes of the clean and enhanced speech signal. In temporal processing methods the characteristics of degradation is not required. They work by accentuating the high SNR regions of the noisy signal. Speech enhancement is then done by exploiting the characteristics of excitation source using linear prediction (Yegnanarayana et al., 2002).

In this work, a method for enhancing speech corrupted with background noise is proposed that is based on adapting the scales and thresholds of CWT coefficients according to the input noise level. The method first decides the number of scales at which the CWT of the noisy signal is to be computed on the basis of the standard deviation of the noise in the noisy signal. The CWT coefficients obtained were then soft thresholded using the proposed adaptive thresholds that were again decided by the standard deviation of the noise in the noisy signal. Experiments are done on speech signals from the TIMIT database corrupted by additive white Gaussian noise (AWGN), pink noise, babble noise, F16 cockpit noise and car interior noise. The performance of the proposed method is compared with some well established methods on the basis of SNR, SSNR, PESQ and SD.

In previous work done by authors in (Sharma and Swami, 2014), it was shown that adaptive thresholding along with a different but fixed set of scales for positive and negative input SNR values when combined with bionic wavelet transform results in significant improvement in enhancement results. The present work is an extension of the previous work in which the number of perceptual scales of the continuous wavelet transform for the noisy signals with positive SNRs were also made adaptive as a function of input noise. Also the algorithm is tested on various real world noises and is compared with some recently published methods. In addition to SNR and SSNR, SD and PESQ are used for these comparisons.

The paper is organized as follows. Section 2, provides a brief description of some speech enhancement methods that are used for comparison with the proposed work. In Section 3, the proposed adaptive scales and thresholds method is elaborated. Comparative results between several related enhancement procedures are demonstrated in Section 4. Section 5 concludes the paper.

2. Background

2.1. Perceptually scaled wavelet packet transform (PWT)

PWT is a form of wavelet packet decomposition that is tailored to match the Bark scale used in speech processing applications. PWT denoising is a combination of bark-scaled wavelet packet decomposition (BS-WPD), a soft-decision gain modification and a "magnitude" decision-directed estimation technique (Cohen, 2001). It attains speech enhancement by introducing some redundancy in the wavelet packet decomposition. The PWT is an overcomplete representation meant specifically for audio signals that achieve higher frequency resolution than the critical band decomposition and a higher time resolution than the conventional wavelet packet decomposition (WPD). Expanding all the high frequency subbands as in critical band wavelet packet decomposition (CB-WPD) results in an increase in computational complexity and at the same time reduces the perceptual quality of unvoiced sounds. Initially, the BS-WPD and CB-WPD splits the audio frequency range 0–8KHz into 21 subbands. Redundancy for speech enhancement is provided by further decomposing the CB-WPD into four nondecimated subbands by a two level overcomplete expansion as shown in Fig. 1. Thus 84 subbands are obtained as outputs of lowpass and highpass wavelet filters without downsampling. This overcomplete auditory representation results in PWT denoising when combined with modified Wiener filtering and the "magnitude" decision directed estimation.

2.2. Wiener filtering (WF)

Wiener filter obtains an estimate of the clean signal from the corrupted signal by minimizing the mean square error between the clean signal and the estimated signal. The Weiner filter H when multiplied with noisy signal gives the estimate of the clean signal. Both signal and noise are assumed to follow a Gaussian distribution and the filter transfer function, H, in the frequency domain is thus a solution to the optimization problem given by