Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 1123–1134
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Multiple statistical models for soft decision in noisy speech enhancement☆

Joon-Hyuk Chang[a,*], Saeed Gazor[b], Nam Soo Kim[c], Sanjit K. Mitra[d]

[a]*School of Electronic Engineering, Inha University, Incheon 402-751, Korea*
[b]*Department of Electrical and Computer Engineering, Queen's University, Kingston, Ont., Canada K7L 3N6*
[c]*School of Electrical Engineering and INMC, Seoul National University, Seoul, Kwanak 151-742, P.O. Box 34, Korea*
[d]*Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106, USA*

Received 29 November 2005; received in revised form 9 June 2006; accepted 4 July 2006

Abstract

Most speech enhancement algorithms are based on the assumption that speech and noise are both Gaussian in the discrete cosine transform (DCT) domain. For further enhancement of noisy speech in the DCT domain, we consider multiple statistical distributions (i.e., Gaussian, Laplacian and Gamma) as a set of candidates to model the noise and speech. We first use the goodness-of-fit (GOF) test in order to measure how far the assumed model deviate from the actual distribution for each DCT component of noisy speech. Our evaluations illustrate that the best candidate is assigned to each frequency bin depending on the Signal-to-Noise-Ratio (SNR) and the Power Spectral Flatness Measure (PSFM). In particular, since the PSFM exhibits a strong relation with the best statistical fit we employ a simple recursive estimation of the PSFM in the model selection. The proposed speech enhancement algorithm employs a soft estimate of the speech absence probability (SAP) separately for each frequency bin according to the selected distribution. Both objective and subjective tests are performed for the evaluation of the proposed algorithms on a large speech database, for various SNR values and types of background noise. Our evaluations show that the proposed soft decision scheme based on multiple statistical modeling or the PSFM provides further speech quality enhancement compared with recent methods through a number of subjective and objective tests.

Keywords: Speech enhancement; DCT; Multiple statistical model; Gaussian; Laplacian; Gamma; GOF; PSFM; SAP; PESQ

1. Introduction

In recent years there has been a great deal of interest in the enhancement of noisy speech for speech coding, recognition, and many other applications. In practice, the presence of noise seriously degrades the performance of speech coder and recognition systems. A number of approaches have been advanced for speech enhancement, e.g., spectral subtraction, Wiener filtering, soft decision estimation, and minimum mean square error (MMSE) estimation [1–3]. Widespread use of these methods is due to the fact that they are fairly straightforward to implement, effective in removing various background noises, and have a low computational load. It has been reported that speech enhancement algorithms based on soft decision gain modification have better performances compared with earlier methods employing hard decisions in which each frame is classified into either speech or non-speech using a voice activity detector (VAD) [4–11]. The most popular algorithms for the enhancement of noisy speech are based on the discrete Fourier transform (DFT) [12]. The transformation decorrelates the speech samples and concentrates the energy of different components more efficiently in separate bins within the transform domain,

Abbreviations: DCT, discrete cosine transform; PSFM, power spectral flatness measure; SAP, speech absence probability; MMSE, minimum mean square error; CLT, central limit theorem; CDF, cumulative distribution function; MS, model selection; GOF, goodness-of-fit; SNR, signal-to-noise ratio; DFT, discrete Fourier transform; VAD, voice activity detector; KLT, karhunen–Loeve transform; KS, Kolmogorov–Smirnov

☆ This work was supported in part by IITA through IT Leading R&D Support Project.

* Corresponding author. Tel.: +82 32 860 7423.
E-mail address: changjh@inha.ac.kr (J.-H. Chang).

0031-3203/$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2006.07.006

i.e., the transformation approximates the Karhunen–Loeve transform (KLT). Almost all of the known speech enhancement algorithms that operate in the DFT domain assume that the coefficients of both the noisy speech and noise are all jointly zero-mean Gaussian distributed random variables in the transform domain [4–7,9,11]. This Gaussian assumption is motivated by the central limit theorem (CLT) as these coefficients are just a weighted sum of a large number of the speech samples. However, through a number of experiments in many transformed domains, it has been found that the clean speech and noise coefficients are more effectively described by other probability density functions (pdf's) such as the Gamma and Laplacian distributions [13,14]. In particular, we have observed that the DFT spectra of noisy speech follows the Laplacian pdf better than the conventional Gaussian pdf [6,13]. According to Soon et al., the DCT has been found to be better in enhancing noisy speech as compared to the DFT [15]. The main reason is that the DCT provides significantly higher speech energy compaction compared to the DFT [5,15]. The amplitude estimator for the DCT is obtained based on the assumption that both the noise and original speech signal amplitudes can be modeled by zero-mean Gaussian distributed random variables in the transform domain. However, it has been pointed out in our previous work [5] that the Laplacian model is more suitable than the conventional Gaussian model for DCT spectra of noisy speech. More recently, Gazor et al. [13] investigated the distribution of clean speech signals under the KLT and DCT and reported that the statistics of DCT coefficients of the clean speech signal (excluding samples of silence intervals) are like those of a Laplacian pdf. This is analogous to the pdf of the DCT coefficients of image signals that is best approximated by the Laplacian distribution [16].

In this paper, we employ multiple distributions (three models) where the distribution of the DCT coefficients of the signals is represented by either the Gaussian pdf, the Laplacian pdf or the Gamma pdf. Specifically, these three pdfs are employed jointly, to represent the distribution of each DCT coefficient. Each model has been independently suggested for speech enhancement [2–6,8–11] and results in better performance compared to other methods under certain conditions. To further improve the performance, our approach selects the best model first depending on the environment using various estimated parameters such as frequency, the SNR, and the PSFM.

In order to establish a set of reliable rules for selection from these models, we use the goodness-of-fit (GOF) test as a distance measure between the empirical cumulative distribution function (CDF) and each of these models. We first evaluate these models under various noise environments to measure the accuracy of each model for each SNR and noise type. Actually, GOF tests are carried out for each frequency bin separately to compare these models as a function of the frequency bins in the DCT domain [4–7,9–11,17]. Following these statistical models, we use a soft Bayesian-decision scheme for the enhancement of the noisy speech in order to achieve a robust performance and lower the speech clipping phenomena in which the probability of speech absence is incorporated to modify the spectral gain for the noise suppression.

More specifically, we carry out the GOF test and establish relevant relations between the PSFM and the statistical characteristics of the signal, such as the pdf of the DCT coefficients resembles best the gamma pdf while the PSFM is low; in contrast, the best model is the Gaussian as the PSFM increases. We exploit these relations and develop novel methods for the selection of statistical models. According to the experimental results, we derive a robust assignment rule for model selection by assigning a pdf to each DCT coefficient of speech and noise from the set of candidates. Through a number of subjective and objective tests, we compare the performances of these methods with some recent related approaches. We have found that the proposed multiple model approaches (employing PSFM) are superior to or at least comparable to existing approaches at all testing conditions.

The organization of this paper is as follows. Section 2 presents three statistical models and statistically evaluates these models based on the GOF test. An algorithm for the enhancement of the DCT coefficients is proposed in Section 3. This enhancement algorithm uses the soft estimate of the SAP. In Section 4, an adaptive estimation procedure for the SAP is described based on these multiple models. In Section 5, a number of subjective and objective quality tests are conducted to evaluate the performance of the algorithm, and finally in Section 6, some concluding remarks are drawn.

2. Multiple models and statistical analysis

In this section, we briefly review multiple statistical models in a speech enhancement framework under noisy environments and evaluate each statistical model via the GOF test. For this, we assume that the additive noise is independent of the speech signal. We denote the M-point DCT of M successive samples of the received noisy speech by

$$X_k(t) = S_k(t) + N_k(t), \quad 0 \leqslant k \leqslant M-1, \tag{1}$$

where k denotes the kth frequency bin index, M is the total number of frequency components, and t is the time frame index, respectively. In Eq. (1) $N_k(t)$ and $S_k(t)$ are the DCT coefficients of the noise, and clean speech, respectively.

The basic assumption commonly adopted in most speech enhancement approaches is to describe each frame of noisy speech signal as either a noise frame or as a speech frame, i.e., the following hypotheses:

$$\begin{cases} \text{speech absent} & H_0 : \mathbf{X}(t) = \mathbf{N}(t), \\ \text{speech present} & H_1 : \mathbf{X}(t) = \mathbf{N}(t) + \mathbf{S}(t), \end{cases} \tag{2}$$

where $\mathbf{X}(t) = [X_0(t), X_1(t), \ldots, X_{M-1}(t)]^T$, $\mathbf{N}(t) = [N_0(t), N_1(t), \ldots, N_{M-1}(t)]^T$, and $\mathbf{S}(t) = [S_0(t), S_1(t), \ldots,$