Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 36 (2016) 58–71

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Speech enhancement using Maximum A-Posteriori and Gaussian Mixture Models for speech and noise Periodogram estimation☆

Sarang Chehrehsa*, Tom James Moir

School of Engineering, Auckland University of Technology (AUT), St. Pauls Street, Auckland, New Zealand

Received 27 February 2015; received in revised form 16 August 2015; accepted 2 September 2015
Available online 11 September 2015

Abstract

In speech enhancement, Gaussian Mixture Models (GMMs) can be used to model the Probability Density Function (PDF) of the Periodograms of speech and different noise types. These GMMs are created by applying the Estimate Maximization (EM) algorithm on large datasets of speech and different noise type Periodograms and hence classify them into a small number of clusters whose centroid Periodograms are the mean vectors of the GMMs. These GMMs are used to realize the Maximum A-Posteriori (MAP) estimation of the speech and noise Periodograms present in a noisy speech observation. To realize the MAP estimation, use of a constrained optimization algorithm is proposed in which relatively good enhancement results with high processing times are attained. Due to the use of constraints in the optimization algorithm, incorrect estimation results may arise due to possible local maxima. A simple analytic MAP algorithm is proposed to attain global maximums in lower calculation times. With the new method the complicated MAP formula is simplified as much as possible to find the maxima, through solving a set of equations and not through conventional numerical methods used in optimization. This method results in excellent speech enhancement with a relatively short processing time.
© 2015 Elsevier Ltd. All rights reserved.

Keywords: Gaussian Mixture Model (GMM); Maximum A-Posteriori (MAP); Wiener filter; Speech enhancement

1. Introduction

Speech is the simplest, most efficient and most frequent way of communication among human beings. There are lots of applications that transmit, amplify and understand these signals such as mobile communication, hearing aids and speech recognition systems. Due to the modern requirement of portability of these applications, they must perform in differing environments with different background noise types and levels. The background noise will highly degrade the performance of these applications, specially the quality and intelligibility of the output speech signal and hence speech enhancement algorithms can play a critical role in these applications.

As discussed in Mohammadiha et al. (2013a), speech enhancement algorithms can be classified into two main categories: unsupervised and supervised algorithms. The simplest and most famous speech enhancement method is

☆ This paper has been recommended for acceptance by Tom Quatieri.

* Corresponding author. Tel.: +64 211874450.
E-mail addresses: sarang.chehrehsa@aut.ac.nz (S. Chehrehsa), tom.moir@aut.ac.nz (T.J. Moir).

http://dx.doi.org/10.1016/j.csl.2015.09.001
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

spectral subtraction (Boll, 1979) in which the spectral amplitude of noise is estimated from the spectral amplitude of noisy speech and subtracted from it to get to the spectral amplitude of clean speech. Since there are no considerations of the speech spectrum in spectral subtraction, it results in an artificial noise called musical noise. To reduce this musical noise some methods discussed in Sim et al. (1998) can be used which are simple but require an efficient Voice Activity Detector (VAD) to estimate the noise spectrum. In these methods the amplitude of the spectrum is used, but in Lu and Loizou (2008) the complex spectrum is considered. In Wiener filtering (Widrow and Mccool, 1975), speech and noise probabilistic properties are used and hence the enhanced speech suffers from less musical noise with respect to spectral subtraction methods. A method called Short Time Spectral Amplitude (STSA) is discussed in Ephraim and Malah (1984), which is based on Minimum Mean Square Error (MMSE) estimation with the assumption that the speech and noise spectral components are statistically independent and Gaussian random variables. The MMSE-STSA method is derived by minimizing a conditional mean square value of the short time spectral amplitude. As discussed in Wolfe and Godsil (2003), when PDFs of speech and noise spectrums are assumed to be Gaussian, the spectral gain will become the Wiener filter gain. This method is based on the a priori SNR estimation on a frame-by-frame basis by a decision directed approach and Maximum Likelihood (ML) estimator with the assumption that the noise variance is known or can be estimated during the silence intervals. There are different methods to calculate the a priori SNR. A decision directed method is discussed in Cappe (1994). A data driven approach to calculate a priori SNR is discussed in Suhadi et al. (2011) in which two trained artificial neural networks, one for speech and one for noise, is used. As confirmed in the literature, the MMSE spectral gain is superior to the spectral subtraction method but is computationally complicated to implement. To overcome this issue, a method discussed in Wolfe and Godsil (2003) called the Maximum A-Posteriori (MAP) method, which can result in relatively good enhancement results is used.

In supervised speech enhancement algorithms we use some additional information about noise and speech such as noise type, speaker identity etc. to improve the enhancement. In supervised methods we create some offline models for speech and noise which are trained using large observed samples of each signal. Some examples of this class of algorithms include the codebook based approaches as discussed in Sreenivas and Kirnapure (1996), where LPC codebooks of speech are used and in Srinivasan et al. (2006), AR coefficient codebooks of speech and noise are used which leads to ML estimates of clean speech. Another well-known and high performance supervised speech enhancement methods are Hidden Markov Model (HMM) based systems and the state-of-the-art approaches are discussed in references (Ephraim, 1992; Sameti et al., 1998; Zhao and Kleijn, 2007). In these methods, the waveform signal is modelled as an autoregressive (AR) process, and hence the waveforms of speech and noise signals are modelled by HMMs. In recent HMM based methods as discussed in Mohammadiha et al. (2013b), Mohammadiha and Leijon (2013) and Veisi and Sameti (2013), distribution of the power spectral coefficients of speech and noise are modelled using HMMs using Gamma distribution. There are also other methods that are based on modelling the spectral amplitude of speech using Gaussian Mixture Modelling (GMM) in which estimates of speech are attained using a MAP criterion as discussed in Hao et al. (2009, 2010) and Fodor and Fingscheidt (2011) and a minimum mean-square error (MMSE) criterion as discussed in Burshteinand and Gannot (2002). The advantage of the supervised approaches such as the HMM based denoising algorithm is that it produces high quality enhanced speech signals. The reason for this is that for each noise type, a system is trained a priori. This is a tedious task in practice and is addressed in Mohammadiha et al. (2013a).

In most supervised model based algorithms, GMM is used for the modelling of spectral amplitudes, log spectral amplitudes or Periodograms with their true power and some of these methods give excellent enhancement results. In this research we are going to use normalized Periodograms (with power equal to one) as the vectors to be modelled. In most Bayesian estimation criterions especially MAP estimation, due to the complexity of the estimation criterion formula, some mathematical distributions like Laplacian or Gamma are used to simplify the formula. Here we aim to use GMM to realize an explicit MAP estimate of speech signals. These are different methods from the previous work discussed in Wolfe and Godsil (2003) and Lotter and Vary (2005) (which use approximated distributions to simplify the MAP formula) and hence our aim is to realize the MAP with no approximations.

In our previous research we used model-based speech enhancement methods such as codebook constrained Wiener filtering and Bayesian estimation. In codebook constrained Wiener filtering, we used the K-means algorithm to classify the large datasets of speech and noise normalized Periodograms into some centroids. In this way the Periodograms are classified based on their variety and shape with normal power (Chehresa and Savoji, 2009, 2010). Despite the large sizes of codebooks, poor enhancement results were attained. Later we used Gaussian Mixture Modelling (GMM) in which we modelled the Probability Density Function (PDF) of speech and noise Periodograms (Chehresa and Savoji,

Download English Version:

https://daneshyari.com/en/article/558220

Download Persian Version:

https://daneshyari.com/article/558220

Daneshyari.com