Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 47 (2005) 265–276
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Multi-frame GMM-based block quantisation of line spectral frequencies

Stephen So, Kuldip K. Paliwal *

School of Microelectronic Engineering, Griffith University, Nathan Campus, Brisbane, QLD 4111, Australia

Received 30 April 2004; received in revised form 21 January 2005; accepted 15 February 2005

Abstract

In this paper, we investigate the use of the Gaussian mixture model-based block quantiser for coding line spectral frequencies that uses multiple frames and mean squared error as the quantiser selection criterion. As a viable alternative to vector quantisers, the GMM-based block quantiser encompasses both low computational and memory requirements as well as bitrate scalability. Jointly quantising multiple frames allows the exploitation of correlation across successive frames which leads to more efficient block quantisation. The efficiency gained from joint quantisation permits the use of the mean squared error distortion criterion for cluster quantiser selection, rather than the computationally expensive spectral distortion. The distortion performance gains come at the cost of an increase in computational complexity and memory. Experiments on narrowband speech from the TIMIT database demonstrate that the multi-frame GMM-based block quantiser can achieve a spectral distortion of 1 dB at 22 bits/frame, or 21 bits/frame with some added complexity.

Keywords: Speech coding; LSF coding; Transform coding; Block quantisation; Gaussian mixture models

1. Introduction

Linear predictive coding (LPC) of speech requires the accurate quantisation of parameters representing the spectral envelope. Speech is windowed into frames and the spectral envelope is parametrically modelled as an all-pole filter, whose coefficients are called linear predictive coding (LPC) parameters. These LPC parameters are generally quantised in terms of line spectral frequencies (LSFs) using a vector quantiser (VQ). Extrapolating from the operating curve of full search VQ suggests that we need about 19 bits/frame to achieve transparent coding of these parameters (Paliwal and Kleijn, 1995), while high

* Corresponding author. Tel.: +61 7 3875 3754/6536; fax: +61 7 3875 5198.
E-mail addresses: s.so@griffith.edu.au (S. So), k.paliwal@griffith.edu.au (K.K. Paliwal).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.007

rate analysis predicts a lower bound of 23 bits/frame[1] (Hedelin and Skoglund, 2000). It is not possible to design codebooks at these rates and in addition, the computational cost of the resulting full search vector quantiser is very high.

Less complex but suboptimal vector quantisers such as multistage and split VQ have been investigated in the speech coding literature (LeBlanc et al., 1993; Paliwal and Atal, 1993), where it was generally observed that 22–24 bits/frame were required to achieve *transparent coding*[2] in speech, with varying degrees of complexity. Further gains in performance can be achieved by exploiting temporal correlation between successive frames. Matrix quantisation (Tsao and Gray, 1985) and its derivatives such as split matrix quantisation (Xydeas and Papanastasiou, 1999) and multimode matrix quantisation (Nurminen et al., 2003; Sinervo et al., 2003) perform better by jointly quantising LSF frames.

The use of Gaussian mixture models (GMM) for the coding of LSFs has been investigated in (Hedelin and Skoglund, 2000; Shabestary and Hedelin, 2002; Subramaniam and Rao, 2000, 2001, 2003). In (Subramaniam and Rao, 2003), a Gaussian mixture model (GMM) is used to parameterise the probability density function (PDF) of the source and optimised Gaussian block quantisers are designed for each cluster (or, mixture component).[3] Using this quantiser in its fixed rate mode, a spectral distortion of approximately 1 dB was achieved at 24 bits/frame. The main advantages of this scheme over vector quantisers include (Subramaniam and Rao, 2003):

1. lower complexity through the use of block quantisers;
2. bitrate scalability; and
3. search complexity and memory requirements being independent of the rate of the system.

A modified quantiser with memory was also described in (Subramaniam and Rao, 2003) that coded the difference between successive frames, similar to differential pulse code modulation (DPCM) with a one-tap predictor. A spectral distortion of 1 dB was achieved at 22 bits/frame (Subramaniam and Rao, 2003). During the coding process, there is frequent use of the spectral distortion (SD) calculation for cluster quantiser selection. While there are approximate high-rate expressions for the spectral distortion calculation (Gardner and Rao, 1995), the number of computations is still comparatively higher than mean squared error (MSE).

In this paper, we investigate a modified version of the fixed-rate GMM-based block quantiser that operates on multiple frames and uses the mean squared error (MSE) distortion criterion.[4] We have found this system to perform better than the single frame as well as predictive quantiser of (Subramaniam and Rao, 2003) in terms of spectral distortion.

The organisation of this paper is as follows. Section 2 introduces some preliminaries such as the line spectral frequency representation of LPC parameters and distortion measures that are commonly used in speech coding. In Section 3, we describe the operation of the multi-frame GMM-based block quantiser as well as its computational and memory requirements. Section 4 details the LPC analysis method and speech database that we have used to evaluate the performance of the quantiser. Following this is a discussion of the performance of the multi-frame GMM-based block quantiser and how it compares with other quantisation schemes. Finally we conclude in Section 6.

[1] This is the lower bound for full-band spectral distortion (0–4 kHz) while for partial-band (0–3 kHz), the bound is 22 bits/frame (Hedelin and Skoglund, 2000).

[2] Transparent coding means that the coded speech is indistinguishable from the original through listening. An objective measure of quality is the *spectral distortion* (SD), which is defined as the root-mean-square difference between the log-spectra of the coded and original speech. Coded speech is generally accepted as being transparent when the average SD is about 1 dB (Paliwal and Atal, 1993).

[3] Therefore, we refer to this quantisation scheme as a GMM-based block quantiser.

[4] This paper is an extended version of (Paliwal and So, 2004) and contains more comparative results.