Chapter 7

MFCC Quantisation in Distributed Speech Recognition

7.1 Abstract

This chapter investigates the application of the multi-frame GMM-based block quantisation scheme to MFCC quantisation in distributed speech recognition and examines how it compares with other schemes. The advantage of the multi-frame GMM-based block quantiser is: superior recognition performance at low bitrates, which is comparable with vector quantisation; fixed and relatively low computational and memory complexity that is independent of bitrate; and bitrate scalability, where the bitrate can be dynamically altered without requiring codebook re-training.

We begin the chapter with some background theory on speech recognition, which covers the basic ideas of feature extraction and pattern recognition using hidden Markov models (HMMs). Following this, we provide a general review of client/server-based speech recognition systems and the various types of modes (NSR and DSR) that have been proposed and reported in the literature. We also briefly describe the Aurora-2 DSR experimental framework, which will be used extensively to evaluate the performance and robustness to noise of the various DSR schemes. The second half of the chapter is dedicated to presenting and discussing results of different quantisation schemes applied to a common DSR framework. The schemes investigated include the multi-frame GMM-based block quantiser, the memoryless GMM-based block quantiser, the non-uniform (Lloyd-Max) scalar

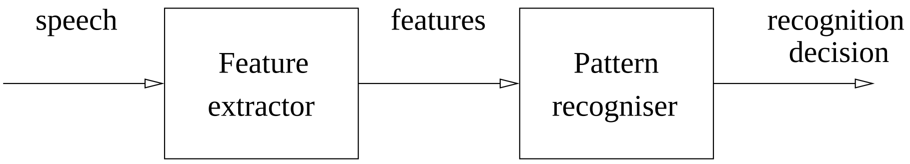

Figure 7.1: A typical speech recognition system (after [178])

quantiser, and vector quantiser. Two sets of experimental results are presented. The first set compares the recognition performance of each quantisation scheme as a function of bitrate in clean and matched conditions. The second set compares the recognition performance of each scheme as a function of SNR in noisy, mismatched conditions.

Publications resulting from this research: [131, 174]

7.2 Preliminaries of Speech Recognition

Figure 7.1 shows a block diagram of a speech recognition system, highlighting the main components in general. In this section, we give only a brief review of each of these components rather than a comprehensive coverage of the algorithms used in modern recognition systems, as the scope of this chapter is focused on the efficient quantisation of MFCC features for distributed speech recognition.

7.2.1 Speech Production

Speech sounds can be broadly classified as either *voiced* or *unvoiced.* Voiced sounds, such as /*iy*/ (as in s*ee*), are periodic and have a harmonic structure that is not present in unvoiced sounds, such as /*s*/, which are aperiodic and noise-like. These are best visualised in Figure 7.2, which shows the waveform and spectrogram of the sentence, *she had your dark suit in greasy wash-water all year*, and highlights the voiced and unvoiced sections in the first word, *she.* Notice that the spectrum for /*sh*/ is flat, similar to that of noise, while the spectrum of /*iy*/ shows a harmonic structure, as characterised by the alternating bands.

Voiced sounds are produced when air from the lungs is excited by vibrating vocal folds in the larynx. A glottal wave, with a fundamental frequency of f_0 and harmonics

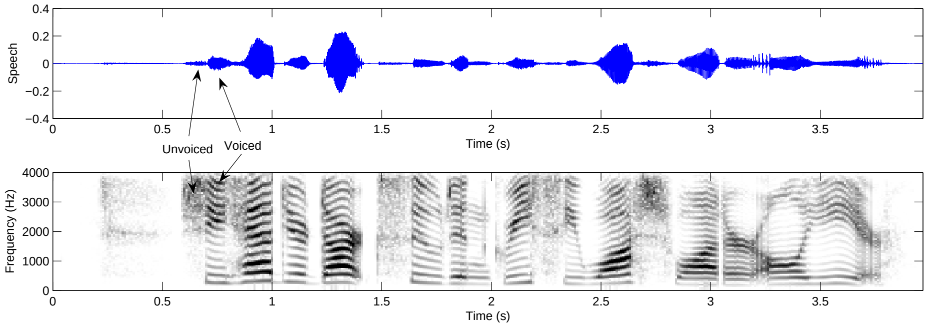

Figure 7.2: Waveform and spectrogram of the sentence, *she had your dark suit in greasy wash-water all year*, highlighting the unvoiced /*s*/ and voiced /*iy*/ sounds in *she*.

at multiples of the fundamental frequency, is generated and this wave passes through the vocal tract, which can be viewed as an acoustic tube that starts at the larynx and terminates at the lips. This tube changes shape to create resonances and anti-resonances that emphasise and de-emphasise certain parts of the spectrum, respectively. *Formants* occur where the spectrum has been emphasised by the resonances of the vocal tract. Along with changes in the articulators (the lips, tongue, jaws, and teeth), different quasi-periodic sounds can be produced [73, 151]. The vocal folds do not vibrate for unvoiced sounds but instead, the vocal tract is constricted by the articulators and air passes through rapidly to produce a noise-like sound [151].

Speech production can be modelled as consisting of a *source* and *filter* component. The source component represents the excitation (which is aperiodic noise for unvoiced sounds and periodic for voiced sounds) while the filter component emphasises parts of the spectrum, just like the vocal tract with its resonances and anti-resonances [151].

7.2.2 Feature Extraction

The role of feature extraction is to compactly represent speech in a way that preserves information that is relevant and important for subsequent recognition [73]. Speech is initially passed through a pre-emphasis filter:

$$H(z) = 1 - \alpha z^{-1} \tag{7.1}$$

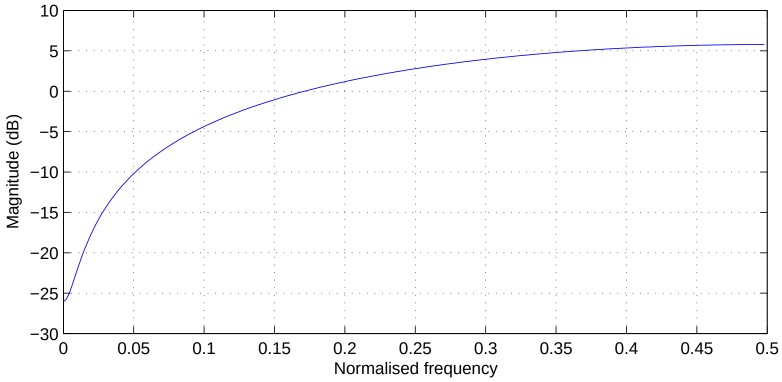

Figure 7.3: Magnitude and phase response of the pre-emphasis filter, $H(z) = 1 - 0.95z^{-1}$

where $\alpha = 0.95$. The magnitude response of this filter is shown in Figure 7.3. The role of the pre-emphasis filter is to remove the spectral tilt (ie. flatten the spectrum) [138], as shown in Figure 7.4(b), where the first formant has been shifted down while the higher frequency formants have been shifted up, allowing them to be analysed at the same level.

Assuming that speech is sampled at 8 kHz, the pre-emphasised speech is then windowed into overlapping segments of 25 ms with 10 ms shift before analysis. A tapered window function, such as the Hamming window, is used to reduce the effects of spectral leakage caused by the blocking process. Acoustic information in speech (eg. formants) manifests itself in the frequency domain[1] (as shown in Figure 7.4), hence the most popular parametric feature sets for speech recognition are spectral-based and can be categorised into two classes: *linear prediction-based* and *Fourier transform-based* [35].

Linear Prediction-Based Features

In linear prediction-based feature extraction, feature vectors are derived from the LPC spectra of speech, which models the vocal tract. Examples include the linear prediction coefficients themselves, reflection coefficients [35], line spectral frequencies [62], etc. In pre-HMM speech recognition systems, the Itakura minimum prediction residual distance measure [76], which calculates the residual error when the tested speech frame is filtered through the reference LPC filter, is used as a distance measure for features based on the LPC spectra. Alternatively, features can be derived from the LPC cepstra and these

[1] The temporal movement of formants, as shown in Figure 7.2, is also useful for speech recognition and is expressed in the delta and acceleration features, which are the first and second derivatives of the spectral-based features [73].

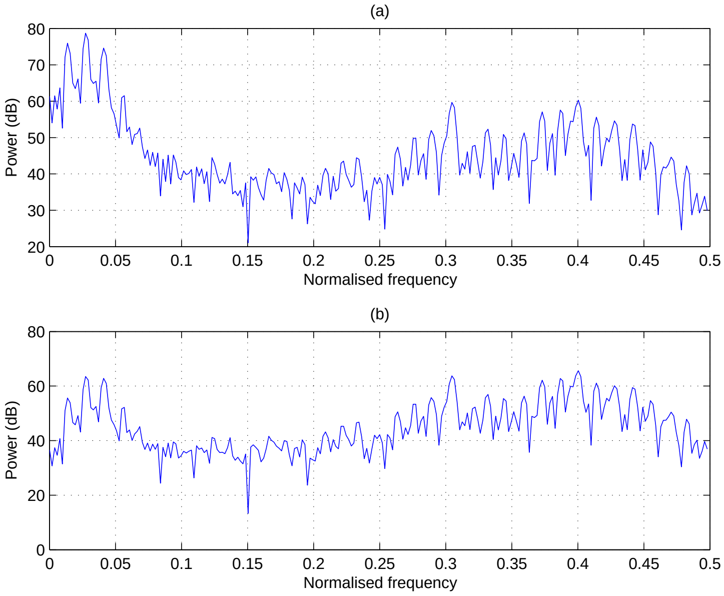

Figure 7.4: The effect of pre-emphasis on the power spectral density of speech: (a) Original PSD of speech; (b) PSD of speech filtered with pre-emphasis filter.

include the linear prediction-based cepstral coefficients (LPCCs) and perceptual linear prediction (PLP) coefficients [67]. The advantage of cepstral-based features is that, because they are derived from an orthogonal basis, simpler Euclidean distance measures can be used [35]. Through a recursive equation, LPCCs are calculated directly from the linear prediction coefficients which represent the speech spectrum in terms of linear frequency. PLP coefficients are calculated in a similar way, where the only difference is that the autocorrelation coefficients used for the linear prediction analysis are derived from the inverse Fourier transform of the Bark frequency-warped power spectral density, obtained using the short-time Fourier transform [73]. The advantage of PLPs over LPCCs is the non-linear frequency scale which matches more closely to the perceptual response of the human auditory system.

Mel Frequency-Warped Cepstral Coefficients

The other class of features are Fourier transform-based and include the Mel frequency-warped cepstral coefficients (MFCCs) and Bark frequency-warped cepstral coefficients (BFCCs), though the former is more commonly used. The discrete Fourier transform is applied to each windowed segment of speech, where the magnitude spectrum is squared to obtain the short-time power spectral density (PSD) or power spectrum of the speech. The PSD is then filtered by a series of M overlapping triangular-shaped filters that are centred on the Mel scale[2]. The filters are overlapped in such a way that the starting and ending frequencies fall on the centres of the previous and next filter, respectively, in order to simulate critical bandwidths [151]. Figure 7.5 shows 20 Mel frequency-warped triangular-shaped filters.

The energy from each filter is accumulated and compressed non-linearly using the natural logarithm to reduce the dynamic range, resulting in a vector of M coefficients for each frame. It has been noted in previous studies that most of the variation in speech can be compactly represented by the first few eigenvectors, whose directional cosines are similar to a cosine series expansion [135, 35]. Hence PCA can be approximated by the discrete cosine transform (DCT), which is applied to the vector of log energies to give the final MFCC vector. The decorrelation aspect of the DCT is a desirable characteristic because of the use of hidden Markov models (HMM) in the subsequent pattern recognition stage. The mixture of Gaussians used in each state of the HMMs use diagonal covariance matrices because of the difficulty in re-estimating off-diagonal components when only a finite training data set is available [138].

The first cepstral coefficient, c_0, may be replaced with the log energy, $\log E$, of the speech frame and this captures the changes in recording level. Also *cepstral mean subtraction* (CMS) is often applied to reduce the effect of convolutional distortion due to changes in the microphone, its distance from the speaker, and the acoustics of the room. In CMS, the mean is removed from the MFCCs [73]. In order to capture temporal changes in the spectra, the approximated first and second derivatives of the MFCC feature vectors are calculated to give the delta and acceleration coefficients, respectively, which are concatenated with the mean-removed MFCCs to give a final feature vector of dimension $3M$ [73]. Typically, speech recognition systems use $M = 13$ MFCC coefficients. Therefore, after concatenating with the delta and acceleration coefficients, the final feature vector has a dimension of 39.

It has been shown in various studies (for example, in [35]) that MFCC features perform

[2] The Melody scale, or more commonly known as the Mel scale, is a non-linear and perceptually-motivated frequency scale. It was derived through perception experiments by Stevens and Volkman [179], where the test subject was asked to manually adjust a stimulus tone so that it perceptually had half the pitch of a given reference tone [157].

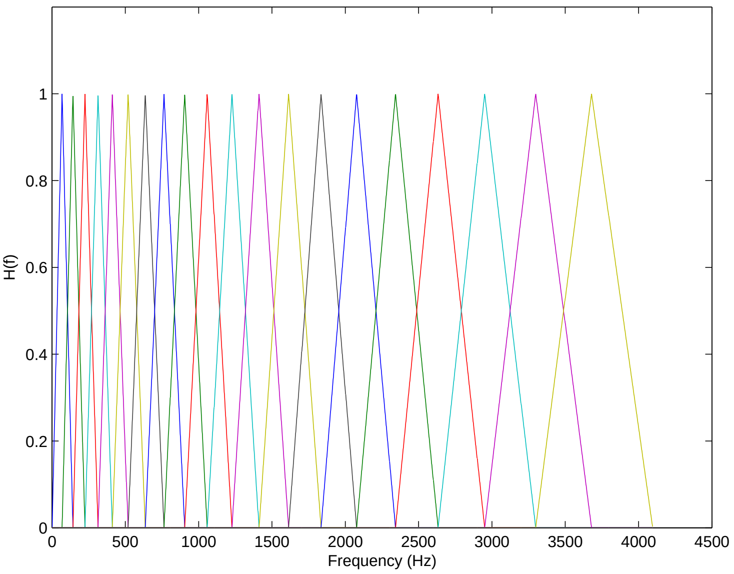

Figure 7.5: Filterbank of Mel frequency-warped triangular-shaped filters used for MFCC calculation from 8 kHz speech ($M = 20$)

better in speech recognition than LPCCs or other linear prediction-based feature sets. Therefore, it is to no surprise that MFCCs are the predominant feature set in modern speech recognition systems.

7.2.3 Pattern Recogniser

The role of the pattern recogniser is to determine which reference template matches the current test feature vector. Of all the pattern recognisers available, the hidden Markov model (HMM) has had the most success in speech recognition. The HMM is a stochastic signal model that consists of a number of probabilistic states [138]. At each instant of time, the model changes from the current state to another state (or, the same state), known as state transition, that is governed by an unobservable or hidden stochastic process. Within each state, there is another stochastic process that produces an output symbol based on a probability distribution, which is usually represented by a Gaussian mixture model.

Typical HMM parameters include the number of states, N, the number of observation vectors per state (for discrete HMMs), M, state transition probability matrix, $\boldsymbol{A}$, output probability matrix, $\boldsymbol{B}$, and initial state distribution, $\boldsymbol{\pi}$ [138]. An HMM is designed from training feature vectors using the Baum-Welch algorithm (also known as the Forward-

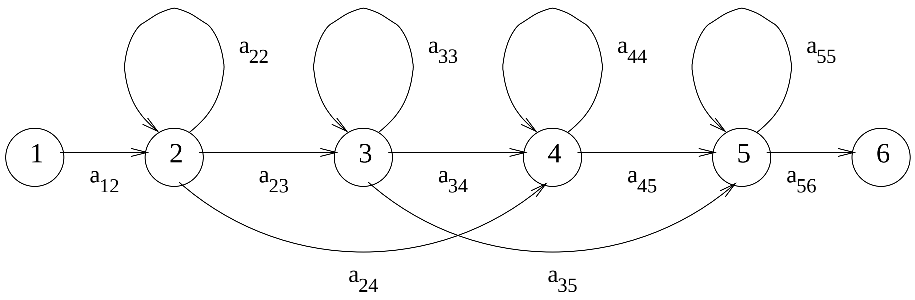

Figure 7.6: A typical 6 state, left-to-right hidden Markov model (HMM) (after [203])

Backward algorithm) which is based on a similar principle to the EM algorithm for GMM estimation [73]. In order to calculate the likelihood, $P(\lambda|\boldsymbol{O})$, of an HMM, λ, generating a given observation sequence, $\boldsymbol{O}$, the Viterbi algorithm is used to select the optimum state sequence that maximises the probability [73]. Therefore, $P(\lambda|\boldsymbol{O})$ can be considered a similarity measure between the given observation sequence and the reference sequence used train the HMM.

In a speech recognition system, an N-state HMM with left-to-right topology, as shown in Figure 7.6, is typically designed for each speech unit (phoneme or word) based on its feature vectors. During the recognition phase, probabilities or likelihoods are calculated for each stored HMM and the one which has the highest likelihood is deemed as matching the given test feature vector. In earlier speech recognition systems, which used discrete HMMs, where each state has a finite set of output symbols, feature vectors are quantised using a vector quantiser before they are matched with all the stored HMM models [138]. In modern speech recognition systems, which use continuous HMMs, the vector quantisation stage is not required [73].

The detailed operation of the HMM is beyond the scope of this work and the reader should refer to Rabiner's HMM tutorial [138] for details.

7.3 Client/Server-Based Speech Recognition

With the increase in popularity of remote and wireless devices such as personal digital assistants (PDAs) and cellular phones, there has been a growing interest in applying automatic speech recognition (ASR) technology in the context of mobile communication systems. Speech recognition can facilitate consumers in performing common tasks, which have traditionally been accomplished via buttons or pointing devices, such as making a call through voice dialing or entering data into their PDAs via spoken commands and sentences. Some of the issues that arise when implementing ASR on mobile devices include: computational and memory constraints of the mobile device; network bandwidth utilisation; and robustness to noisy operating conditions.

Mobile devices generally have limited storage and processing ability which makes implementing a full on-board ASR system impractical. The solution to this problem is to perform the complex speech recognition task on a remote server that is accessible via the network. Various modes of this client/server approach have been proposed and reported in the literature. The most common ones are shown in Figure 7.7 and are discussed in the following subsections.

7.3.1 Network Speech Recognition

In the *Network Speech Recognition* (NSR) mode [85], the user's speech is compressed using conventional speech coders (such as the GSM speech coder) and transmitted to the server which performs the recognition task. In speech-based NSR (Figure 7.7(a)), the server calculates ASR features from the decoded speech to perform the recognition. In bitstream-based NSR (Figure 7.7(b)), the server uses ASR features that are derived from linear predictive coding (LPC) parameters taken directly from the bitstream. Numerous studies have been reported in the literature evaluating and comparing the performance of these two forms of NSR [48, 69, 74, 83, 99, 140, 189, 51].

Literature Review of Speech-Based NSR

Euler and Zinke [48] investigated the effect of three CELP-based speech coders, LD-CELP, RPE-LTP, and TETRA-CELP at 16, 13, and 4.8 kbps, respectively, on isolated word recognition and speaker verification. Narrowband speech was coded and decoded using the CELP coders, and 12 LPCCs and their delta coefficients extracted from the decoded speech. They found that the speech coders operating at 13 kbps and lower decreased the recognition performance in matched and mismatched conditions.

Lilly and Paliwal [99] examined the influence of six speech coders at bitrates ranging from 40 kbps to 4.8 kbps. The tandeming of speech coders and its effect was also investigated. High bitrate ADPCM coders were found to have minimal effect on the recognition while low bitrate CELP coders achieved the lowest recognition accuracy. The same trend was also observed when tandeming. This may be due to the influence of the quantisation noise shaping in CELP coders, which is designed to improve perceptual quality. Spectral information that is important for recognition purposes, is de-emphasised by the noise shaping filter, which exploits spectral masking[3]. MFCCs were also found to be more robust to the effects of speech coders than LPCCs.

Digalakis *et al.* [39] compared the effects of G.721 ADCPM, GSM RTE-LTP, and mu-law on speech recognition performance. Similar to what was observed in [99], the G.721 ADPCM coder incurred minimal degradation while mu-law and GSM achieved the lowest accuracy.

Turunen and Vlaj [189] examined five speech coders and through comparisons and tandeming experiments, identified features that affected the recognition performance. Their first observation was the degrading effect on recognition performance of *postfiltering*, which is applied to smooth the decoded speech and improve its subjective quality. Secondly, the accuracy of the vocal tract model, as represented by the LPC parameters, plays an important role in recognition results. The G.728 LD-CELP does not transmit LPC envelope information but uses a 50th-order backward all-pole predictor. It achieves about 2% less recognition accuracy than the G.729 CS-ACELP coder, which explicitly transmits LPC parameters in the bitstream. Also, other vocal tract models, such as that in the G.727 ADPCM coder which uses a pole-zero model, performed as good as G.729 [189].

Hirsch [69] investigated the influence of the AMR-NB coders at various bitrates and compared the recognition performance with full-rate and half-rate GSM on speech corrupted with noise. Feature extraction was performed using the standard ETSI Aurora frontend and also the advanced noise-robust Aurora frontend. As expected, degradation in recognition performance was observed for coded speech. The performance of the noise-robust Aurora frontend, however, was about 16% higher than when using the standard frontend.

[3] In spectral masking, noise that is in the presence of strong tones will tend to be masked. Therefore, we can afford to increase the amount of noise due to quantisation in the formant regions without affecting the perceptual quality. Noise shaping filters tend to increase the quantisation error in these regions while decreasing, by a similar amount, errors in the spectral valleys [15].

Literature Review of Bitstream-Based NSR

Kim and Cox [83] extracted LPCCs from the spectral envelope information contained in the bitstream. Because speech recognition systems operate on frames sampled at 100 Hz while speech coders process frames at 50 Hz, LSFs from the bitstream were interpolated to the higher frame rate before the LPCCs were extracted. Cepstral liftering was applied and the log energy coefficient calculated from the residual signal. In order to improve the recognition performance further, voiced/unvoiced information from the speech coder was added to the feature.

Huerta and Stern [74] reported their study which examined the derivation of cepstral feature vectors from various parts of the GSM speech coder bitstream and compared the recognition performance with speech-based NSR. One method involved converting the LAR coefficients to LP coefficients and deriving cepstral coefficients (LPCCs). Another involved deriving the cepstrals from the residual signal, as represented by the RPE-LTP parameters. Though the residual signal usually contains only speaker dependent information such as pitch, periodicity, and global waveform information, it still carries some information relevant to speaker independent speech recognition because of the low order (8th) of the LPC analysis [74]. Their results showed the LAR-derived cepstral features to achieve similar performance to speech-based NSR, which had a degraded performance compared with baseline MFCC-based recognition of the original speech. The residual-derived cepstral features did not perform as well though. However, when the LAR-derived and residual-derived cepstral coefficients were concatenated or added to form new features, the recognition accuracy surpassed that of speech-based NSR and was nearly identical to the baseline MFCC-based performance.

Raj *et al.* [140] used a more principled method of combining LPC-derived and residual-derived cepstral coefficients and reported their results for the GSM, CELP and LPC coders. The spectral envelope parameters (LAR coefficients for GSM and LPC or LSFs for CELP), were converted to LP coefficients and LPCCs were derived. The log power spectrum of the residual signal was also calculated and represented as a 32-dimensional vector. The LPCCs and residual log spectra features were concatenated and the extended feature vector reduced in dimensionality using linear discriminant analysis (LDA), whose classes were similar to phoneme classes. The new features achieved the same recognition performance as the baseline system (for GSM and CELP) and were better than speech-based NSR and LPCC-derived features in all cases.

Gallardo-Antolin *et al.* [51] investigated another bitstream-based NSR scheme for GSM speech that was robust to various bit errors such as random errors, burst errors, and frame substitutions. They derived their feature vectors by converting LAR coefficients to LP coefficients and from these, an LPC power spectrum was calculated. MFCCs were then derived from the LPC power spectrum, with the log energy coefficients calculated from analysis of the decoded speech. For error-free speech, the bitstream-derived features performed similarly to those derived from decoded speech. However, as the bit error rate (BER) was increased, the former maintained respectable recognition scores while the latter's performance drops significantly.

7.3.2 Distributed Speech Recognition

In *Distributed Speech Recognition* (DSR), shown in Figure 7.7(c), the ASR system is distributed between the client and server. Here, the feature extraction of speech is performed at the client. These ASR features are compressed and transmitted to the server via a dedicated channel, where they are decoded and input into the ASR backend. Studies have shown that DSR generally performs better than NSR [85] because, in the latter model, speech is processed for optimal perceptual quality and this does not necessarily result in optimal recognition performance [178]. Various schemes for compressing the ASR features have been proposed in the literature. The disadvantage is that DSR requires some modifications to the existing mobile communications infrastructure, such as the addition of a dedicated channel for the transmission of the compressed MFCC feature bitstream.

Digalakis *et al.* in [39] evaluated the use of uniform and non-uniform scalar quantisers as well as product code vector quantisers (split vector quantiser) for compressing MFCCs between 1.2 and 10.4 kbps. They used a greedy-based bit allocation algorithm, where bits were added to each component and the word error rate (WER) was evaluated. The component which resulted in the largest improvement in recognition performance was chosen to receive the allocated bit. This procedure was continued until all bits had been allocated [39]. They concluded that split vector quantisers achieved word error rates (WER) similar to that of scalar quantisers while requiring less bits. Also, PDF-optimised non-uniform scalar quantisers performed better than uniform scalar quantisers, which suggested that the PDF of MFCCs were far from being uniformly distributed. Also,

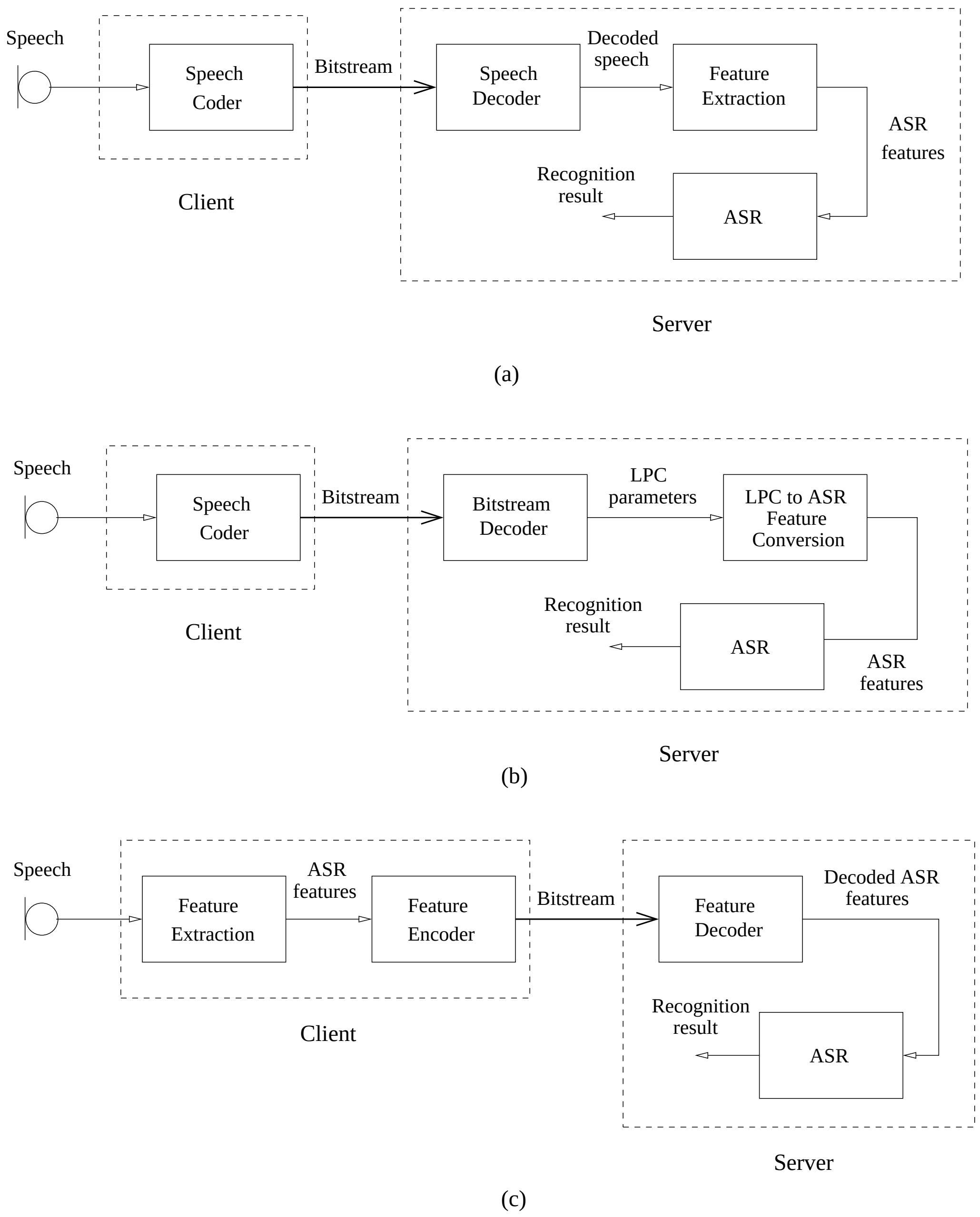

Figure 7.7: Client-server-based speech recognition modes: (a) Speech-based network speech recognition (NSR); (b) Bitstream-based network speech recognition; (c) Distributed speech recognition (DSR)

PDF-optimised scalar quantisation with non-uniform bit allocation performed significantly better than one with uniform bit allocation. They concluded that 2 kbps (20 bits/frame) was the required bitrate for split vector quantisation to achieve unquantised recognition performance.

Ramaswamy and Gopalakrishnan [144] investigated the application of tree-searched multistage vector quantisers with first-order linear prediction operating at 4 kbps (40 bits/frame). The current MFCC feature vector was subtracted from the previous quantised frame to give a residual vector. The first 12 coefficients of the residual vector were then quantised using a two-stage multistage vector quantiser, while the last coefficient, representing c_0, was scalar quantised. Their system achieved near identical recognition performance as the unquantised baseline system, with only minor degradation.

Transform coding, based on the discrete cosine transform (DCT), was investigated in [86] at 4.2 kbps. In this scheme, feature vectors of dimension 14 (13 MFCCs plus the energy coefficient) were processed. For each cepstral coefficient, eight temporally consecutive coefficients were grouped together and processed by the DCT, which exploited temporal correlation between the MFCC frames. The first DCT coefficient was quantised using PCM (12 bits) or DPCM (6 bits) with first order predictor, while the rest of the DCT coefficients were quantised using PCM (18 bits). The energy coefficient was encoded using PCM (12 bits) or DPCM (3 bits). The DPCM coding of the transform coefficients made the scheme insensitive to environmental variations [86].

Zhu and Alwan [206] used a two-dimensional DCT, where 12 successive MFCC frames were stacked together to form a block of 12×12. Zonal sampling was performed, where a fraction of the lowest energy components were set to zero and the remaining transform coefficients were scalar quantised and entropy coded with runlength and Huffman coding. This scheme is similar to the JPEG scheme for image coding. The advantage of this scheme, compared with that of [86], is that both intraframe as well as interframe correlation are exploited by the 2D-DCT. This leads to better energy compaction and hence allow for more data reduction. Noise-robust feature sets, such as peak isolated MFCC (MFCCP) [180] and variable frame-rate peak isolated MFCCs (VFR_MFCCP) [207] were also tested. Their results showed that, firstly, the quantised MFCCs always performed slightly worse than the unquantised MFCCs at all SNR levels. Secondly, the quantised noise-robust features at 624 bps resulted in recognition accuracies that even surpassed the unquantised

MFCCs at low SNRs.

The ETSI DSR standard [47] uses split vector quantisers to compress the MFCC vectors at 4.4 kbps (44 bits/frame). Feature vectors of dimension 14 (13 MFCCs and $\log E$) are split into pairs of subvectors, with the energy parameters, c_0 and $\log E$ belonging to the same pair. A weighted Euclidean distance measure is used for the energy parameter subvector.

Srinivasamurthy *et al.* [178] exploited correlation across consecutive MFCC features by using a DPCM scheme followed by entropy coding. Their scheme is a scalable one, where the bitstream is multiresolution or embedded. That is, a coarsely quantised, *base layer* is transmitted. If higher recognition performance is required, the client can transmit further *enhancement layers* which are combined with the base layer by the server to obtain higher quality features [178].

Bitrate Scalability

Even though vector quantisers generally give better recognition performance using less bits, they are not scalable in bitrate when compared with scalar quantiser-based schemes, such as DPCM and transform coders. In other words, the vector quantiser is designed to operate at a specific bitrate only and will need to be re-trained for other bitrates. *Bitrate scalability* is a desirable feature in DSR applications, since one may need to adjust the bitrate adaptively, depending on the network conditions. For instance, if the communications network is heavily congested, then it may be more acceptable to sacrifice some recognition performance by operating at a lower bitrate in order to offset long response times. In addition to this, the computational complexity of vector quantisers can be quite high, when compared with scalar quantiser-based schemes. This form of scalability contrasts with that mentioned by Sriniwasamurthy *et al.* [178], where the bitstream is embedded with a base (coarse) layer, followed by successive enhancement layers. In order to distinguish between the two forms of scalability, we term this latter form as *bitstream scalability.* In this study, we investigate *bitrate scalable* quantisation schemes only.

7.4 The ETSI Aurora-2 Experimental Framework

The purpose of the ETSI Aurora-2 experiment is to provide a common framework for evaluating noise-robust speech recognition systems. It consists of a clean speech database, a noise database, a standard MFCC-based frontend, and scripts for performing the various training and test sets. The recognition engine that is used is the HMM Toolkit (HTK) software [203].

The TIDigits database [96] forms the basis of the clean speech database, where the original 20 kHz speech was downsampled to 8 kHz and filtered using the frequency characteristic of ITU G.712 (300–3400 Hz). Aurora-2 also provides a database of eight background noises, which were deemed to be commonly encountered in real-life operating conditions for DSR. These noises were recorded at the following places [70]:

- Suburban train (subway)
- Crowd of people (babble)
- Car
- Exhibition hall (exhibition)
- Restaurant
- Street
- Airport
- Train station

This noise is added to the filtered clean speech at various SNRs to simulate noise corruption.

There are two training modes: training with clean speech[4] only and training with clean and noisy (multicondition) speech [70]. In multicondition training, the noises added are subway, babble, car, and exhibition. When training with clean speech only, the best recognition performance is achieved in matched conditions, ie. when testing with clean speech as well. However, when the speech to be tested has background noise, then multicondition training is desirable, as it includes the distorted speech in the training data [70].

[4] Note that all clean speech is filtered using the G.712 frequency characteristic before training.

For the testing, there are three test sets, known as test set A, B, and C. In test set A and B, 4004 test utterances from the TIDigits database are divided into four subsets of 1001 utterances each and four different types of noises are added to each subset at varying levels of SNRs (∞, 20, 15, 10, 5, 0, -5 dB)[5]. Therefore, there are a total of $4 \times 7 = 28$ recognition accuracies reported in test set A and B. In test set C, only two subsets of 1001 utterances and two noises are used, giving a total of 14 recognition accuracies.

In test set A, the subway, babble, car, and exhibition noises are added to each subset and these are the same noises used in multicondition training, hence test set A evaluates the system in matched conditions. In test set B, the other four noises, namely restaurant, street, airport, and train station, are used instead. Because these noises were not present in the multicondition training, then test set B evaluates the system in mismatched conditions (mismatched noise). Test set C contains two utterance subsets only (of the four) with the noises, subway and street, added. Both the speech and noise are filtered using the MIRS frequency characteristic before they are added, hence test set C evaluates the system in mismatched conditions (mismatched frequency characteristic) [70].

Whole word HMMs are used for modelling the digits with the following parameters [69]:

- 16 states per word (with 2 dummy states at beginning and end);
- left-to-right topology without skips over states;
- 3 Gaussian mixtures per state; and
- diagonal covariance matrices

For more details on the HTK reference recogniser and Aurora frontend, the reader should refer to [70, 47].

7.5 Setup of Experiments for Evaluating Different Quantisation Schemes in a DSR Framework

We have evaluated the recognition performance of various quantisation schemes using the publicly available HMM Toolkit (HTK) 3.2 software on the ETSI Aurora-2 database

[5] An SNR of ∞ dB means that no noise is added and we are testing with clean speech.

[70]. Training was done on clean data only (no multicondition training) and testing was performed using test set A. In order to see the recognition performance as a function of bitrate, we focus on the results of testing on *clean speech* (SNR of ∞ dB), where the four word recognition accuracies for each type of noise are averaged to give the final score for the specific quantisation scheme. In addition to this, the effect of different types of noise at varying levels of SNR on the recognition performance is also investigated at the bitrates of 1.2 kbps and 0.6 kbps for each quantisation scheme.

The ETSI DSR standard Aurora frontend [47] was used for the MFCC feature extraction. As a slight departure from the ETSI DSR standard, we have used 12 MFCCs (excluding the zeroth cepstral coefficient, c_0, and logarithmic frame energy, $\log E$) as the feature vectors to be quantised. This was done to maintain consistency in the bitrate-scalable GMM-based block quantisation scheme by avoiding arbitrary bit allocation, as c_0 and $\log E$ are sensitive to changes in recording level of a speech utterance and are generally coded independently [47, 86, 144].

It is well known that lower order cepstral coefficients are particularly sensitive to undesirable variations caused by factors such as transmission, speaker characteristics, and vocal efforts, etc. [80]. As bits are distributed on the basis of the variance of each MFCC, the bit allocations will be particularly sensitive to these spectral variations. In our scalar quantiser experiments, we have found this to degrade the performance of the recognition as too many bits are given to the lower order MFCCs. The bit allocation formula is derived through constrained minimisation of the MSE. However, quantisation based on the reduction of MSE between the original and quantised MFCC feature vector does not necessarily correlate to an improvement in recognition performance. Therefore, in order to reduce the effect of these variations on bit allocation, we have applied the cepstral liftering technique of [80] to the MFCCs using the following lifter window function, $w(n)$ [80]:

$$\begin{aligned} w(n) &= 1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right) \\ &\quad \text{where } n = 1, 2, \ldots, L \end{aligned} \tag{7.2}$$

where L is the feature length. This cepstral liftering procedure is analogous to the weighted distance measure used in LSF quantisation for speech coding. That is, the components of the vector that play a larger role in affecting the final result, which in our case is recognition performance, are emphasised by the lifter window. Cepstral mean subtraction

Table 7.1: Average word recognition accuracy as a function of bitrate and number of clusters for the memoryless GMM-based block quantiser (baseline accuracy = 98.01%)

Bitrate (kbps)	Recognition accuracy (in %)				
	2 cluster	4 cluster	8 cluster	16 cluster	32 cluster
0.3	23.46	19.99	16.69	8.06	8.06
0.4	43.52	53.25	57.67	23.25	9.07
0.6	68.66	79.73	85.72	87.59	82.03
0.8	86.24	90.32	91.45	93.70	94.48
1.0	90.53	94.18	95.03	95.49	96.05
1.2	93.88	95.85	95.94	96.40	96.68
1.5	95.96	96.46	96.96	97.17	97.23
1.7	97.02	96.98	97.16	97.28	97.38
2.0	97.34	97.17	97.54	97.58	97.69
2.2	97.58	97.33	97.62	97.70	97.69
2.4	97.58	97.54	97.69	97.90	97.74
3.0	97.90	97.81	97.87	97.83	97.93
4.4	97.99	97.99	98.09	98.04	98.03

(CMS) is applied to the decoded 12 MFCC features, which are concatenated with their corresponding delta and acceleration coefficients, giving the final feature vector dimension of 36 for the ASR system. The HTK parameter type is `MFCC_D_A_Z`. The baseline average recognition accuracy using unquantised MFCC features is 98.01%.

In the training of the single frame and multi-frame GMM-based block quantiser, 20 iterations of the EM algorithm were used to generate a 16 and 32 cluster GMM.

7.6 Recognition Performance of the Memoryless GMM-Based Block Quantiser

Table 7.1 shows the recognition accuracy for the memoryless GMM-based block quantiser at various bitrates and number of clusters. At 2 kbps, the recognition accuracy is roughly the same as the unquantised scheme. Between 2 kbps and 800 bps, the recognition performance gradually decreases where it can be seen that the higher cluster schemes maintain a higher accuracy. This may be attributed to the more accurate modelling of the source PDF by using more clusters. Since it has been shown in other studies [128] that using more clusters generally will reduce the quantisation distortion incurred at a fixed bitrate, it would be expected to indirectly lead to better recognition.

Table 7.2: Average word recognition accuracy as a function of bitrate and number of frames for 16 cluster multi-frame GMM-based block quantiser (baseline accuracy = 98.01%)

Bitrate (kbps)	Recognition accuracy (in %)			
	2 frames	3 frames	4 frames	5 frames
0.3	78.26	89.59	91.30	92.96
0.4	91.07	94.32	95.05	95.36
0.6	95.52	96.62	97.05	96.78
0.8	96.92	97.27	97.41	97.52
1.0	97.40	97.61	97.74	97.71
1.2	97.56	97.74	97.75	97.89
1.5	97.78	97.80	97.86	97.83
1.7	97.80	97.96	97.99	97.96
2.0	97.99	97.89	98.06	97.97
2.2	98.03	97.97	97.94	98.04

At bitrates below 800 bps, the recognition performance drops dramatically, where we have the situation of higher clusters leading to steeper decreases. For MFCC frames containing no information (all zero), the recognition accuracy is 8.06%. This situation may be explained by the shortage of bits to be allocated to all clusters. A 16 cluster quantiser, for instance, requires at least 4 bits in total to be able to uniquely identify each cluster (assuming a uniform allocation of levels) while a 32 cluster block quantiser requires at least 5 bits[6]. Therefore, the single frame GMM-based block quantiser performs poorly when the number of bits approaches $\log_2 m$, where m is the number of clusters.

7.7 Recognition Performance of the Multi-Frame GMM-Based Block Quantiser

Table 7.2 shows the average word recognition accuracy of the 16 cluster multi-frame GMM-based block quantiser for different bitrates and number of frames. It can be observed that this quantiser achieves an accuracy close to the unquantised, baseline system at 1 kbps or 10 bits/frame, which is half the bitrate of the single-frame GMM-based block quantiser. For bitrates lower than 600 bps, the performance gradually rolls off.

[6] In reality, quantiser levels are non-uniformly allocated in this scheme, so most of the available quantiser levels will be allocated to only a fraction of the cluster block quantisers. The decoder will be able to determine which cluster block quantisers are operational since it performs an identical bit allocation using the same stored models as the encoder.

Table 7.3: Average word recognition accuracy as a function of bitrate and number of clusters for 5 frame multi-frame GMM-based block quantiser (baseline accuracy = 98.01%)

Bitrate (kbps)	Recognition accuracy (in %)	
	16 clusters	32 clusters
0.2	82.94	87.70
0.3	92.96	94.20
0.4	95.36	96.03
0.6	96.78	97.06
0.8	97.52	97.58
1.0	97.71	97.57
1.2	97.89	97.89
1.5	97.83	97.93
1.7	97.96	97.96
2.0	97.97	97.95

In terms of quantiser distortion, the multi-frame GMM-based block quantiser generally performs better as more frames are concatenated together because interframe memory can be exploited by the KLT. Also, because the dimensionality of the vectors is high, the block quantisers operate at a higher rate. Comparing Table 7.2 with the 16 cluster column of Table 7.1, it can be observed that there is a trend between using more frames to reduce MFCC frame distortion and improving the recognition accuracy, at low bitrates. In other words, the average recognition accuracy gets progressively better as more and more frames are jointly quantised. At 300 bps, the recognition accuracy of jointly quantising five frames is roughly 14% higher than quantising 2 frames only. However, this comes at the expense of higher delay, computational, and memory requirements.

Compared with the results of the single frame GMM-based block quantiser in Table 7.1, the multi-frame scheme does not suffer from a dramatic drop in recognition accuracy at low bitrates. Unlike the single frame scheme, where there was a shortage of bits to distribute among clusters, the multi-frame GMM-based block quantiser is able to provide enough bits, thanks to the increased dimensionality of the vectors. For example, at 300 bps, a 16 cluster, single frame GMM-based block quantiser has a total bit budget of 3 bits. On the other hand, a 16 cluster, 2 frame multi-frame GMM-based block quantiser has 6 bits while a 3 and 4 frame scheme has 9 and 12 bits, respectively. Therefore, the multi-frame GMM-based block quantiser can operate at lower bitrates while maintaining good recognition performance.

Table 7.4: Average word recognition accuracy as a function of bitrate for non-uniform scalar quantiser (baseline accuracy = 98.01%)

Bitrate (kbps)	Recognition accuracy (in %)
0.6	38.17
0.8	72.31
1.0	86.68
1.2	93.27
1.5	95.45
1.7	96.17
2.0	96.97
2.2	97.21
2.4	97.40
3.0	97.76
4.4	97.96

Table 7.3 shows the average word recognition accuracy of a 16 cluster and 32 cluster multi-frame GMM-based block quantiser, where the number of frames is fixed at 5. As expected, using more clusters to reduce the quantised MFCC distortion has led to an improvement in recognition accuracy, at the cost of an increase in complexity and memory.

7.8 Comparison with the Recognition Performance of the Non-Uniform Scalar Quantiser

For the scalar quantisation experiment, each MFCC was quantised using a non-uniform Gaussian Lloyd-Max scalar quantiser whose bit allocation was calculated using the high resolution formula of (2.43). We have chosen this method over the WER-based greedy algorithm of [39] because of its computational simplicity and this allows us to scale any bitrate with ease.

Table 7.4 shows the average recognition accuracy of the non-uniform scalar quantiser. It can be seen that the accuracy decreases linearly in the range of 4.4 to 1.2 kbps and drops rapidly below this range. Comparing Table 7.4 with Tables 7.1 and 7.2, the GMM-based block quantisers use less bits than the non-uniform scalar quantiser to achieve a certain level of recognition accuracy. This may be attributed to the effectiveness of the KLT as a decorrelator and energy compactor, as well as the better modelling of the source PDF.

Table 7.5: Average word recognition accuracy, computational complexity (in kflops/frame), and memory requirements (ROM) as a function of bitrate for vector quantiser (baseline accuracy = 98.01%)

Bitrate (kbps)	Recognition accuracy (in %)	kflops/frame	ROM (in floats)
0.4	76.94	0.77	192
0.6	91.83	3.07	768
0.8	95.65	12.29	3072
1.0	96.85	49.51	12288
1.2	97.01	196.7	49152

7.9 Comparison with the Recognition Performance of the Unconstrained Vector Quantiser

An unconstrained, full-search vector quantiser was used to quantise single MFCC frames. In terms of minimising quantiser distortion, the vector quantiser is considered the optimum coding scheme [55], hence it will serve as an informal upper recognition bound for single frame quantisation and highlight the effectiveness of the multi-frame GMM-based block quantiser in exploiting interframe memory. Table 7.5 shows the average recognition accuracies at several bitrates as well as the computational and memory requirements of the vector quantiser. Comparing this with Table 7.1, the vector quantisation scheme achieves higher recognition than the single frame GMM-based block quantiser for all bitrates, which is consistent with the fact that the vector quantiser will always incur the least distortion of all quantisation schemes for a given dimension. Comparing with the performance of the multi-frame GMM-based block quantiser in Table 7.2, it can be observed that this scheme gives higher recognition accuracies than the vector quantiser for all bitrates considered. Even the 2 frame, multi-frame GMM-based block quantiser does better than the vector quantiser. Hence this shows that there is a considerable amount of correlation between MFCC frames that can be exploited by quantisation schemes.

As can be seen from Table 7.5, the computational complexity and memory requirements of the vector quantiser are dependent on the bitrate and can be quite high at medium bitrates like 1.2 kbps. On the other hand, the complexity of the GMM-based block quantiser, as shown in Table 7.6, is constant for all bitrates. Also of note is that, unlike the GMM-based block quantiser, the vector quantiser is *not bitrate scalable.*

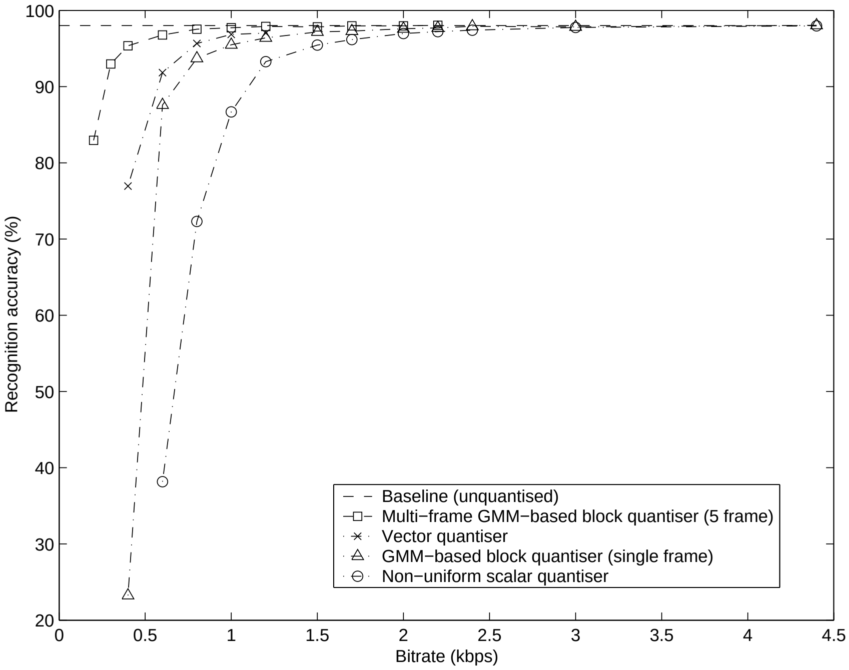

Figure 7.8: Summary of average word recognition accuracies for all quantisation schemes considered

Table 7.6: Bitrate independent computational complexity (in kflops/frame) and memory requirements (ROM) of the multi-frame GMM-based block quantiser as a function of number of concatenated vectors, p, and number of clusters, m

m	p	kflops/frame	ROM (floats)
16	1	13.65	3136
	2	22.86	10624
	3	32.07	22720
	4	41.28	39424
	5	50.50	60736
32	1	27.30	4416
	2	45.71	14976
	3	64.14	31936
	4	82.57	55296
	5	101.0	121216

7.10 Effect of Additive Noise on Recognition Performance

The effect of undesirable noise on the recognition performance is important and relevant to DSR systems, since the operator will most likely be immersed in background environmental noise that will also be captured by his/her mobile device. The Aurora-2 recognition task provides various types of background noise that is added to the clean speech at various SNR levels ($20, 15, 10, 5, 0, -5$ dB). In test set A, the four noises added are suburban train (subway), babble, car, and exhibition hall [70].

In this section, we evaluate the word recognition accuracy for all quantisation schemes, on speech corrupted with additive noise, as a function of SNR. The recognition models are trained on *clean speech* only (no multicondition training). The bitrates tested are 1.2 kbps (12 bits/frame) and 0.6 kbps (or 6 bits/frame) for all quantisation schemes. The notation we have used to abbreviate the quantisation schemes are as follows:

- GMM-5 is the five frame, multi-frame GMM-based block quantiser;
- GMM-1 is the memoryless GMM-based block quantiser;
- VQ is the unconstrained vector quantiser; and
- SQ is the non-uniform scalar quantiser.

Tables 7.7, 7.8, 7.9, and 7.10 shows the word recognition accuracy at 1.2 kbps when speech is corrupted with subway, babble, car, and exhibition noise, respectively. The results for the original, unquantised scheme are given for comparative purposes. We can see that the multi-frame GMM-based block quantiser (GMM-5) generally achieves the highest recognition accuracies of all quantisation schemes with the scalar quantiser performing the worst. The vector quantiser (VQ) is theoretically the best quantiser for a given dimension for minimising distortion, and this correlates generally to recognition performance, where we observe it outperforming the memoryless GMM-based block quantiser (GMM-1). Figure 7.9 shows the recognition accuracy for each quantisation scheme operating at 1.2 kbps, plotted against SNR for each of the noises. We can see that the quantisation schemes which exploit memory (GMM-5, GMM-1 and VQ) maintain a recognition that is close to the baseline, as opposed to the memoryless scalar quantiser which degrades rapidly as noise is added (particularly at SNRs of 20 dB and 15 dB).

Table 7.7: Word recognition accuracy for speech corrupted with subway noise at varying SNRs (in dB) at 1.2 kbps.

Quantisation	Recognition accuracy (in %)						
scheme	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
Unquantised	98.07	94.14	86.67	66.17	38.62	23.43	16.12
GMM-5	97.64	92.48	82.68	58.89	32.61	21.86	15.23
VQ	97.11	92.26	81.30	59.32	31.62	19.65	14.03
GMM-1	96.44	89.13	77.40	50.78	27.11	19.37	13.82
SQ	92.85	68.47	48.42	30.61	22.35	17.38	12.56

Table 7.8: Word recognition accuracy for speech corrupted with babble noise at varying SNRs (in dB) at 1.2 kbps

Quantisation	Recognition accuracy (in %)						
scheme	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
Unquantised	98.07	95.92	90.69	74.94	45.56	22.91	12.64
GMM-5	98.13	94.98	87.30	65.36	36.73	20.80	12.12
VQ	97.16	92.93	85.01	65.11	36.19	19.62	10.67
GMM-1	96.58	91.48	81.92	60.94	34.61	19.35	10.94
SQ	93.80	67.87	47.64	29.87	21.31	16.51	10.28

Table 7.9: Word recognition accuracy for speech corrupted with car noise at varying SNRs (in dB) at 1.2 kbps

Quantisation	Recognition accuracy (in %)						
scheme	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
Unquantised	97.97	95.59	88.88	68.42	36.09	20.61	13.30
GMM-5	97.88	93.80	83.21	55.92	29.38	18.52	12.73
VQ	97.02	93.14	83.12	54.94	27.02	19.27	11.78
GMM-1	96.39	91.05	78.08	49.42	25.17	18.01	11.24
SQ	93.44	70.44	45.87	27.86	22.19	16.94	11.72

Table 7.10: Word recognition accuracy for speech corrupted with exhibition noise at varying SNRs (in dB) at 1.2 kbps

Quantisation	Recognition accuracy (in %)						
scheme	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
Unquantised	97.93	93.34	85.56	62.79	33.42	19.01	10.74
GMM-5	97.90	92.84	81.70	54.83	28.60	18.94	10.86
VQ	96.73	91.36	77.63	50.45	26.87	16.94	10.77
GMM-1	96.24	89.45	75.84	45.63	25.42	17.46	11.66
SQ	92.97	72.79	48.32	28.63	20.73	13.82	8.89

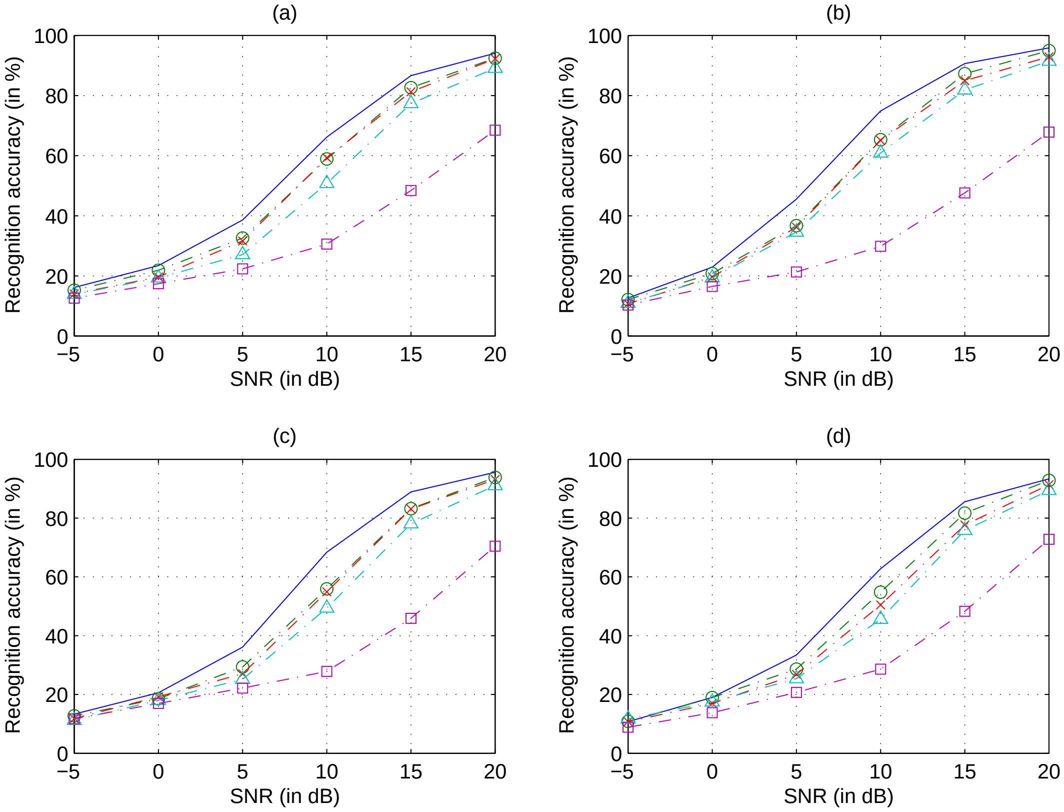

Figure 7.9: Plot of recognition accuracy versus SNR for all quantisation schemes at 1.2 kbps: (a) subway noise; (b) babble noise; (c) car noise; and (d) exhibition noise. (Solid lines are unquantised, circles are GMM-5, crosses are VQ, triangles are GMM-1, squares are SQ)

Table 7.11: Word recognition accuracy for speech corrupted with subway noise at varying SNRs (in dB) at 0.6 kbps.

Quantisation	Recognition accuracy (in %)						
scheme	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
Unquantised	98.07	94.14	86.67	66.17	38.62	23.43	16.12
GMM-5	96.38	88.73	74.33	48.17	26.93	18.73	13.57
VQ	94.40	82.22	71.29	48.30	26.44	15.84	11.21
GMM-1	84.41	77.56	64.14	44.24	25.15	16.43	11.36
SQ	8.32	8.29	8.29	8.26	8.14	8.11	8.07

Table 7.12: Word recognition accuracy for speech corrupted with babble noise at varying SNRs (in dB) at 0.6 kbps

Quantisation	Recognition accuracy (in %)						
scheme	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB
Unquantised	98.07	95.92	90.69	74.94	45.56	22.91	12.64
GMM-5	97.10	91.54	77.90	53.78	30.05	18.02	11.25
VQ	91.66	83.25	72.79	52.39	29.96	16.35	10.34
GMM-1	89.94	76.12	62.61	43.02	25.94	15.90	10.13
SQ	8.25	8.22	8.16	8.13	8.16	8.13	8.13

Tables 7.11, 7.12, 7.13, and 7.14 shows the word recognition accuracy at 0.6 kbps when speech is corrupted with subway, babble, car, and exhibition noise, respectively. Similar to the previous case at the higher bitrate of 1.2 kbps, the multi-frame GMM-based block quantiser generally achieves higher recognition accuracies than the other schemes, with the scalar quantiser being the worst. There are cases where the GMM-1 scheme achieves a slightly higher recognition performance, such as for an SNR of 0 dB of subway noise, but this discrepancy is insignificant. Figure 7.10 shows the recognition accuracy at 0.6 kbps as a function of SNR. It is particularly interesting to note that the difference in recognition performance between the multi-frame and memoryless GMM-based block quantiser at 0.6 kbps is larger than that observed at 1.2 kbps (in Figure 7.9), for medium to high SNRs. For instance, when speech is corrupted with babble noise with the SNR at 15 dB, the multi-frame GMM-based block quantiser achieves a recognition accuracy that is about 15% higher than the memoryless GMM-based block quantiser at 0.6 kbps, while at 1.2 kbps, that difference is only 5%. This shows that there is an advantage in using more efficient quantisation schemes at moderately high SNRs and low bitrates.

Another interesting observation, especially from the low bitrate results, is the dimin-

Table 7.13: Word recognition accuracy for speech corrupted with car noise at varying SNRs (in dB) at 0.6 kbps

Quantisation	Recognition accuracy (in %)						
scheme	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
Unquantised	97.97	95.59	88.88	68.42	36.09	20.61	13.30
GMM-5	96.51	89.02	72.89	44.92	24.22	17.00	11.03
VQ	91.92	84.22	70.00	44.02	23.11	15.81	10.86
GMM-1	87.59	76.86	59.86	36.92	21.00	14.70	11.09
SQ	8.29	8.23	8.29	8.26	8.23	8.23	8.23

Table 7.14: Word recognition accuracy for speech corrupted with exhibition noise at varying SNRs (in dB) at 0.6 kbps

Quantisation	Recognition accuracy (in %)						
scheme	∞ dB	20 dB	15 dB	10 dB	5 dB	0 dB	−5 dB
Unquantised	97.93	93.34	85.56	62.79	33.42	19.01	10.74
GMM-5	97.13	89.60	73.25	43.04	24.13	16.94	9.60
VQ	92.35	84.60	70.01	42.58	22.80	14.47	10.86
GMM-1	87.41	78.96	59.86	34.16	20.18	12.74	9.16
SQ	7.93	7.87	7.87	7.84	7.81	7.81	7.81

ishing advantage, as the SNR degrades, of the more efficient quantisation schemes, like GMM-5, over the less efficient ones, such as GMM-1. For example, testing on clean speech and speech with an SNR of 20 dB, the GMM-5 scheme at 0.6 kbps scheme mostly achieves roughly 10% or better recognition performance over GMM-1 at the same bitrate. This advantage diminishes as more and more noise is added, with the difference reduced to roughly 2% or less at an SNR of −5 dB. A similar trend can also be seen in Figure 7.10, when comparing VQ with GMM-1, where the advantages of VQ, in terms of lower MSE distortion, diminish as more noise is added. We can therefore conclude that, firstly, the advantages of using more efficient quantisation schemes manifest themselves more at SNRs of 10 dB and higher. Also for DSR in noisy environments where the SNR is very low, the noise robustness of the underlying speech recognition system becomes the dominant factor, rather than MFCC quantisation efficiency, when it comes to recognition performance. This is, in fact, consistent with the results from [206], where the recognition performance as a result of using quantised MFCCs, was always worse than when using the unquantised MFCCs, for all levels of SNR. By quantising more noise-robust features, such as those used in [206] (MFCCPs and VFR_MFCCPs), better recognition accuracy can be achieved.

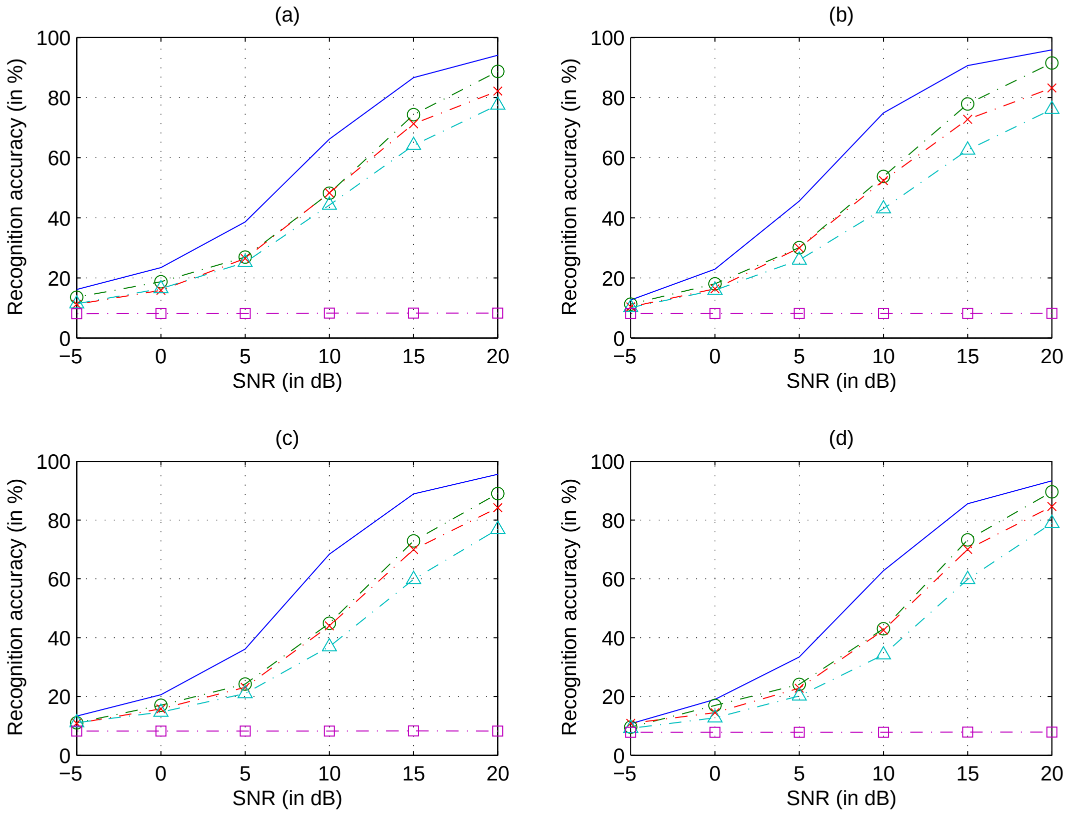

Figure 7.10: Plot of recognition accuracy versus SNR for all quantisation schemes (excluding SQ) at 0.6 kbps: (a) subway noise; (b) babble noise; (c) car noise; and (d) exhibition noise. (Solid lines are unquantised, circles are GMM-5, crosses are VQ, triangles are GMM-1, squares are SQ)

7.11 Chapter Summary

In this chapter, we provided a brief review of automatic speech recognition with particular emphasis on the speech features such as the Mel frequency-warped cepstral coefficients. Following this, we reviewed the literature that investigated various modes of client/server-based speech recognition systems, such as network speech recognition and distributed speech recognition. The experimental database used for evaluating the performance of our MFCC quantisation schemes as well as the parameters for the recognition task were described given in detail. Next, we presented our results on MFCC quantisation in a DSR framework using multi-frame GMM-based block quantisers and compared its performance against the memoryless GMM-based block quantiser, the non-uniform scalar quantiser, and the unconstrained vector quantiser. The multi-frame GMM-based block quantiser achieved better recognition at lower bitrates, exhibiting negligible degradation of 1% (WER of 2.5%) in recognition performance over the baseline system at 800 bps and 5% (WER of 7%) at 300 bps. Unlike vector quantisation schemes, the multi-frame GMM-based block quantiser is scalable in bitrate and has a complexity that is independent of bitrate. The performance of the multi-frame GMM-based block quantiser in the presence of noise was also evaluated. It was found that the recognition performance of relatively high SNRs was influenced mostly by the quantisation scheme. However, at low SNRs, the effect of quantisation efficiency diminishes and recognition performance is dependent on the noise robustness of the underlying features.

Chapter 8

Conclusions and Future Research

8.1 Chapter Summary and Conclusions

This dissertation has examined block and vector quantisation schemes that are efficient, in terms of rate-distortion and computational requirements, and reported on their performance in four unique application areas. In this section, we give a summary as well as present the findings and conclusions of each chapter.

8.1.1 Chapter 2: Efficient Block Quantisation

This chapter provided a general introduction to block quantisation, which is an example of a transform coder. The decorrelating properties of the Karhunen-Loève transform and its role in block quantisation were described. We also reviewed the discrete cosine transform as a useful alternative transform to the KLT. For sources which have Gauss-Markov properties, the DCTs decorrelating ability is similar to that of the KLT, hence the DCT is popularly used in image coding.

We provided a literature review of adaptive transform coding schemes, which resolve the problems of data non-stationarity by partitioning the vector space into local regions and designing transforms adapted to the statistics of each region. Additionally, a simple scheme using K-means clustering and local block quantisers was described and this formed a useful baseline for evaluating the recent schemes that utilise Gaussian mixture models for estimating the PDF. Following this, we gave a detailed summary of the GMM-based block quantisation scheme of [183].

We presented our modification to the GMM-based block quantiser, that replaces the KLT with a DCT. Due to the data independence property and fixed orthogonal bases of the DCT, the complexity of the new GMM-DCT-based block quantiser is considerably lowered. This modified scheme is expected to be competitive with the KLT-based GMM-based block quantiser for image coding, since images tend to have Gauss-Markov statistics and are highly correlated. We also described our multi-frame GMM-based block quantiser, that exploits interframe correlation using the KLT by concatenating successive frames into larger ones.

A new bit encoding technique was introduced that allows the use and encoding of fractional bits in a fixed-rate block quantiser. This scheme uses the concept of a generalised positional number system and is simple in implementation. To complement this fractional bit technique, we also described some heuristic algorithms for dealing with bit allocation issues.

8.1.2 Chapter 3: Efficient Vector Quantisation

This chapter provided a general review of vector quantisation, its advantages over the scalar quantiser, and its limitations, with regards to its exponential growth of complexity as a function of the number of bits and dimensionality. Product code vector quantisers, such as the split and multistage vector quantiser, alleviate the complexity issue by dividing the quantisation process into codebooks of lower dimensionality, or sequential and independent stages, respectively. These structural constraints though cause suboptimal quantisation performance. We have also identified and analysed the main source of suboptimality in the split vector quantiser (SVQ), namely the vector splitting which degrades the memory advantage, the shape advantage, and the space-filling advantage. In order to address at least two of these suboptimalities, we have introduced a new type of product code vector quantiser called the switched split vector quantiser (SSVQ), which consists of a hybrid of a full-dimension, unconstrained switch vector quantiser and numerous split vector quantisers. The first stage (ie. switch vector quantiser) allows the SSVQ to exploit global statistical dependencies as well as match the marginal PDF shape of the data, which would otherwise have not been exploited by normal SVQ. Also, the tree structured characteristic of the switch vector quantiser provides a dramatic reduction in search complexity. We have shown via computer simulations of 2-D vector quantisation how SSVQ is superior to SVQ in terms of quantisation performance and computational complexity. The only disadvantage of SSVQ is the increase in memory requirements.

8.1.3 Chapter 4: Lossy Image Coding

In this chapter, we have presented a comprehensive literature review of image coding techniques, which includes vector quantisation, transform coding, and subband and wavelet-based coding. Fundamental to image subband coding and image processing in general, is the non-expansive filtering of data with a finite length. Similar to the procedure given in [106], the symmetric extension method was examined in depth with examples provided for even and odd tapped filters as well as filters with unequal lengths.

The remainder of the chapter was dedicated to the results and discussion of various quantisation schemes, such as the block quantiser based on the KLT and DCT and the GMM-based block quantiser. It was shown that the GMM-based block quantiser achieved higher PSNRs and better subjective quality than the traditional fixed-rate block quantiser/transform coder at a given bitrate, which demonstrates the advantages of accurate source PDF estimation and the use of multiple decorrelating transforms. Because images are highly correlated and have Gauss-Markov properties, replacing the KLT with the data dependent DCT should result in comparable performance. Through PSNRs and visual inspection, we showed that the GMM-DCT-based block quantiser is comparable in quantisation performance, with only a fraction of the complexity. Next, a novel and low complexity method of encoding fractional bits in a fixed-rate framework and heuristic algorithms for compensating quantiser levels in bit allocation were evaluated and shown to improve the PSNR slightly. Finally, we presented a method of pre-processing an image using the wavelet transform before block quantisation that reduces block artifacts and improves the image quality.

8.1.4 Chapter 5: LPC Parameter Quantisation in Narrowband Speech Coding

In this chapter, we first reviewed the basics of speech coding, such as speech production and the modelling of speech using linear prediction analysis. The operation of various speech coders was also described and this highlighted the role and importance of LPC quantisation. Different LPC parameter representations, that are both robust to quantisation and provide simple checks for filter stability, were covered. The line spectral frequencies are one of the more popular representations and were thus used in our evaluation of various quantisation schemes.

The first quantisation scheme that we evaluated was the multi-frame GMM-based block quantiser, which has the advantage of bitrate scalability and bitrate independent complexity. By extending the decorrelating transform to exploit the linear dependencies between multiple frames, the multi-frame GMM-based block quantiser was able to achieve transparent coding at bitrates as low as 21 bits/frame, though the computational complexity and memory requirements become an issue. This quantisation scheme was compared with scalar quantisers, the split vector quantiser, the multistage vector quantiser, and the single-frame GMM-based block quantiser, and was generally found to perform better in terms of spectral distortion, bitrate, and complexity.

The switched split vector quantiser was also evaluated as an LSF quantiser. Transparent coding was achieved at bitrates as low as 22 bits/frame, though the memory requirements of the two-part SSVQ were relatively high. It was determined that the three-part SSVQ, with transparent coding at 23 bits/frame, was well-balanced in terms of quantisation performance and complexity. Compared with other single-frame quantisers, the SSVQ achieved generally better spectral distortion performance. One aspect that the SSVQ excelled was the low computational complexity.

8.1.5 Chapter 6: LPC Parameter Quantisation in Wideband Speech Coding

This chapter began with the definition of wideband speech and described its advantages over toll-quality narrowband speech, such as improved naturalness and the ability to distinguish between fricatives, as well as provide better presence of the speaker, all of which can alleviate listener fatigue. We have also shown through the visual inspection of LPC-based spectral envelopes that, due to the extra bandwidth, a higher order LPC analysis is required to capture most of the short-term correlation information in the speech. Following this, we have given a review of the state-of-the-art coding schemes for wideband speech as well as the industry standard coders such as the ITU-T G.722 (subband/ADPCM coder) and ITU-T G.722.2 (AMR-WB ACELP coder).

Since the focus of this chapter is primarily on spectral quantisation for wideband LPC-based speech coders such as CELP, we provided a review of quantisation schemes that have been reported in the wideband speech coding literature. We have also evaluated some of these schemes such as PDF-optimised scalar quantisers, the unconstrained vector quantiser, and the GMM-based block quantiser on the two competing LPC parameter representations: line spectral frequencies (LSFs) and immittance spectral pairs (ISPs). Our experimental results have shown that ISPs are superior to LSFs by 1 bit/frame in independent quantisation schemes, such as scalar quantisers; while LSFs are the superior representation in joint vector schemes, such as the vector quantiser and GMM-based block quantiser. Through the extrapolation of the operating distortion-rate curve of unconstrained vector quantisation, we also derived an informal lower bound of 35 bits/frame and 36 bits/frame, for the transparent coding of wideband LSFs and ISPs, respectively. We speculate that this may be due to the fact that the last ISP parameter, which is not really a 'frequency', is not correlated with the other ISPs and hence impacts on the memory advantage of block and vector quantisation schemes, which aim to minimise the unweighted MSE for each vector as a whole. Furthermore, because this parameter is a reflection coefficient, it does not possess the error localisation properties of LSFs, but rather propagates errors throughout the entire spectrum. Therefore, additional measures may need to be taken, such as independent quantisation of the last ISP, or use of a weighted distance measure, when vector quantising ISPs.

Finally, we presented and discussed the results of the switched split vector quantiser (SSVQ) and the multi-frame GMM-based block quantiser, for coding wideband LSF and ISF vectors. The SSVQ was able to achieve transparent coding at 43 bits/frame and 44 bits/frame, when using an unweighted mean-squared-error (MSE) on LSFs and ISFs, respectively. The spectral distortion performance of the SSVQ on LSFs was improved by using a weighted MSE that emphasised LSFs that were located near peaks in the power spectrum. The resulting scheme was transparent at 42 bits/frame. The multi-frame GMM-based block quantiser was able to achieve transparent coding at 37 bits/frame and 38 bits/frame with a moderate computational complexity for LSFs and ISFs, respectively. These two quantisation experiments again confirm our finding that LSFs are superior to ISFs by about 1 bit/frame in joint vector quantisation schemes.

8.1.6 Chapter 7: MFCC Quantisation in Distributed Speech Recognition

In this chapter, we provided a brief review of automatic speech recognition with particular emphasis on the speech features such as the Mel frequency-warped cepstral coefficients. Following this, we reviewed the literature that investigated various modes of client/server-based speech recognition system, such as network speech recognitions and distributed speech recognition. The experimental database used for evaluating the performance of our MFCC quantisation schemes as well as the parameters for the recognition task were described in detail. Next, we presented our results on MFCC quantisation in a DSR framework using multi-frame GMM-based block quantisers and compared its performance against the memoryless GMM-based block quantiser, the non-uniform scalar quantiser, and the unconstrained vector quantiser. The multi-frame GMM-based block quantiser achieved better recognition at lower bitrates, exhibiting negligible degradation of 1% (WER of 2.5%) in recognition performance over the baseline system at 800 bps and 5% (WER of 7%) at 300 bps. Unlike vector quantisation schemes, the multi-frame GMM-based block quantiser is scalable in bitrate and has a complexity that is independent of bitrate. The performance of the multi-frame GMM-based block quantiser in the presence of noise was also evaluated. It was found that the recognition performance of relatively high SNRs was influenced mostly by the quantisation scheme. However, at low SNRs, the effect of quantisation efficiency diminishes and recognition performance is dependent on the noise robustness of the underlying features.

8.2 Suggestions for Future Research

This dissertation has examined block and vector quantisation schemes that are efficient, in terms of rate-distortion and computational requirements. Improvements in the rate-distortion efficiency have been derived from compensating the suboptimalities of each quantisation scheme, whether it be through accurate estimation of the source via parametric modelling (in the case of the GMM-based block quantiser), or exploitation of dependencies before applying constrained quantisation (in the case of the switched split vector quantiser). We have evaluated these schemes in four different applications, where efficient quantisation is required. In order to continue this line of research, this section lists some possible directions for further investigations.

In Chapter 2, two modern transform coding paradigms were discussed. They can be classified as either hard or soft clustering. The hard clustering paradigm takes the form of adaptive transform coders, where the vector space is partitioned into disjoint regions and a local transform is designed. Archer and Leen [13] developed the optimal adaptive transform coder, where the partitioning of the vector space, the local transform design, and quantiser design are performed jointly, in order to minimise distortion. The soft clustering paradigm involves modelling the source of the vectors using a mixture of individual and overlapping Gaussian sources, and designing optimal block quantisers for each source. It would be interesting to compare and contrast these two transform coding paradigms.

In Chapter 3, we identified the sources of suboptimality in the split vector quantiser and proposed the switched split vector quantiser, which compensates for the losses in the memory and shape advantages, by using a switch vector quantiser. As we have discussed in this chapter, the switched vector quantiser aims to exploit global dependencies in the vector space initially, which would otherwise have been neglected by an initial vector split. A further step would be to exploit the dependencies within each local cluster of vectors to increase the efficiency of the local split vector quantiser. Therefore, one possible path that warrants further research is to use different sized splitting for each of the local split vector quantisers. This variable splitting algorithm should adapt to the unique statistics of each cluster in order to minimise distortion. Another possible path is to develop a method of decorrelating the subvectors within each local split vector quantiser. That is, we can improve the efficiency of the split vector quantiser by removing correlation between each of the subvectors. A transform that can perform this *partial decorrelation* was discovered during the course of this research and this could be applied to the SSVQ.

In Chapter 5, we evaluated various quantisation schemes for LPC parameter coding used in narrowband speech coding. The training and test speech were assumed to be clean and noise-free. However, in a real-life scenario, the negative effects of additive background noise (such as babble and car noises) on the output quality of a speech coder cannot be neglected. In addition to this, the transmission channel was assumed to be perfect and lossless. However, this is not always possible in practice. Therefore, it is necessary to further investigate the effects of background noise and frame erasure conditions on the spectral distortion performance for each of the quantisation schemes considered in this dissertation.

In Chapter 6, we compared the relative performance of line spectral frequencies (LSFs) and immittance spectral pairs (ISPs) in quantisation experiments for wideband speech. We showed that ISPs outperformed LSFs in independent scalar quantisation while the trend was reversed for joint block and vector quantisation schemes. The unique nature of the last ISP warrants further investigation into how joint vector quantisation schemes can be applied in an optimal way. Similar to the case in MFCC quantisation, where the logarithmic energy coefficient, $\log E$, is quantised separately, separate quantisation of the last ISP may be required, which can be handled in a $(15, 1)$ split vector quantisation scheme. A weighted Euclidean distance measure needs to be developed for this type of quantisation scheme that takes into account the nature of the last ISP and how deviations affect the reconstructed power spectrum. In this chapter, we have also derived an informal lower bound on the number of bits required to transparently code LSFs and ISPs, by extrapolating the operating distortion-rate curve of the vector quantiser. We pointed out, however, that issues with 'over-training' were present that affected the tightness of this bound. Further work in determining a lower bound would involve undertaking the process outlined by Hedelin and Skoglund [66], where a GMM with bounded support is derived from training data, and used to generate artificial vectors. These vectors can then be used to train the vector quantiser. Lastly, as we have pointed out earlier, the effects of background noise and frame erasure conditions need to be investigated as well.

In Chapter 7, we evaluated the recognition performance of the various quantisation schemes in the task of quantising Mel frequency-warped cepstral coefficients (MFCCs). The distance measure that was used to design and search the quantiser codebook was mean-squared-error that is weighted with a fixed lifter window. It would be interesting to derive a weighted distance measure, that incorporates both fixed and dynamic weights, to emphasise parts of the speech that may be beneficial for recognition. Also, evaluating these quantisation schemes on feature sets that have been determined to be more robust to noise deserves further attention. The Aurora-2 connected-digits recognition task has a relatively small vocabulary and thus is not a very complex recognition task. It is proposed that the DSR evaluation be extended to databases with a larger vocabulary, such as the Aurora-3 and the DARPA Resource Management (RM) databases.

Bibliography

[1] "IEEE transactions, journals, and letters: information for authors", IEEE Periodicals, Transactions/Journals Department, 2000, pp. 4–5.

[2] "3rd generation partnership project; Technical specification group services and system aspects; Mandatory speech codec speech processing functions; Adaptive multi-rate (AMR) speech codec; Transcoding functions (Release 5)", Technical Specification TS 26.090, 3rd Generation Partnership Project (3GPP), June 2002.

[3] "3rd generation partnership project; Technical specification group services and system aspects; speech codec speech processing functions; AMR wideband speech codec; Transcoding functions (Release 5)", Technical Specification TS 26.190, 3rd Generation Partnership Project (3GPP), Dec 2001.

[4] "3rd generation partnership project; Technical specification group services and system aspects; speech codec speech processing functions; AMR wideband speech codec; Transcoding functions (Release 5)", Technical Specification TS 26.190, 3rd Generation Partnership Project (3GPP), Dec 2001.

[5] J.-P. Adoul and R. Lefebvre, "Wideband speech coding", in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Ed. Amsterdam: Elsevier, 1995, pp. 289–309.

[6] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using vector quantization in the wavelet domain", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1990, pp. 2297–2300.

[7] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform", *IEEE Trans. Image Processing*, vol. 1, no. 2, pp. 205–220, Apr. 1992.

[8] M. Antonini, T. Gaidon, P. Mathieu, and M. Barlaud, "Wavelet transform and image coding", in *Wavelets in Image Communication*, M. Barlaud, Ed. Amsterdam: Elsevier, 1994, pp. 65–188.

[9] C. Archer and T.K. Leen, "Optimal dimension reduction and transform coding with mixture principal components", in *Proceedings of International Joint Conference on Neural Networks*, July 1999.

[10] C. Archer and T.K. Leen, "From mixtures of mixtures to adaptive transform coding", in *Proceedings of International Joint Conference on Neural Networks*, July 1999, pp. 925–931.

[11] C. Archer and T.K. Leen, "Adaptive transform coding as constrained vector quantization", in *Proceedings of the IEEE Workshop*, Dec. 2000.

[12] C. Archer and T.K. Leen, "The coding-optimal transform", in *Proceedings of the Data Compression Conference*, IEEE Computer Society Press, 2001.

[13] C. Archer and T.K. Leen, "A generalized Lloyd-type algorithm for adaptive transform coder design", *IEEE Trans. Signal Processing*, vol. 52, no. 1, pp. 255–264, Jan. 2004.

[14] B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave", *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, Aug. 1971.

[15] B.S. Atal and M.R. Schroeder, "Predictive coding of speech signals and subjective error criteria", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, no. 3, pp. 247–254, Jun. 1979

[16] B.S. Atal and J.R. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates", in *Proc. IEEE. Int. Conf. Acoust., Speech, Signal Processing*, May 1982, pp. 614–617.

[17] B.S. Atal, R.V. Cox, and P. Kroon, "Spectral quantization and interpolation for CELP coders", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, Scotland, May 1989, pp. 69–72.

[18] A. Bernard and A. Alwan, "Low-bitrate distributed speech recognition for packet-based and wireless communication", *IEEE. Trans. Speech Audio Processing*, vol. 10, no. 8, pp. 570–579, Nov. 2002.

[19] B. Bessette, R. Salami, R. Lefebvre, M. Jelfnek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)", *IEEE Trans. Speech Audio Processing*, vol. 10, no. 8, pp. 620–636, Nov. 2002.

[20] V. Bhaskaran and K. Konstantinides, *Image & Video Compression Standards*, 2nd Edition, Kluwer International Series, Boston, 1997.

[21] B. Bhattacharya, W.P. LeBlanc, S.A. Mahmoud, and V. Cuperman, "Tree searched multi-stage vector quantization of LPC parameters for 4 kb/s speech coding", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1992, pp. I-105–I-108.

[22] J.A. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models", Technical Report, University of Califorina, Berkeley, ICSI-TR-97-021, 1997.

[23] Y. Bistritz and S Pellerm, "Immittance spectral pairs (ISP) for speech encoding", in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1993, pp. II-9–II-12.

[24] G. Biundo, S. Grassi, M. Ansorge, F. Pellandini and P.A. Farine, "Design techniques for spectral quantization in wideband speech coding", in *Proc. of 3rd COST 276 Workshop on Information and Knowledge Management for Integrated Media Communication*, Budapest, Oct. 2002, pp. 114-119.

[25] P.J. Burt and E.H. Adelson, "The Laplacian pyramid as a compact image code", *IEEE Trans. Commun.*, vol. 31(4), pp. 532–540, Apr. 1983.

[26] A. Buzo, A.H. Gray Jr., R.M. Gray, and J.D. Markel, "Speech coding based upon vector quantization", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 562–574, Oct. 1980.

[27] J.P. Campbell Jr., V.C. Welch, and T.E. Tremain, "An expandable error-protected 4800 bps CELP coder (U.S. federal standard 4800 bps voice coder)", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, U.K., 1989, pp. 735–738.

[28] W. Chen and C.H. Smith, "Adaptive coding of monochrome and color images", *IEEE Trans. Commun.*, vol. COM-25(11), pp. 1285–1292, Nov. 1977.

[29] J.H. Chen and D. Wang, "Transform predictive coding of wideband speech signals", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 275–278.

[30] P. Combescure, J. Schnitzler, K. Fischer, R. Kirchherr, C. Lamblin, A. le Guyader, D. Massaloux, C. Quinquis, J Stegmann, P. Vary, "A 16, 24, 32 kbit/s wideband speech codec based on ATCELP", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 5–8.

[31] R.V. Cox, "Speech Coding Standards", in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Ed. Amsterdam: Elsevier, 1995, pp. 49–78.

[32] R.E. Crochiere, S.A. Webber, and J.L. Flanagan, "Digital coding of speech in subbands", *Bell System Technical Journal*, vol. 55, pp. 1069–1085, Oct. 1976.

[33] A. Crossman, "A variable bit rate audio coder for videoconferencing", in *IEEE Workshop on Speech Coding for Telecommunications*, pp. 7–8, 1993.

[34] G. Davis and A. Nosratinia, "Wavelet-based image coding: an overview", *Applied and Computational Control, Signals, and Circuits*, vol. 1, no. 1, pp. 205–269, 1998.

[35] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[36] J. Dejener, "Digital speech compression: putting the GSM 06.10 RPE-LTP algorithm to work", 1994. Available: http://www.ddj.com/documents/s=1012/ddj9412b/

[37] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *J. Royal Stat. Soc.*, vol. 39, pp. 1–38, 1977.

[38] R.A. DeVore and B.J. Lucier, "Wavelets", Technical Report, University of South Carolina.

[39] V.V. Digalakis, L.G. Neumeyer and M. Perakakis, "Quantization of cepstral parameters for speech recognition over the world wide web", *IEEE J. Select. Areas Commun.*, vol. 17, no. 1, pp. 82–90, Jan 1999.

[40] R. Dony and S. Haykin, "Optimally adaptive transform coding", *IEEE Trans. Image Processing*, vol. 4, no. 10, pp. 1358–1370, 1995.

[41] H. Dudley, "The vocoder", *Bell Labs Rec.*, 19, pp. 122, 1939.

[42] E.R. Duni, A.D. Subramaniam, and B.D. Rao, “Improved quantization structures using generalised HMM modelling with application to wideband speech coding”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 2004, pp. 161-164.

[43] M. Effros, P. Chou, and R.M. Gray, “Weighted universal image compression”, *IEEE Trans. Image Processing*, vol. 8, no. 10, pp. 1317–1328, 1999.

[44] M. Effros, H. Feng, and K. Zeger, “Suboptimality of the Karhunen-Loève transform for transform coding”, *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1605–1619, Aug. 2004.

[45] T. Eriksson, J. Lindén, and J. Skoglund, “Exploiting interframe correlation in spectral quantization–a study of different memory VQ schemes”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 765–768.

[46] D. Esteban and C. Galand, ‘Application of quadrature mirror filters to split band voice coding schemes’, in *Proc. IEEE Int. Conf. Acoustic, Speech, Signal Processing*, 1977, pp.191-195.

[47] “Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms”, Tech. Rep. Standard ES 201 108 v1.1.3, European Telecommunications Standards Institute (ETSI), September 11 2003.

[48] S. Euler and J. Zinke, “The influence of speech coding algorithms on automatic speech recognition”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Adelaide, Australia, Apr. 1994, pp. I-621–I-624.

[49] A. Fingerhut, *U.S. Department of Defense LPC-10 Voice Coder*, Release 1.5, Washington University, Oct 1997. Available: http://www.arl.wustl.edu/ jaf/lpc/lpc10-1.5.tar.gz

[50] D. Gabor, “Theory of communication”, *Proc. IEE*, 1936.

[51] A. Gallardo-Antolin, F. Diaz-de-Maria and F. Valverde-Albacete, “Recognition from GSM digital speech”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 1443–1446.

[52] W.R. Gardner and B.D. Rao, "Theoretical analysis of the high-rate vector quantization of LPC parameters", *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 367–381, Sept. 1995.

[53] A. Gersho, "Asymptotic optimal block quantization", *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 373–380, July 1979.

[54] A. Gersho and B. Ramamurthi, "Image coding using vector quantization", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1982, pp. 428–431.

[55] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Massachusetts: Kluwer, 1992.

[56] V.K. Goyal, "Theoretical foundations of transform coding", *IEEE Signal Processing Mag.*, vol. 18, no. 5, Sept. 2001.

[57] V.K. Goyal, "Transform coding with backward adaptive updates", *IEEE Trans. Inform. Theory*, vol. 46, no. 4, pp. 1623–1633, July 2000,

[58] A. Graps, "An introduction to wavelets', *IEEE Computational Science and Engineering*, vol. 2, no. 2, pp. 50–61, 1995.

[59] R.M. Gray and D.L. Neuhoff, "Quantization", *IEEE Trans. Inform. Theory*, Vol. 44, No. 6, pp. 2325–2383, Oct. 1998.

[60] A. Gray and J. Markel, "Quantization and bit allocation in speech processing", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 459–473, 1976.

[61] G. Guibé, H.T. How and L. Hanzo, "Speech spectral quantizers for wideband speech coding", *European Transactions on Telecommunications*, 12(6), pp. 535–545, 2001.

[62] F.S. Gurgen, S. Sagayama, and S. Furui, "Line spectrum frequency-based distance measures for speech recognition", in *Proc. Int. Conf. Spoken Language Processing*, Kobe, Japan, Nov. 1990, pp. 521–524.

[63] W. Fisher, V. Zue, J. Bernstein, and D. Pallet, "An acoustic-phoenetic data base", *J. Acoust. Soc. Am.*, vol. 81, Suppl. 1, 1987

[64] E. Harborg, J.E. Knudsen, A. Fuldseth and F.T. Johansen, "A real-time wideband CELP coder for a videophone application", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994, pp. 121–124.

[65] M.H. Hayes, *Statistical Digital Signal Processing and Modeling*, New York: Wiley, 1996, p. 40.

[66] P. Hedelin and J. Skoglund, “Vector quantization based on Gaussian mixture models”, *IEEE Trans. Speech Audio Processing*, Vol. 8, No. 4, pp. 385–401, July 2000.

[67] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech”, *J. Acoust. Soc. Am.*, 87(4), pp. 1738–1752, Apr. 1990.

[68] M.L. Hilton, B.D. Jawerth and Ayan Sengupta, “Compressing still and moving images with wavelets’, *Multimedia Systems*, vol. 2, no. 3, 1994.

[69] H.G. Hirsch, “The influence of speech coding on recognition performance in telecommunication networks”, in *Proc. Int. Conf. Spoken Language Processing*, Denver, USA, Sept. 1998.

[70] H.G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, *ISCA ITRW ASR2000*, Paris, France, Sept. 2000.

[71] H. Hotelling, “Analysis of a complex of statistical variables into principal components”, *J. Educ. Psychology*, vol. 24, pp. 417–441, 498–520, 1933.

[72] J.J.Y. Huang and P.M. Schultheiss, “Block quantization of correlated Gaussian random variables”, *IEEE Trans. Commun. Syst.*, vol. CS-11, pp. 289–296, Sept. 1963.

[73] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, New Jersey: Prentice Hall, 2001.

[74] J.M. Huerta and R.M. Stern, “Speech recognition from GSM codec parameters”, in *Proc. Int. Conf. Spoken Language Processing*, vol. 4, 1998, pp.1463–1466.

[75] F. Itakura, “Line spectrum representation of linear predictive coefficients of speech signals”, *J. Acoust. Soc. Amer.*, vol. 57, p. S35, Apr. 1975.

[76] F. Itakura, “Minimum prediction residual principle applied to speech recognition”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 1, pp. 67–72, Feb. 1975.

[77] F. Itakura and S. Saito, “Speech analysis-synthesis based on the partial autocorrelation coefficient”, *Proc. JSA*, pp. 199–200, 1969.

[78] A.K. Jain, “Image data compression: a review”, *Proc. IEEE*, vol. 69(3), pp. 349–389, Mar. 1981.

[79] J.D. Johnston, “A filter family designed for use in quadrature mirror filters banks”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1980, pp. 291-294.

[80] B.H. Juang and A.H. Gray, Jr., “Multiple stage vector quantisation for speech coding’, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1982, pp. 597-600.

[81] B.H. Juang, L.R. Rabiner and J.G. Wilpon, “On the use of bandpass liftering in speech recognition”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, pp. 947–954, July 1987.

[82] H. Karhunen, “Uber lineare methoden in der wahrscheinlichkeitsrechnung”, *Ann. Acad. Sci. Fenn.*, Ser. A.I. 34, Helsinki, 1947.

[83] H.K. Kim and R.V. Cox, “A bitstream-based front-end for wireless speech recognition on IS-136 communications system”, *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 558–568, July 2001.

[84] H. Kiya, K. Nishikawa, and M. Sagawa, “Property of circular convolution for subband image coding”, *IEICE Trans. Fundamentals*, vol. E75-A, no. 7, pp. 852–860, July 1992.

[85] I. Kiss, “A comparison of distributed and network speech recognition for mobile communication systems”, in *Proc. Int. Conf. Spoken Language Processing*, 2000.

[86] I. Kiss and P. Kapanen, “Robust feature vector compression algorithm for distributed speech recognition”, in *Proc. Eurospeech*, 1999, pp. 2183–2186.

[87] N.P. Koestoer, “Robust linear prediction analysis for speech coding”, PhD dissertation, Griffith University, 2002.

[88] B. Kolman, *Introductory Linear Algebra with Applications*, New York: Macmillan, 1990, pp. 439–444.

[89] K.P. Kramer and M.V. Mathews, “A linear coding for transmitting a set of correlated signals”, *IRE Trans. Inform. Theory* (Corresp), vol. IT-17, pp. 751–752, Nov. 1971.

[90] V. Krishnan, D.V. Anderson and K.K. Truong, "Optimal multistage vector quantization of LPC parameters over noisy channels", *IEEE Trans. Speech Audio Processing*, Vol. 12, No. 1, pp. 1–8, Jan. 2004.

[91] P. Kroon and W.B. Kleijn, "Linear-prediction based analysis-by-synthesis coding" in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Ed. Amsterdam: Elsevier, 1995, pp. 79–119.

[92] C. Laflamme, J.-P. Adoul, R. Salami, S. Morissette, and P. Mabilleau, "16 kbps wideband speech coding technique based on algebraic CELP", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 13–16.

[93] W.P. LeBlanc, B. Bhattacharya, S.A. Mahmoud, and V. Cuperman, "Efficient search and design procedures for robust multi-stage VQ for LPC parameters for 4 kb/s speech coding", *IEEE Trans. Speech Audio Processing*, vol. 1, no. 4, pp. 373–385, Oct. 1993.

[94] R. Lefebvre, R. Salami, C. Laflamme, and J.-P. Adoul, "8 kbit/s coding of speech with 6 ms frame-length", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993, pp. II-612–II-615.

[95] R. Lefebvre, R. Salami, C. Laflamme, and J.-P. Adoul, "High quality coding of wideband audio signals using transform coded excitation (TCX)", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994, pp. I-193–I-196.

[96] R.G. Leonard, "A database for speaker-independent digit recognition", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1984, vol. 3, pp. 328–331.

[97] J. Le Roux and C. Gueguen, "A fixed point computation of partial correlation coefficients", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 3, pp. 257–259, June 1977.

[98] A.S. Lewis and G. Knowles, "Image compression using the 2-D wavelet transform", *IEEE Trans. Image Processing*, vol. 1, no. 2, pp. 244–250, Apr. 1992.

[99] B.T. Lilly and K.K. Paliwal, "Effect of speech coders on speech recognition performance", in *Proc. Int. Conf. Spoken Language Processing*, 1996, vol. 4, pp. 2344–2347.

[100] Y. Linde, A. Buzo, and R.M. Gray, “An algorithm for vector quantizer design”, *IEEE Trans. Commun.*, vol. COM-28, no. 1, pp. 84–95, Jan. 1980.

[101] S.P. Lloyd, “Least square quantization in PCM”, *IEEE Trans. Inform. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[102] M. Loéve, “Fontions aléatoires de seconde ordre”, in P. Levý, *Processus Stochastiques et Mouvement Brownien*, Paris, France: Hermann, 1948.

[103] T.D. Lookabaugh and R.M. Gray, “High-resolution quantization theory and the vector quantizer advantage”, *IEEE Trans. Inform. Theory*, vol. 35, no. 5, pp. 1020–1033, Sept 1989.

[104] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, *Proc. of the Fifth Berkeley Symposium on Math., Stat., and Prob.*, vol. 1, 1967, pp. 281–296.

[105] P.C. Mahalanobis, “On the generalized distance in statistics”, in *Proc. Indian Nat. Inst. Sci.* (Calcutta), 1936, vol. 2, pp. 49–55.

[106] S.A. Martucci, “Signal extension and noncausal filtering for subband coding of images”, *SPIE: Visual Communications and Image Processing*, vol. 1605, pp.137–148, 1991.

[107] J. Makhoul, “Linear prediction: a tutorial review”, *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[108] J. Makhoul, S. Roucos, and H. Gish, “Vector quantization in speech coding”, *Proc. IEEE*, vol. 73, pp. 1551–1588, Nov. 1985.

[109] S. Mallat, “A theory of multiresolution signal decomposition: the wavelet representation”, *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 11, no. 7, pp. 674–693, July 1989.

[110] H.S. Malvar and D.H. Staelin, “The LOT: transform coding without blocking effects”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37(4), pp. 553–559, Apr. 1989.

[111] J. Max, “Quantising for minimum distortion”, *IRE Trans. Inform. Theory*, vol. IT-6, pp. 7–12, Mar. 1960.

[112] A. Mertins, *Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications*, John Wiley and Sons, 1999.

[113] R. Meston, "GSM vocoders improve speech transmission", 2004. Available: http://www.eetasia.com/ARTICLES/2004NOV/B/2004NOV01_MSD_PD_TA.pdf

[114] W.B. Mikhael and V. Krishnan, "Energy-based split vector quantizer employing signal representation in multiple transform domains", *Digital Signal Processing*, vol. 11, no. 4, pp. 359-370, Oct. 2001.

[115] J.R. Movellan, "Tutorial on principal component analysis", Technical Report, University of California, San Diego, 2003. Available: http://mplab.ucsd.edu/tutorials/pdfs/PCA.pdf

[116] A.N. Netravali and J.O. Limb, "Picture coding: A review", *Proc. IEEE*, vol. 68, no. 3, pp. 366–406, Mar. 1980.

[117] P. Noll, "Digital audio coding for visual communications", *Proc. IEEE*, vol. 83, no. 6, pp. 925–943, June 1995.

[118] A. Ortega and M. Vetterli, "Adaptive scalar quantization without side information", *IEEE Trans. Image Proc.*, vol. 6, pp. 665–676, May 1997.

[119] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression', *IEEE Signal Processing Magazine*, vol. 15, no. 6, Nov 1998.

[120] M.D. Paez and T.H. Glisson, "Minimum mean-squared-error quantization in speech PCM and DPCM systems", *IEEE Trans. Commun.*, vol. COM-20, pp. 225–230, Apr. 1972.

[121] E. Paksoy, K. Srinivasan, and A. Gersho, "Variable rate speech coding with phonetic segmentation", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1993, pp. 155–158.

[122] K.K. Paliwal and B.S. Atal, "Efficient vector quantisation of LPC parameters at 24 bits/frame", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 661–664.

[123] K.K. Paliwal and B.S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame", *IEEE Trans. Speech Audio Processing*, vol. 1, no. 1, pp. 3–14, Jan. 1993.

[124] K.K. Paliwal and W.B. Kleijn, "An introduction to speech coding" in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Ed. Amsterdam: Elsevier, 1995, pp. 1–47.

[125] K.K. Paliwal and W.B. Kleijn, "Quantization of LPC parameters" in *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal, Ed. Amsterdam: Elsevier, 1995, pp. 443–466.

[126] K.K. Paliwal and S. So, "Low complexity GMM-based block quantisation of images using the discrete cosine transform", accepted by ICIP 2003.

[127] K.K. Paliwal and S. So, "Low complexity Gaussian mixture model-based block quantisation of images", in *Proc. Microelectronic Engineering Research Conference*, Brisbane, Australia, Nov. 2003.

[128] K.K. Paliwal and S. So, "A fractional bit encoding technique for the GMM-based block quantisation of images", *Digital Signal Processing*, vol. 15, pp. 255–275, May 2005.

[129] K.K. Paliwal and S. So, "Low complexity GMM-based block quantisation of images using the discrete cosine transform", *Signal Processing: Image Communication*, vol. 20, pp. 435–446, June 2005.

[130] K.K. Paliwal and S. So, "Multiple frame block quantisation of line spectral frequencies using Gaussian mixture models", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Montreal, Canada, 2004, pp. I-149–152.

[131] K.K. Paliwal and S. So, "Scalable distributed speech recognition using multi-frame GMM-based block quantization", in *Proc. Int. Conf. Spoken Language Processing*, Jeju, Korea, Oct. 2004.

[132] J. Pan and T.R. Fischer, "Vector quantization-lattice vector quantization of speech LPC coefficients", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, 1994, pp. 513–516.

[133] J. Pan, “Two-stage vector quantization-pyramidal lattice vector quantization and application to speech LSP coding”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 1996, pp. 737–740.

[134] J.W. Paulus and J. Schnitzler, “16 kbit/s wideband speech coding based on unequal subbands”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1996, pp. 255–258.

[135] L.C.W. Pols, “Spectral analysis and identification of Dutch vowels in monosyllabic words”, Doctoral dissertation, Free University, Amsterdam, The Netherlands, 1966.

[136] J.G. Proakis and D.G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, 3rd Edition, New Jersey: Prentice-Hall, 1996.

[137] S. Quackenbush, “A 7 kHz bandwidth, 32 kbps speech coder for ISDN”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 1–4.

[138] L.R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition”, *Proc. IEEE*, vol. 77, no. 2, pp. 257–286.

[139] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, New Jersey: Prentice Hall, 1993.

[140] B. Raj, J. Migdal and R. Singh, “Distributed speech recognition with codec parameters”, in *Proc. ASRU*, Trento, Italy, Dec. 2001.

[141] B. Ramamurthi and A. Gersho, “Classified vector quantization of images”, *IEEE Trans. Commun.*, vol. COM-34, pp. 1105–1115, Nov. 1986.

[142] T.A. Ramstad, S.O. Aase and J.H. Husøy, *Subband Compression of Images: Principles and Examples*, vol. 6, Amsterdam: Elsevier Science B.V., 1995.

[143] K.R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, and Applications*, Academic Press, 1990.

[144] G.N. Ramaswamy and P.S. Gopalakrishnan, “Compression of acoustic features for speech recognition in network environments”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1998, pp. 977–980.

[145] H.C. Reeve, III, and J.S. Lim, “Reduction of blocking effect in image coding”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1983, pp. 1212–1215.

[146] D.A. Reynolds and R.C. Rose, “Robust text-independent speaker identification using Gaussian mixture models”, *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[147] E.A. Riskin, “Optimal bit allocation via the generalized BFOS algorithm”, *IEEE Trans. Inform. Theory*, 37(2), pp. 400–402, 1991.

[148] G. Roy and P Kabal, “Wideband CELP speech coding at 16 kbits/sec”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 17–20.

[149] M.J. Sabin and R.M. Gray, “Product code vector quantizers for waveform and voice coding”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, no. 3, pp. 474–488, June 1984.

[150] A. Said and W.A. Pearlman, ‘A new fast and efficient image codec based on set partitioning in hierarchical trees’, *IEEE Trans. Circuits and Systems for Video Technology*, vol. 6, June 1996.

[151] C. Sanderson, “Automatic person verification using speech and face information”, PhD dissertation, Griffith University, 2002.

[152] K. Sayood, *Introduction to Data Compression*, San Francisco: Morgan Kaufmann Publishers, 1996.

[153] M.R. Schroeder and B.S. Atal, “Code-excited linear prediction (CELP): high quality speech at very low bit rates”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1985, pp. 937–940.

[154] A. Segall, “Bit allocation and encoding for vector sources”, *IEEE Trans. Inform. Theory*, vol. IT-22, no. 2, pp. 162–169, Mar. 1976.

[155] C.E. Shannon, “A mathematical theory of communication”, *Bell Sys. Tech. J*, vol. 27, pp. 379-423, 625-656, 1948.

[156] C.E. Shannon, “Coding theorems for a discrete source with a fidelity criterion”, *IRE National Convention Record*, part 4, pp. 142-163, 1959.

[157] B.J. Shannon and K.K. Paliwal, “A comparative study of filter bank spacking for speech recognition”, in *Proc. Microelectronic Engineering Research Conference*, Brisbane, Australia, Nov. 2003.

[158] J.M. Shapiro, “Embedded image coding using zerotrees of wavelet coefficients”, *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.

[159] J.M. Shapiro, “An embedded wavelet hierarchical image coder’, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, March 1992.

[160] Y. Shin, S. Kang, T.R. Fischer, C. Son, and Y. Lee, “Low-complexity predictive trellis coded quantization of wideband speech LSF parameters”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2004, pp. 145-148.

[161] J. Shlens, “A tutorial on principal component analysis”, Technical Report, University of California, San Diego, 2003. Available: http://www.snl.salk.edu/ shlens/pub/notes/pca.pdf

[162] Y. Shoham, “Low-delay code-excited linear predictive coding of wideband speech at 32 kbps”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1991, pp. 9–12.

[163] U. Sinervo, J. Nurminen, A. Heikkinen, and J. Saarinen, “Evaluation of split and multistage techniques in LSF quantization”, in *Proc. Norsig 2001*, Trondheim, Norway, Oct. 2001, pp. 18–22.

[164] M.J.T. Smith and S.L. Eddins, “Analysis/synthesis techniques for subband image coding”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 8, pp. 1446–1456, Aug. 1990.

[165] S. So and K.K. Paliwal, “Efficient block coding of images using Gaussian mixture models”, in *Proc. Fourth Australasian Workshop on Signal Processing and Applications 2002*, Brisbane, Australia, Sept. 2002, pp. 71–74.

[166] S. So and K.K. Paliwal, “Efficient vector quantisation of line spectral frequencies using the switched split vector quantiser”, in *Proc. Int. Conf. Spoken Language Processing*, Jeju, Korea, Oct. 2004.

[167] S. So and K.K. Paliwal, “Multi-frame GMM-based block quantisation of line spectral frequencies”, to appear in *Speech Commun.*, 2005.

[168] S. So and K.K. Paliwal, “Multi-frame GMM-based block quantisation of line spectral frequencies for wideband speech coding”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. I, Philadelphia, USA, 2005, pp. 121–124.

[169] S. So and K.K. Paliwal, "Efficient product code vector quantisation using the switched split vector quantiser", submitted to *Digital Signal Processing*, Nov. 2004.

[170] S. So and K.K. Paliwal, "A comparative study of LPC parameter representations and quantisation schemes in wideband speech coding", submitted to *Digital Signal Processing*, May 2005.

[171] S. So and K.K. Paliwal, "Comparison of LPC parameter representations for wideband speech coding", to be submitted to *IEEE Signal Processing Lett.*, 2005.

[172] S. So and K.K. Paliwal, "Switched split vector quantisation of line spectral frequencies for wideband speech coding", to appear in *Proc. European Conf. Speech Communication and Technology* (Eurospeech), Lisbon, Portugal, Sept 2005.

[173] S. So and K.K. Paliwal, "A comparison of LSF and ISP representations for wideband LPC parameter coding using the switched split vector quantiser", to appear in *Proc. IEEE Int. Symp. Signal Processing and Applications* (ISSPA), Sydney, Australia, Aug 2005.

[174] S. So and K.K. Paliwal, "Scalable distributed speech recognition using Gaussian mixture model-based block quantisation", submitted to *Speech Commun.*, 2005.

[175] F.K. Soong and B.H. Juang, "Line spectrum pair (LSP) and speech data compression", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, San Diego, California, Mar 1984, pp. 37–40.

[176] F.K. Soong and B.H. Juang, "Optimal quantization of LSP parameters", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, New York, pp. 394–397, Apr. 1988.

[177] T. Sporer, K. Brandenburg and B. Edler, "The use of multirate filterbanks for coding of high quality digital audio", in *Proc. EUSIPCO*, vol. 1, 1992, pp. 211–214.

[178] N. Srinivasamurthy, A. Ortega and S. Narayanan, "Efficient scalable encoding for distributed speech recognition", submitted to *IEEE Trans. Speech and Audio Processing*, 2003. Available: http://biron.usc.edu/˜snaveen/papers/Scalable_DSR.pdf

[179] S.S. Stevens and J. Volkman, "The relation of pitch to frequency", *Journal of Psychology*, 53, pp. 329, 1940

[180] B. Strope and A. Alwan, “A model of dynamic auditory perception and its application to robust word recognition”, *IEEE Trans. Speech Audio Processing*, vol. 5, no. 2, pp. 451–464, Sept. 1997.

[181] J.K. Su and R.M. Mersereau, “Coding using Gaussian mixture and generalized Gaussian models”, in *Proc. IEEE Int. Conf. Image Processing*, Lausanne, Switzerland, 1996, pp. 217–220.

[182] A.D. Subramaniam and B.D. Rao, “PDF optimized parametric vector quantization with applications to speech coding”, *34th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, Nov. 2000.

[183] A.D. Subramaniam and B.D. Rao, “PDF optimized parametric vector quantization of speech line spectral frequencies”, *IEEE Trans. Speech Audio Processing*, vol. 11, no. 2, pp. 130–142, Mar. 2003.

[184] A.D. Subramaniam, “Gaussian mixture models in compression and communication”, PhD dissertation, University of California, San Diego, CA, 2003.

[185] N. Sugamura and F. Itakura, “Speech analysis and synthesis methods developed at ECL in NTT–from LPC to LSP–”, *Speech Commun.*, vol. 5, pp. 199–215, Jun. 1986.

[186] R.J. Tocci, *Digital Systems: Principles and Applications*, 6th Edition, Prentice Hall, New Jersey, pp. 6–9, 1995.

[187] C. Tsao and R.M. Gray, “Matrix quantizer design for LPC speech using the generalized Lloyd algorithm”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 537–545, June 1985.

[188] A. Tucker, *A Unified Introduction to Linear Algebra: Models, Methods, and Theory*, New York: Maxwell-Macmillan, 1989, pp. 502–504.

[189] J. Turunen and D. Vlaj, “A study of speech coding parameters in speech recognition”, in *Proc. Eurospeech*, 2001, pp. 2363–2366.

[190] A. Ubale and A. Gersho, “A multi-band CELP wideband speech coder”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994, pp. 1367-1370.

[191] C. Valens, “A really friendly guide to wavelets”, Technical Report, 1999. Available: http://perso.wanadoo.fr/polyvalens/clemens/download/arfgtw.pdf

[192] M. Vetterli, “Multi-dimensional subband coding: some theory and algorithms”, *Signal Processing*, vol. 6, pp. 97-112, Apr. 1984.

[193] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*, Prentice-Hall, New Jersey, 1995.

[194] M. Vetterli, “On Fourier and wavelets: representation, approximation, and compression”, presented at Wavelet and Multifractal Analysis 2004, Cargese, Corsica, July 2004. Available: http://www.inrialpes.fr/is2/people/pgoncalv/WAMA2004/lectures/Vetterli-lecture.pdf.

[195] R. Viswanathan and J. Makhoul, “Quantization properties of transmission parameters in linear predictive systems”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 309–321, 1975.

[196] G.K. Wallace, “The JPEG still picture compression standard”, *Communications of the ACM*, vol. 34, no. 4, pp. 30-44, Apr. 1991.

[197] S. Wang and A. Gersho, “Phonetically-based vector excitation coding of speech at 3.6 kbit/s”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, May, 1989, pp. I-349–352.

[198] P.A. Wintz, “Transform picture coding”, *Proc. IEEE*, vol. 60(7), pp. 809–820, July 1972.

[199] R.C. Wood, “On optimum quantization”, *IEEE Trans. Inform. Theory*, vol. IT-15, no. 2, pp. 248–252, Mar. 1969.

[200] J.W. Woods and S.D. O’Neil, “Sub-band coding of images”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1986, pp. 1005–1008.

[201] J.W. Woods and S.D. O’Neil, “Subband coding of images”, *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-34(5), pp. 1278-1288, Oct. 1986.

[202] T. Wuppermann and F. de Bont, “Feasibility study of 32 kb/s wideband speech and music coding with a low-delay filterbank”, in *IEEE Workshop on Speech Coding for Telecommunications*, pp. 11–12, 1993.

[203] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2.1)*, Cambridge University Engineering Department, 2002.

[204] Y. Zhang and C.J.S. deSilva, "An isolated word recognizer using the EM algorithm for vector quantization", *IREECON 1991*, Sydney, Australia, pp. 289–292.

[205] J. Zhou, Y. Shoham and A. Akansu, "Simple fast vector quantization of the line spectral frequencies", in *Proc. Int. Conf. Spoken Language Processing*, Vol. 2, 1996, pp. 945–948.

[206] Q. Zhu and A. Alwan, "An efficient and scalable 2D DCT-based feature coding scheme for remote speech recognition", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Aug 2001, pp. 113-116.

[207] Q. Zhu and A. Alwan, "On the use of variable frame rate analysis in speech recognition", in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 2000, vol. 3, pp. 1783–1786.