Scalable distributed speech recognition using Gaussian mixture model-based block quantisation

Stephen So , Kuldip K. Paliwal

Reporter:黃重翔
Professor:陳嘉平

Abstract

- In this paper, we investigate the use of block quantisers based on Gaussian mixture models (GMMs) for the coding of Mel frequency-warped cepstral coefficient (MFCC) features in distributed speech recognition (DSR) applications.
- Specifically, we consider the multiframe scheme, where temporal correlation across MFCC frames is exploited by the Karhunen–Loe `ve transform of the block quantiser.

Abstract

- Compared with vector quantisers, the GMM-based block quantiser has relatively low computational and memory requirements which are independent of bitrate.
- More importantly, it is bitrate scalable, which means that the bitrate can be adjusted without the need for re-training.

Abstract

- Static parameters such as the GMM and transform matrices are stored at the encoder and decoder and bit allocations are calculated on-the-fly without intensive processing.
- We have evaluated the quantisation scheme on the Aurora-2 database in a DSR framework.
- We show that jointly quantising more frames and using more mixture components in the GMM leads to higher recognition performance.

Abstract

- The multi-frame GMM-based block quantiser achieves a word error rate (WER) of 2.5% at 800 bps, which is less than 1% degradation from the baseline (unquantised) word recognition accuracy, and graceful degradation down to a WER of 7% at 300 bps.

- Keywords: Distributed speech recognition; Gaussian mixture models; Block quantisation; Aurora-2

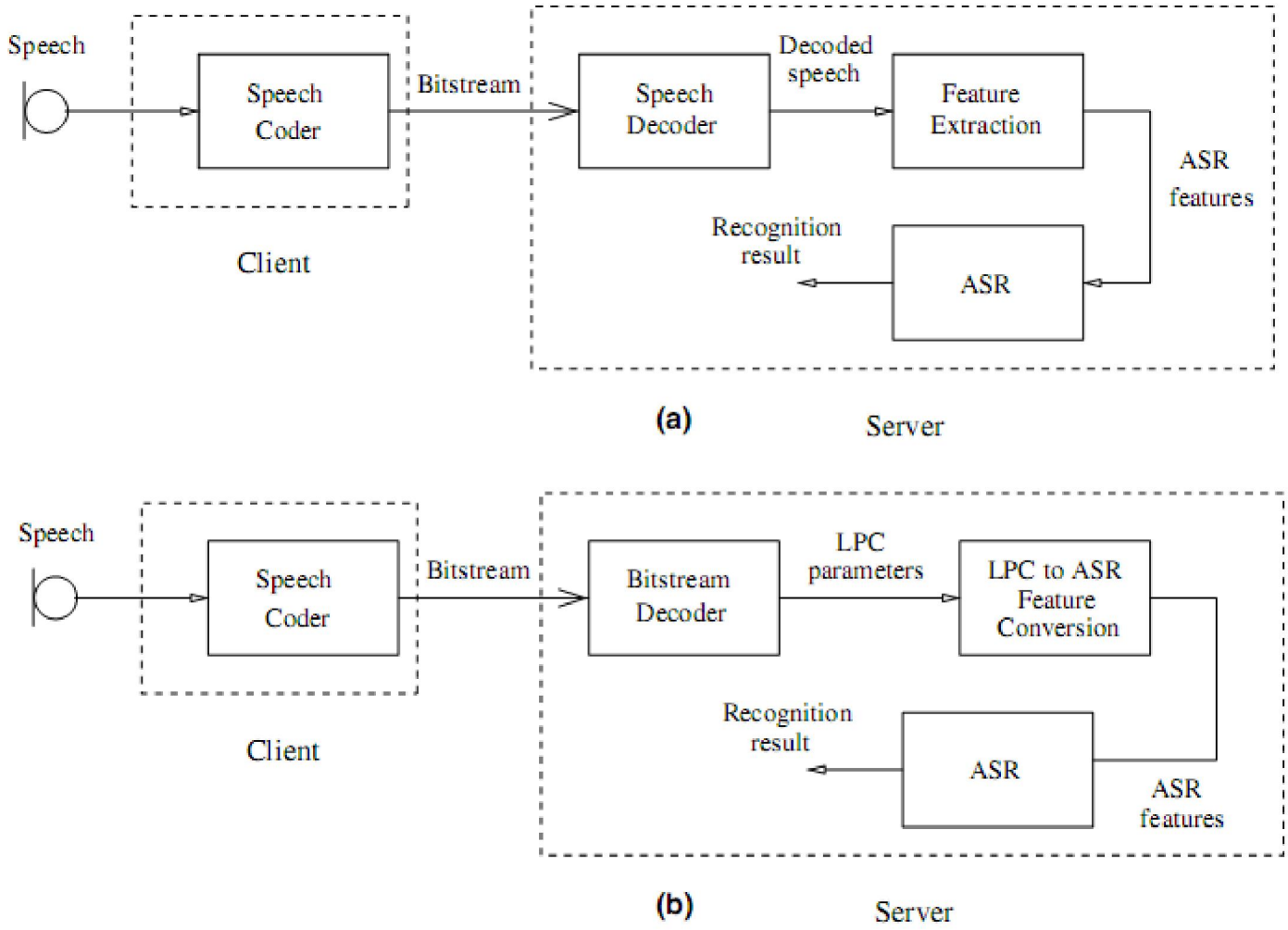
Outline

- Introduction
- Multi-frame GMM-based block Quantization
- Experimental Setup
- Results and Discussion
- Conclusion and Further Work

Introduction

- 隨著遠端與無線裝置如PDA與行動手機的增加，在行動通訊系統的內容中加入自動語音辨識（Automatic speech recognition, ASR）技術開始受到注意。
- 行動裝置大致上都在儲存量與程序執行能力受限，而這造成機上ASR系統並不實際。解決方法爲在遠端伺服器進行複雜的語音辨識任務，在網路上是可行的。

Introduction



Introduction

- Network speech recognition (NSR)中，使用者語音被常規語音編碼器壓縮並傳送至進行辨識任務的伺服器。
- 在位元串爲主的NSR中，伺服器使用直接從位元串取得的線性預測編碼（Linear predictive coding, LPC）參數獲得ASR特徵。

Introduction

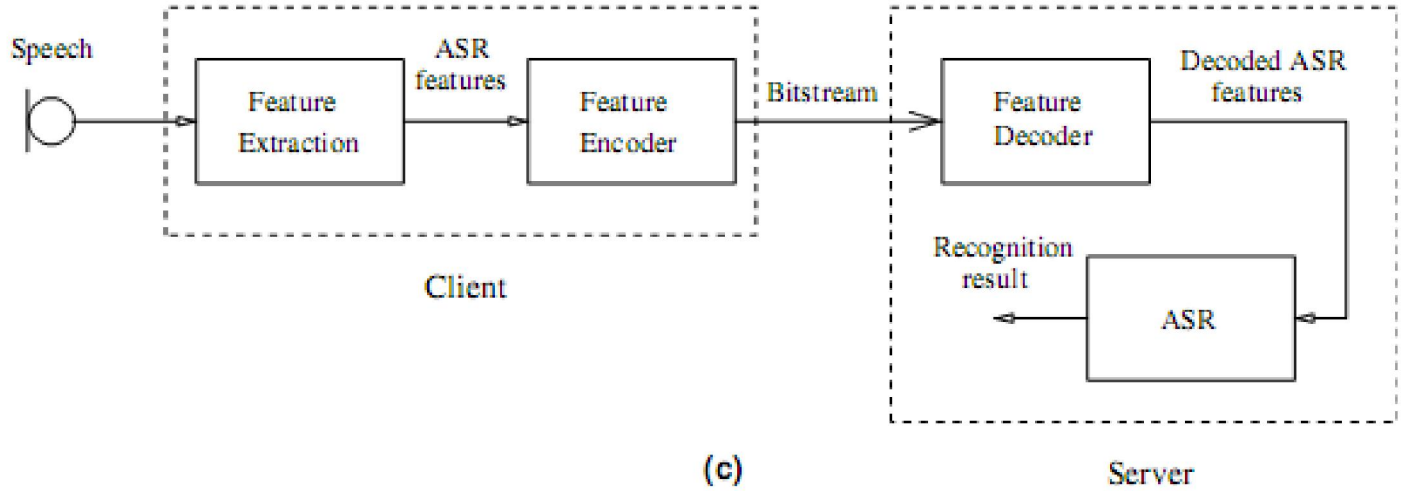


Fig. 1. Client–server-based speech recognition modes: (a) speech-based network speech recognition (NSR), (b) bitstream-based network speech recognition, and (c) distributed speech recognition (DSR).

Introduction

- 分佈式語音辨識 (Network speech recognition, DSR)中，ASR系統分散於客戶端與伺服端。此處語音的特徵萃取是在客戶端進行，而ASR特徵經過壓縮並且經由一專用頻道傳送至伺服器，這個專用頻道進行ASR特徵的解碼並輸入ASR後端。

Introduction

- 雖然向量量化器通常因較少位元能提供較好的辨識效能，然而向量量化器是設計只用於具體的位元率，在其他位元率必須重新訓練。
- 新的量化方法有以下三個優點
  - 簡潔表示和位元率非相關的機率密度函數（Probability density functions PDF）
  - 具立即性位元分配的位元率可調整性
  - 低搜尋複雜度與和系統速率非相關的記憶體需求

Multi-frame GMM-based block Quantization

- 以GMM參數性模組化來源的機率密度函數且接著對每個Gaussian mixture元件設計區塊量化器。
- 我們設計一修改過的框架，使用連接了p個向量
- 多框架GMM區塊量化器可以分爲三個層次：PDF估算，位元分配與最小失真區塊量化。

Multi-frame GMM-based block Quantization

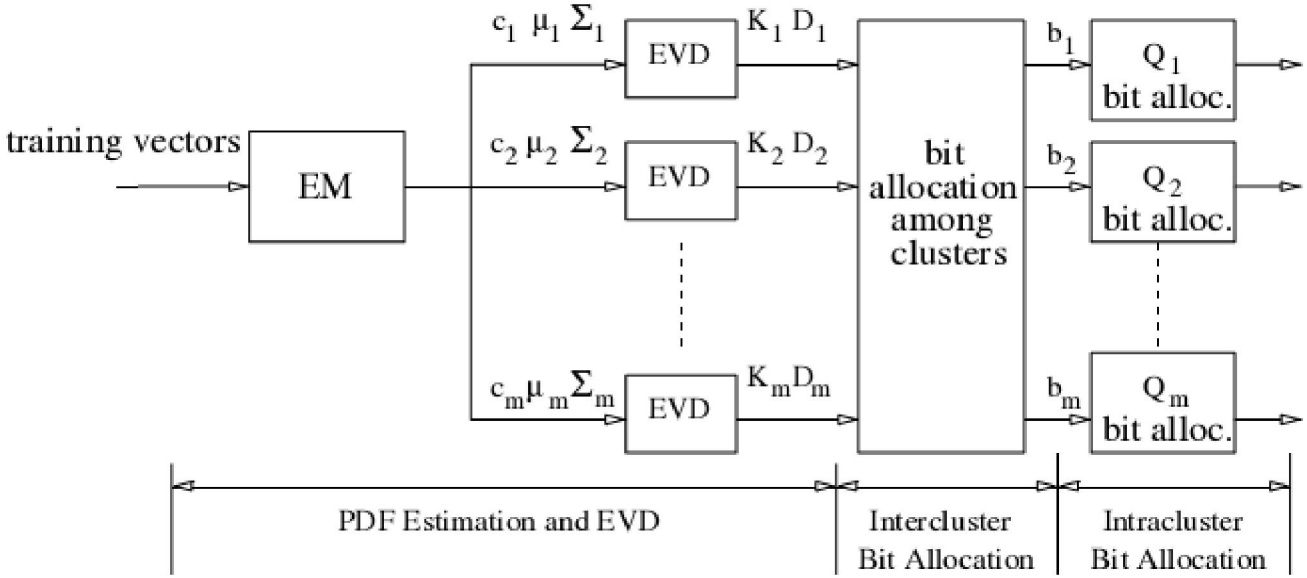


Fig. 2. PDF estimation and bit allocation from training data.

Multi-frame GMM-based block Quantization

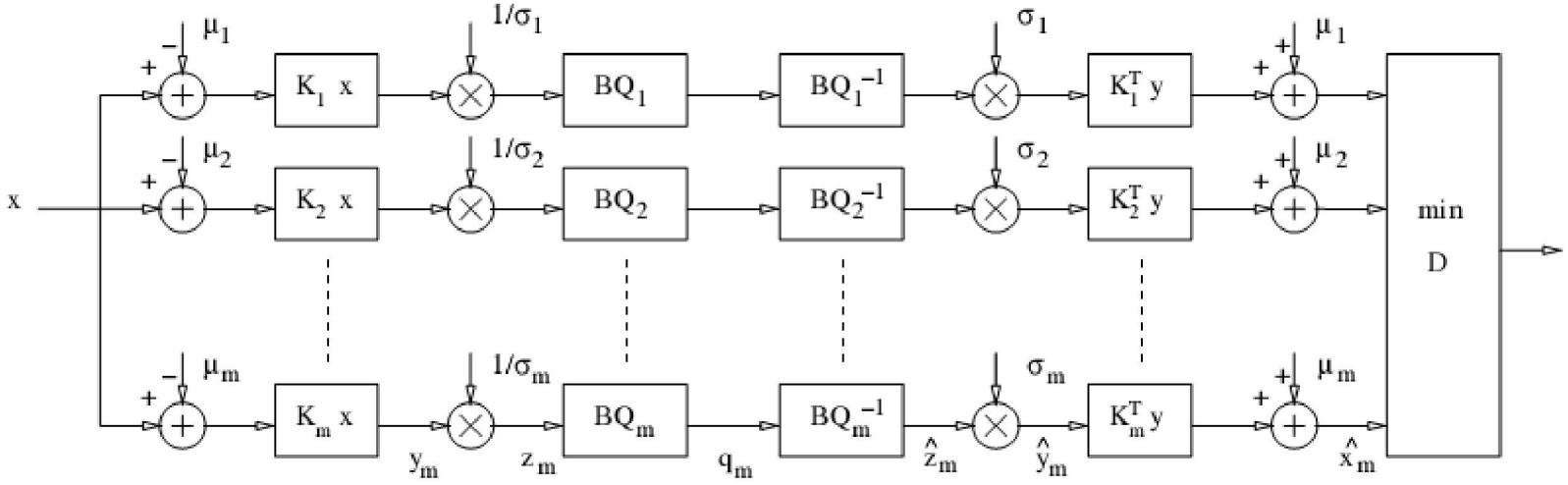


Fig. 3. Block diagram of the GMM-based block quantiser (BQ—block quantiser).

PDF Estimation Using GMMs

- PDF模型與Karhunen-Loeve transform（KLT）正交矩陣爲GMM區塊量化器的唯二穩定且位元率非相關的參數。
- PDF模型以混合的多變元高斯參數表示

$$G(\boldsymbol{x}|\boldsymbol{M}) = \sum_{i=1}^{m} c_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$\boldsymbol{M} = [m, c_i, \ldots, c_m, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m]$$

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

PDF Estimation Using GMMs

- 在m個共變異矩陣計算出特徵值分解（EVD），產生m個特徵值向量，單一框架共有nm個特徵值，而多重框架共有npm個特徵值。

$$\{\lambda_{i,j};\ i=1,\ldots,n;\ j=1,\ldots,m\}$$

$$\{\lambda_{i,j};\quad i=1,\ldots,np;\quad j=1,\ldots,m\}$$

Bit Allocation

- 量化每一向量共有 b_{tot} 個位元，分配到對每個混合元件的各個量化器。
- 量化器被修改的層級總數爲

$$2^{b_{\text{tot}}} = \sum_{i=1}^{m} 2^{b_i}$$

- 平均失真可用以下方法近似

$$D_{\text{tot}} = \sum_{i=1}^{m} c_i D_i(b_i)$$

Bit Allocation

- 用高解析近似法計算對一區塊量化器的總失真爲

$$D_i(b_i) = Knp\Lambda_i 2^{-2\frac{bi}{np}}$$

$$\Lambda_i = \left[\prod_{j=1}^{np} \lambda_{i,j}\right]^{\frac{1}{np}} \quad \text{for } i = 1, 2, \ldots, m$$

- 最小化失真而得到的位元分配算式

$$2^{b_i} = 2^{b_{\text{tot}}} \frac{(c_i\Lambda_i)^{\frac{np}{np+2}}}{\sum_{i=1}^{m}(c_i\Lambda_i)^{\frac{np}{np+2}}} \quad \text{for } i = 1, 2, \ldots, m$$

Bit Allocation Within Mixture Components

- 對所有混合元件分配位元後，進行對n個元件的位元分配，其中被修正的位元總數爲

$$b_i = \sum_{j=1}^{np} b_{i,j} \quad \text{for } i = 1, 2, \ldots, m$$

- 對混合元件i的平均失真：

$$D_i = \frac{1}{np} \sum_{j=1}^{np} \lambda_{i,j} K 2^{-2b_i,j} \quad \text{for } i = 1, 2, \ldots, m$$

- 接著可得位元分配公式

$$b_{i,j} = \frac{b_i}{np} + \frac{1}{2} \log_2 \frac{\lambda_{i,j}}{[\prod_{j=1}^{np} \lambda_{i,j}]^{\frac{1}{np}}} \quad \text{for } i = 1, 2, \ldots, m$$

Minimum Distortion Block Quantisation

- 爲了進行對向量x的量化，首先取得一般化變異數向量

$$\boldsymbol{z}_i = \frac{\boldsymbol{K}_i(\boldsymbol{x} - \boldsymbol{\mu}_i)}{\boldsymbol{\sigma}_i}$$

- 接著可以求得再構成向量

$$\hat{\boldsymbol{x}}_i = \boldsymbol{K}_i^{\mathrm{T}} \boldsymbol{\sigma}_i \hat{\boldsymbol{z}}_i + \boldsymbol{\mu}_i$$

- 最後得到最小化失真之下的混合元件k

$$k = \underset{i}{\operatorname{argmin}}\, d(\boldsymbol{x} - \hat{\boldsymbol{x}}_i)$$

- 這裡爲了編碼MFCC向量而使用均方誤差（MSE）作爲選擇適當的區塊量化器的失真採樣。

Quantiser Index Encoding

- 每一量化向量都由一數辨認哪一個混合元件被用於編碼，總共有 $b - \log_2 m$ 位元可用來量化每個相當於 $2^b/m$ 個量化器層級的向量。因此量化器的指數可以被分作m段。

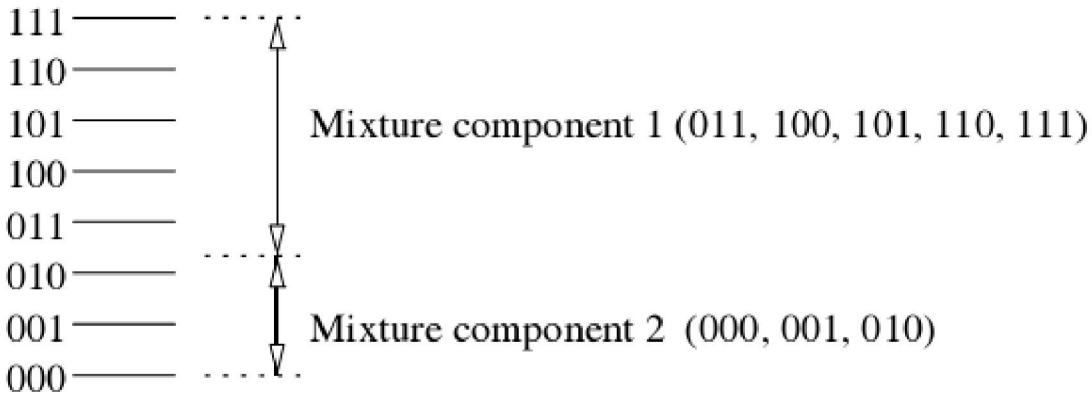


Fig. 4. Example of quantiser level encoding and mixture component number partitioning.

Bitrate Scalability, Computational Complexity and Memory Requirements

Table 1
Bitrate independent computational complexity (in kflop/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 mixture components, m

m	P	kflop/frame	ROM (floats)
16	1	13.46	3136
	2	22.66	10,624
	3	31.88	22,720
	4	41.09	39,424
	5	50.53	60,736
32	1	26.91	4416
	2	45.33	14,976
	3	63.75	31,936
	4	82.18	55,296
	5	100.6	121,216

Experimental Setup

- 實驗中使用ETSI Aurora-2 語料庫在HTK3.2上估計出對數種量化方法的辨識效能。只在乾淨語音中比較辨識效能與位元率。
- 爲減少位元分配的變異使用視窗函數w(n)

$$w(n) = 1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right) \quad \text{where } n = 1, 2, \ldots, L$$

Recognition Performance of the Single Frame GMM-based Block Quantiser

Table 2
Average word recognition accuracy as a function of bitrate and number of mixture components (MC) for single frame GMM-based block quantiser (baseline accuracy = 98.0%)

Bitrate (kbps)	Recognition accuracy (in %)				
	2 MC	4 MC	8 MC	16 MC	32 MC
0.3	23.5	20.0	16.7	8.1	8.1
0.4	43.5	53.3	57.7	23.3	9.1
0.6	68.7	79.7	85.7	87.6	82.0
0.8	86.2	90.3	91.5	93.7	94.5
1.0	90.5	94.2	95.0	95.5	96.1
1.2	93.9	95.9	95.9	96.4	96.7
1.5	96.0	96.5	97.0	97.2	97.2
1.7	97.0	97.0	97.2	97.3	97.4
2.0	97.3	97.2	97.5	97.6	97.7
2.2	97.6	97.3	97.6	97.7	97.7
2.4	97.6	97.5	97.7	97.9	97.7
3.0	97.9	97.8	97.9	97.8	97.9
4.4	98.0	98.0	98.1	98.0	98.0

Recognition Performance of the Multi-frame GMM-based Block Quantiser

Table 3
Average word recognition accuracy as a function of bitrate and number of frames for 16 mixture component multi-frame GMM-based block quantiser (baseline accuracy = 98.0%)

Bitrate (kbps)	Recognition accuracy (in %)			
	2 Frames	3 Frames	4 Frames	5 Frames
0.3	78.3	89.6	91.3	93.0
0.4	91.1	94.3	95.1	95.4
0.6	95.5	96.6	97.1	96.8
0.8	96.9	97.3	97.4	97.5
1.0	97.4	97.6	97.7	97.7
1.2	97.6	97.7	97.8	97.9
1.5	97.8	97.8	97.9	97.8
1.7	97.8	98.0	98.0	98.0
2.0	98.0	97.9	98.1	98.0
2.2	98.0	98.0	97.9	98.0

Recognition Performance of the Multi-frame GMM-based Block Quantiser

Table 4
Average word recognition accuracy as a function of bitrate and number of mixture components (MC) for 5 frame multi-frame GMM-based block quantiser (baseline accuracy = 98.0%)

Bitrate (kbps)	Recognition accuracy (in %)	
	16 MC	32 MC
0.2	83.0	87.7
0.3	93.0	94.2
0.4	95.6	96.0
0.6	96.8	97.1
0.8	97.5	97.6
1.0	97.7	97.6
1.2	97.9	97.9
1.5	97.8	98.0
1.7	98.0	98.0
2.0	98.0	98.0

Comparison with the Recognition Performance of the Non-uniform Scalar Quantiser

Table 5
Average word recognition accuracy as a function of bitrate for non-uniform scalar quantiser (baseline accuracy = 98.0%)

Bitrate (kbps)	Recognition accuracy (in %)
0.6	38.2
0.8	72.3
1.0	86.7
1.2	93.3
1.5	95.5
1.7	96.2
2.0	97.0
2.2	97.2
2.4	97.4
3.0	97.8
4.4	98.0

Comparison with the Recognition Performance

Table 6
Average word recognition accuracy, computational complexity (in kflop/frame), and memory requirements (ROM) as a function of bitrate for vector quantiser (baseline accuracy = 98.0%)

Bitrate (kbps)	Recognition accuracy (in %)	kflop/frame	ROM (in floats)
0.4	77.0	0.77	192
0.6	92.0	3.07	768
0.8	95.7	12.29	3072
1.0	96.9	49.51	12,288
1.2	97.0	196.7	49,152

Summary of the Recognition Performance of All Quantisation Schemes

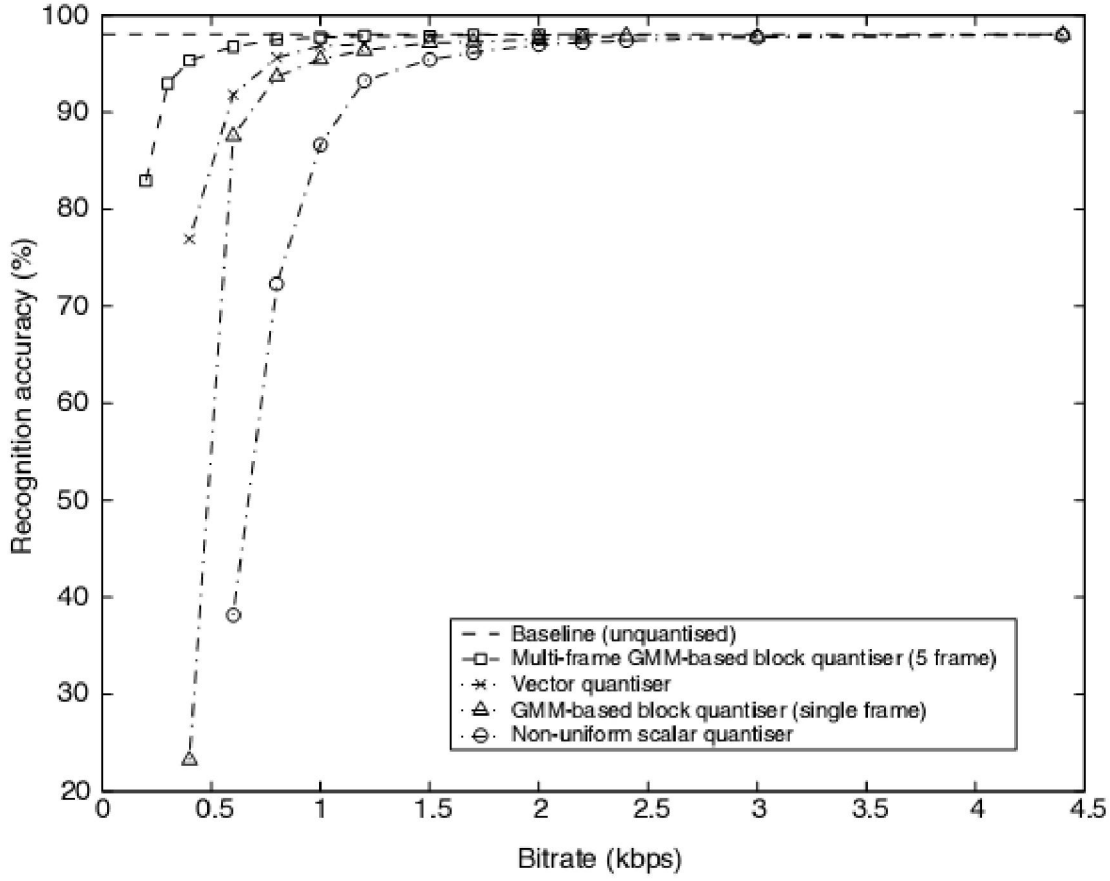


Fig. 5. Summary of average word recognition accuracies for all quantisation schemes considered.

Conclusion and Further Work

- 本論文提出使用多框架GMM區塊量化器來量化在DSR程式中的MFCC特徵。在向量量化器，位元率可調整性與辨識效能上因簡化計算複雜度而得到好的表現。
- 目前的階段還沒有加入額外噪音，因此未來必須在Aurora-2提供的數種噪音環境之下進行測試與改良。
- 伺服端在程序進行與記憶體上的問題並未考慮。綜合以上，提高辨識準確度與限制伺服端的ASR頻道個數是未來的目標。