Improved Noise-Robustness in Distributed Speech Recognition via Perceptually-Weighted Vector Quantisation of Filterbank Energies

Stephen So and Kuldip K. Paliwal

School of Microelectronic Engineering,
Griffith University, Brisbane, Australia, 4111.
s.so@griffith.edu.au, k.paliwal@griffith.edu.au

Abstract

In this paper, we examine a coding scheme for quantising feature vectors in a distributed speech recognition environment that is more robust to noise. It consists of a vector quantiser that operates on the logarithmic filterbank energies (LFBEs). Through the use of a perceptually-weighted Euclidean distance measure, which emphasises the LFBEs that represent the spectral peaks, the vector quantiser codebook provides *a priori* knowledge of the spectral characteristics of clean speech and is used to quantise features from noise-corrupted speech. Our comparative results from the ETSI Aurora-2 recognition task show that the perceptually-weighted vector quantisation of LFBEs achieves higher recognition accuracies for noisy speech than the unweighted vector quantisation, memoryless and multi-frame GMM-based block quantisation and scalar quantisation of Mel frequency-warped cepstral coefficients.

1. Introduction

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.

In *Distributed Speech Recognition* (DSR), depicted in Figure 1, 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. Various schemes for compressing the ASR features have been proposed in the literature. Digalakis *et al.* [1] evaluated the use of uniform and non-uniform scalar quantisers as well as product code vector quantisers for compressing Mel frequency-warped cepstral coefficients (MFCCs) between 1.2 and 10.4 kbps. They concluded that split vector quantisers achieved similar word error rates (WER) than scalar quantisers while requiring less bits. Ramaswamy and Gopalakrishnan [2] investigated the application of tree-searched multi-stage vector quantisers with first order linear prediction operating at 4 kbps. Transform coding, based on the discrete cosine transform (DCT), was investigated in [3] at 4.2 kbps. The ETSI DSR standard [4] uses split vector quantisers to compress the MFCC vectors at 4.4 kbps. Srinivasamurthy *et al.* [5] exploited correlation across consecutive MFCC features by using a DPCM scheme followed by entropy coding.

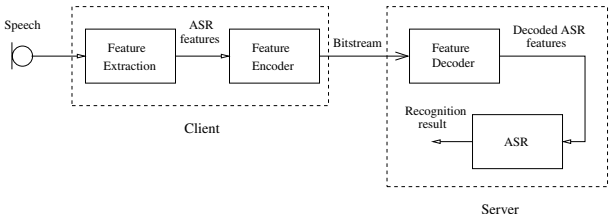

Figure 1: A typical distributed speech recognition system

Using a two-dimensional discrete cosine transform for coding the ASR features, Zhu and Alwan [6] improved the robustness of DSR to noise by using peak-isolated MFCCs (MFCCPs). MFCCPs [7] are derived by applying half-wave rectification to the spectrum reconstructed from a bandpass liftered [8] cepstral vector. They are more robust to noise because of the preservation and emphasis of local spectral peaks, whose frequency locations are known to be important for the discrimination of vowels [7]. Noise-robustness is an important consideration in DSR, since the user at the client side will mostly be immersed in various environmental sounds that will be picked up by the microphone.

In this paper, we present a vector quantiser for logarithmic filterbank energies (LFBEs), which uses a perceptually-weighted Euclidean distance measure[1] to emphasise local spectral peaks and hence improve the noise-robustness of distributed speech recognition. We show that by training the codebook on clean speech features and vector quantising the features of noise-corrupted speech, improved robustness can be achieved on the ETSI Aurora-2 recognition task. We also compare the recognition performance of the perceptually-weighted vector quantiser with four other quantisation schemes of MFCCs that have been previously reported [10]: the unweighted vector quantiser, GMM-based block quantiser, multi-frame GMM-based block quantiser, and non-uniform scalar quantiser.

2. Vector quantisation of ASR features with perceptual weighting

2.1. Logarithmic filterbank energies

Figure 2 shows the process of extracting LFBEs from speech. It can be seen that that by adding a discrete cosine transform to the LFBEs, Figure 2 becomes an MFCC extraction process. There-

[1]This weighted distance measure is similar to the one used in the vector quantisation of line spectral frequencies for speech coding [9].

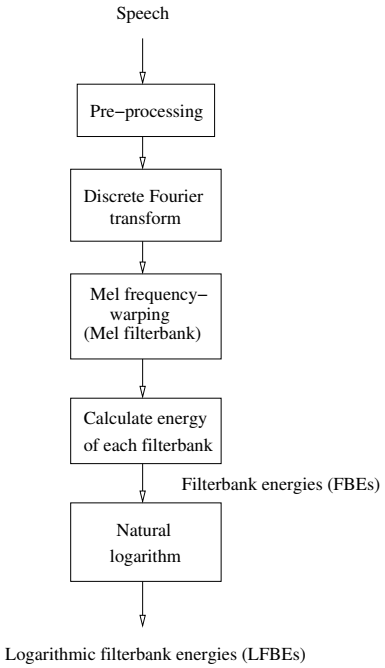

Figure 2: Extraction of logarithmic filterbank energies (LFBEs) from speech

fore, LFBEs contain the same information as MFCCs, as well as being more correlated. In terms of quantisation, it is desirable to decorrelate the vectors before scalar quantisation. This presents no problem for the vector quantiser as it can exploit linear and non-linear dependencies among the vector components (also known as the memory advantage) [11].

In the present study of vector quantising ASR features, LFBEs are favoured over MFCCs because we want to exploit the non-linear spectral sensitivities to increase the noise-robustness in recognition. It is well known that the frequency locations of spectral peaks play a significant role in the discrimination of vowels [7]. A weighted distance measure, to be discussed in a later section, allows the vector quantiser to shift the emphasis to certain parts of the vector. This is only possible when applying vector quantisation in the Fourier spectral domain. Therefore, we use the LFBEs as our ASR feature vector set[2].

2.2. Definition of vector quantisation

The basic definition of a vector quantiser Q of dimension n and size K is a mapping of a vector from n dimensional Euclidean space, $\mathcal{R}^n$, to a finite set, $\mathcal{C}$, containing K reproduction *code-vectors* [11]:

$$Q : \mathcal{R}^n \rightarrow \mathcal{C} \tag{1}$$

where $\mathcal{C} = \{\boldsymbol{y}_i; i \in \mathcal{I}\}$ and $\boldsymbol{y}_i \in \mathcal{R}^n$ [11]. Associated with each reproduction code-vector is a partition of $\mathcal{R}^n$, called a *region* or *cell*, $\mathcal{S} = \{S_i; i \in \mathcal{I}\}$ [11]. The most popular form of vector quantiser is the *Voronoi* or *nearest neighbour* vector quantiser [11], where for each input source vector, $\boldsymbol{x}$, a search is done through the entire codebook to find the nearest code-vector, $\boldsymbol{y}_i$, which has the minimum distance:

$$\boldsymbol{y}_i = Q[\boldsymbol{x}] \quad \text{if } d(\boldsymbol{x}, \boldsymbol{y}_i) < d(\boldsymbol{x}, \boldsymbol{y}_j) \quad \text{for all } i \neq j \tag{2}$$

where $d(\boldsymbol{x}, \boldsymbol{y})$ is the distance measure between the vectors, $\boldsymbol{x}$ and $\boldsymbol{y}$. Generally, the most common distance measure used in vector quantisers is the mean-squared-error (MSE).

2.3. The perceptually-weighted distance measure

The distance measure can be weighted appropriately to vary the emphasis of the quantisation. That is, certain parts of the vector are quantised more finely than others. In coding terms, weighted distance measures perform quantisation noise shaping, which shifts the errors incurred by the quantiser to regions that are deemed as less significant in affecting the fidelity of the reconstructed vector. For example, perceptual weighting filters are used to improve reconstructed speech quality in code-excited linear predictive (CELP) speech coders. Paliwal and Atal [9] introduced a weighted Euclidean distance measure to emphasise specific line spectral frequencies (LSFs) that were situated near the spectral peaks.

The weighted distance measure, $d_w(\boldsymbol{E}, \hat{\boldsymbol{E}})$, between the original LFBE vector, $\boldsymbol{E}$, and the LFBE code-vector, $\hat{\boldsymbol{E}}$, is defined as:

$$d_w(\boldsymbol{E}, \hat{\boldsymbol{E}}) = \sum_{i=1}^{n} \left[w_i (E_i - \hat{E}_i) \right]^2 \tag{3}$$

where n is the vector dimensionality, w_i is the weight of the ith component, E_i and $\hat{E}_i$ are the ith component of the original and code-vector, respectively. In order to emphasise a vector component, E_i, such that it is quantised more finely, the weight, w_i, should be made higher. In the LFBE vector quantiser, it is desirable to emphasise the LFBEs that represent the spectral peaks. Therefore, w_i is set to be a scaled version of the FBE, e^{E_i}:

$$w_i = \left[e^{E_i} \right]^r \tag{4}$$

Through experimentation, we have found 0.5 to be a good value for r.

2.4. Training of clean LFBE codebooks

The Linde-Buzo-Gray (LBG) algorithm [12] is used to train the vector quantiser codebook on zero-mean LFBEs features from clean speech. The zero-mean requirement is needed to ensure consistency in the quantisation, which means that the first cepstral coefficient, c_0, is not used in the recognition task.

Generally, it is beneficial to use the weighted distance measure in the training of the codebook as well as in the vector quantiser encoding. However, because of the mismatch between the training and testing feature vectors, which are extracted from clean and noise-corrupted speech, respectively, we use only the unweighted mean-squared-error for codebook training.

2.5. Encoding of LFBEs from noise-corrupted speech

Zero-mean LFBE features are extracted from the noise-corrupted speech and vector quantised using the clean LFBE codebook. The perceptually-weighted distance measure, expressed in (3), is used to determine the closest matching code-vector. Note that the perceptual weights, expressed in (4), are calculated based on the FBEs from the noise-corrupted speech.

[2]Note that as a consequence of our choice of LFBEs over MFCCs, the vector dimension has increased from 13 to 23. The effect of using less LFBEs to reduce the dimensionality is a topic of further study.

3. Experimental setup

We have evaluated the recognition performance of the perceptually-weighted vector quantiser using the publicly available HMM Toolkit (HTK) 3.2 software on the ETSI Aurora-2 database [13]. Training was done on clean data only (no multi-condition training) and testing was performed using test set A. The effect of different types of noise at varying levels of SNR on the recognition performance was investigated at the bitrate of 1.2 kbps for each quantisation scheme. The parameters for the HMM models are the same as those stated in [13].

The ETSI DSR standard Aurora frontend [4] was used for the LFBE feature extraction, producing 23 LFBEs. 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 for the recognition task. This was done to maintain consistency with the other quantisation schemes that were evaluated in the comparative study [10]. Cepstral liftering [8] was applied to the MFCCs using the following window function, $w(n)$:

$$w(n) = 1 + \frac{L}{2}\sin\left(\frac{\pi n}{L}\right) \quad \text{where } n = 1, 2, \ldots, L \tag{5}$$

where L is the feature length. The MFCCs are then concatenated with their corresponding delta and acceleration coefficients, resulting in a final feature vector dimension of 36.

4. Results and discussion

Tables 2 to 4 show the word recognition accuracies of various quantisation schemes as well as the baseline (without quantisation). These quantisation schemes are the: perceptually-weighted vector quantiser (PWVQ), unweighted vector quantiser (VQ), five frame multi-frame GMM-based block quantiser [10] (GMM5), memoryless GMM-based block quantiser (GMM1), and Lloyd-Max scalar quantisers with non-uniform bit allocation (SQ). The ASR features used are shown as a suffix to the quantiser abbreviations. The top half of each table compares the two vector quantisation schemes while the bottom half contains accuracies for other quantisation schemes that have been investigated in [10].

We can see from the Tables that the proposed perceptually-weighted vector quantisation scheme (PWVQ-LFBE) is more robust to noise than the unweighted vector quantisation of MFCCs (VQ-MFCC). At SNRs of 10 and 15 dB, the PWVQ-LFBE scheme achieves up to 6 to 10% improvement over VQ-MFCC. When compared with the other quantisation schemes, the PWVQ-LFBE scheme maintains good robustness to noise, achieving a higher recognition accuracy than GMM5-MFCC, which is an interframe coding scheme. There is a slight degradation in the recognition performance for clean speech (SNR of ∞ dB) though it is often less than 1%.

5. Conclusion and further work

In this paper, we proposed and evaluated a scheme for quantising ASR feature vectors for distributed speech recognition that is more robust to noise. Noise-robustness is an important consideration in DSR because the user will most likely be immersed in environmental noises. Logarithmic filterbank energies (LFBEs) are extracted from speech and coded using a vector quantiser, which uses a perceptually-weighted distance measure. This distance measure emphasises LFBEs that represent the spectral peaks. Experiments performed using the ETSI Aurora-2 recognition task show that the proposed vector quantisation scheme is more robust to noise than other schemes at 1.2 kbps.

6. References

[1] 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.

[2] 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.

[3] I. Kiss and P. Kapanen, "Robust feature vector compression algorithm for distributed speech recognition", in *Proc. European Conf. Speech Communication Technology*, 1999, pp. 2183–2186.

[4] "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms", Tech. Rep. Standard ES 201 108, European Telecommunications Standards Institute (ETSI), April 11 2000.

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

[6] 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.

[7] 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. 5, pp. 451–464, Sept. 1997.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

Table 1: Word recognition accuracy for speech corrupted with subway noise at varying SNRs (in dB) at 1.2 kbps.

Quantisation scheme	Recognition accuracy (in %)						
	∞ 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
PWVQ-LFBE	96.62	93.28	85.54	65.37	35.22	20.63	13.88
VQ-MFCC	97.11	92.26	81.30	59.32	31.62	19.65	14.03
GMM5-MFCC	97.64	92.48	82.68	58.89	32.61	21.86	15.23
GMM1-MFCC	96.44	89.13	77.40	50.78	27.11	19.37	13.82
SQ-MFCC	92.85	68.47	48.42	30.61	22.35	17.38	12.56

Table 2: Word recognition accuracy for speech corrupted with babble noise at varying SNRs (in dB) at 1.2 kbps

Quantisation scheme	Recognition accuracy (in %)						
	∞ 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
PWVQ-LFBE	97.01	93.89	88.54	71.80	43.50	21.19	11.52
VQ-MFCC	97.16	92.93	85.01	65.11	36.19	19.62	10.67
GMM5-MFCC	98.13	94.98	87.30	65.36	36.73	20.80	12.12
GMM1-MFCC	96.58	91.48	81.92	60.94	34.61	19.35	10.94
SQ-MFCC	93.80	67.87	47.64	29.87	21.31	16.51	10.28

Table 3: Word recognition accuracy for speech corrupted with car noise at varying SNRs (in dB) at 1.2 kbps

Quantisation scheme	Recognition accuracy (in %)						
	∞ 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
PWVQ-LFBE	97.05	93.80	86.79	64.39	32.03	19.71	12.17
VQ-MFCC	97.02	93.14	83.12	54.94	27.02	19.27	11.78
GMM5-MFCC	97.88	93.80	83.21	55.92	29.38	18.52	12.73
GMM1-MFCC	96.39	91.05	78.08	49.42	25.17	18.01	11.24
SQ-MFCC	93.44	70.44	45.87	27.86	22.19	16.94	11.72

Table 4: Word recognition accuracy for speech corrupted with exhibition noise at varying SNRs (in dB) at 1.2 kbps

Quantisation scheme	Recognition accuracy (in %)						
	∞ 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
PWVQ-LFBE	97.16	92.93	83.62	59.24	23.82	18.54	10.12
VQ-MFCC	96.73	91.36	77.63	50.45	26.87	16.94	10.77
GMM5-MFCC	97.90	92.84	81.70	54.83	28.60	18.94	10.86
GMM1-MFCC	96.24	89.45	75.84	45.63	25.42	17.46	11.66
SQ-MFCC	92.97	72.79	48.32	28.63	20.73	13.82	8.89