INTERSPEECH 2009 BRIGHTON

Differential Vector Quantization of Feature Vectors for Distributed Speech Recognition

Jose Enrique Garcia, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

Communications Technology Group (GTC)
Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
jegarlai,ortega,amiguel,lleida@unizar.es

Abstract

Distributed speech recognition arises for solving computational limitations of mobile devices like PDAs or mobile phones. Due to bandwidth restrictions, it is necessary to develop efficient transmission techniques of acoustic features in Automatic Speech Recognition applications. This paper presents a technique for compressing acoustic feature vectors based on Differential Vector Quantization. It is a combination of Vector Quantization and Differential encoding schemes. Recognition experiments have been carried out, showing that the proposed method outperforms the ETSI standard VQ system, and classical VQ schemes for different codebook lengths and situations. With the proposed scheme, bit rates as low as 2.1 kbps can be used without decreasing the performance of the ASR system in terms of WER compared with a system without quantization.

Index Terms: speech recognition, distributed systems, vector quantization

1. Introduction

Distributed Speech recognition (DSR) Systems are based on the client-server paradigm, in which the Front-End (usually the client) extracts feature vectors from speech and sends them out to the Back-End (the server) where the decoding process takes place. This distributed architecture is necessary for running high performance speech recognition applications over mobile devices, due to their computational limitations.

Sometimes, due to the bandwith limitations it is convenient to develop efficient and robust compression techniques for feature vector transmission. Nevertheless, although a large bandwidth were available, a server could be serving a high number of clients, so the available bandwidth should be divided for all of them and a non-efficient transmission scheme would reduce the system performance due to network latencies.

In this work, a technique for MFCC compression is proposed. This method makes use of Differential Vector Quantization (DVQ) that tries to exploit temporal correlation between adjacent frames, due to both, the overlapping of the windowing step and the relatively slow variation of speech production.

This paper is organized as follows, in Section 2 vector quantization techniques are analysed. Section 3 presents the proposed DVQ technique, and other VQ based techniques evaluated in this work. In Section 4 experimental results of the proposed technique are presented, and finally Section 5 shows the conclusions.

This work has been partially funded by the national project TIN2008-06856-C05-04.

2. Vector Quantization

Vector Quantization (VQ) is a source coding method employed for representing in a compact way a value collection. It can be shown that if mutual information of individual features is greater than zero, the use of VQ improves quantization performance compared with scalar quantization.

It was in the eighties when the practical realization was possible thanks to Linde, Buzo and Gray work [1]. This method arises as a generalization of the Lloyd's algorithm [2], which can be seen as an heuristic solution of the k-means problem [3].

The k-means algorithm describes a procedure for clustering data into k classes. It constitutes a method for obtaining a codebook, in such a way that in the quantization process an input vector is represented with the closest codeword, in a minimum distortion sense. The codeword index is then sent to the decoder which is able to reconstruct the original vector with a quantization error. One of the most commonly distortion measure used is the Euclidean distance

The k-means algorithm has some similarity with the Expectation Maximization (EM) algorithm for Gaussian mixtures [4] under some constraints:

- The mixture components have diagonal covariance matrix with unitary elements, and Euclidean distance is used as distortion measure.
- The weights of the components have all the same value.
- While k-means algorithm makes hard assignments of the elements to the clusters, the EM algorithm performs a computation of the membership probability of a cluster.

However, the EM algorithm can be modified to allow hard allocations of elements, reaching the k-means solution [5].

3. Feature Vector compression techniques

Optimal bandwidth resource allocation is essential for distributed speech applications. For improving the performance of the quantizer supplied by the ETSI standard ES 201 108 V1.1.2 [6] (which offers a bit rate of 4.4 kbps) several feature vector compression techniques are presented in this work.

On the one hand, for evaluating the efficiency of the ETSI standard quantizer, several codebooks of different lengths that use the same pairs defined by the standard were generated making use of the k-means algorithm.

On the other hand, two alternative methods for compressing feature vectors are proposed and evaluated. The first one is a combination of the well known differential pulse code modulation (DPCM) with vector quantization that can be called Differential Vector Quantization (DVQ). This codification approach tries to exploit temporal correlation between adjacent frames,

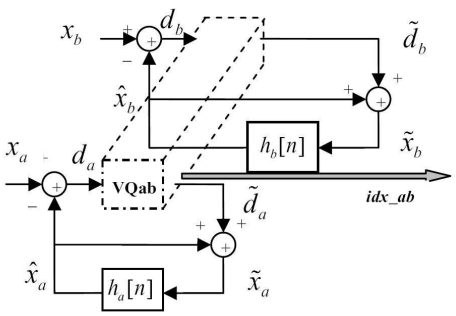

Figure 1: *Block Diagram of the proposed DVQ scheme.*

due to the window overlapping in the feature extraction process and also to the slow variation of speech production.

On the other hand a method of codebook generation based on two classes is also proposed. This method consist on the generation of two codebook families, a family is responsible for the quantization of high energy frames, and the other for the low energy ones. Using this approach, more specific codebooks are generated and the overall quantization error can be reduced.

3.1. ETSI 201 108 V1.1.2 Front-End

ETSI 201 108 V1.1.2 standard presents a collection of algorithms for extracting acoustic features, and their posterior transmission for distributed speech recognition systems. The feature extraction algorithm offers 13 cepstral coefficients, and the log-energy coefficient, with window shifts of 10 ms. Furthermore, it defines a compression algorithm for reducing the transmission rate. This compression is based on vector quantization of feature vectors pairs, resulting at 7 quantized pairs, in which C0 is jointly quantized with log-energy, and the rest of cepstral coefficients are quantized in adjacent pairs. The bit rate obtained using this VQ is 4.4 kbps without error protection.

3.2. Differential vector quantization (DVQ) of MFCC

Differential vector quantization (DVQ) is a compression scheme that makes use of linear prediction jointly with vector quantization of the residual prediction error and has been successfully used in digital video and audio compression [7]. In the proposed scheme, DVQ uses linear prediction over each cepstral coefficient separately, while quantization is performed taking the error prediction by pairs (as defined by the ETSI standard).

The block diagram of the proposed DVQ technique is shown in Fig. 1. Each pair of cepstral coefficients is denoted by the tuple $\mathbf{x} = (x_a, x_b)$. Over this tuple, two predictions are extracted $\hat{\mathbf{x}} = (\hat{x}_a, \hat{x}_b)$ made from quantized values of the previous frame, getting the prediction error pairs

$$\mathbf{d} = (d_a, d_b) = \mathbf{x} - \hat{\mathbf{x}}. \quad (1)$$

Subsequently, these values are quantized resulting in $\tilde{\mathbf{d}} = (\tilde{d_a}, \tilde{d_b})$ and the quantized prediction error

$$\tilde{\mathbf{d}} = \mathbf{d} + \mathbf{e_q}, \quad (2)$$

where $\mathbf{e_q} = (e_{q_a}, e_{q_b})$ is the quantization error.

The prediction errors in (2) will be used to obtain a pair of quantized coefficients using the predicted value of the next frame. Linear prediction filters are denoted as $h_a[n]$ and $h_b[n]$, and for this work, they have been substituted by a single delay line of one sample to reduce the computational cost. Therefore, the predicted value for each coefficient is directly the difference between the current value and the value of the previous frame.

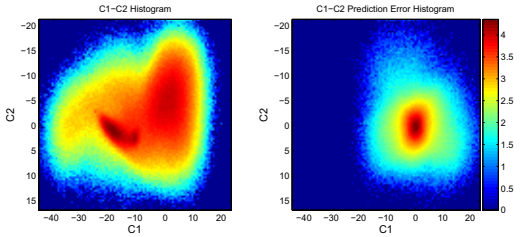

Figure 2: *Histograms of the C1-C2 cepstral coefficient pair (left side) and its corresponding prediction errors (right side).*

One of the main advantages of applying this predictive scheme is that prediction error will not affect forthcoming frames, so the quantization error of a single coefficient is equal to the quantization error of the prediction error.

Similar MFCC quantization schemes have been proposed, in [9] the system is composed of a linear predictor followed by a two stage vector quantizer. This approach presents the problem that quantization error will affect the quantization of forthcoming frames (on the contrary to the proposed DVQ), which can degrade recognition performance. If adaptive filtering were used, the filter coefficients could not be computed with the Backward approach because signals in reception would be different than signals in transmission. Therefore, the coefficients of the predictor filter should be sent, as in the Forward approach, increasing the bandwidth required.

The system proposed in [8] consists of a DPCM codification of each cepstral coefficient, performing uniform scalar quantization (USQ) over the prediction error, followed by an entropy coder. Instead of using USQ, VQ can be used, obtaining better results if the mutual information of each feature pair is greater than zero. A study of mutual information over the prediction error of the cepstral coefficients pairs used in differential vector quantization was carried out, showing that mutual information was positive for all the pairs.

Comparing VQ with DVQ, the later has the advantage that variance and dynamic range is dramatically reduced in all coefficients, so quantization error will be consequently reduced for the same codebook length (same bandwidth), and recognition performance is expected to be improved. As an example of such a dynamic range reduction, in Fig.2 histograms of the cepstral coefficient pair C1-C2 (left side) and its prediction errors (right side) have been plotted.

4. Performance Evaluation

In order to evaluate the performance of the proposed DVQ scheme compared with the rest of quantization techniques (VQ, two class VQ, and the ETSI VQ), a collection of measures and recognition experiments were carried out. In these experiments the quantization error and recognition accuracy for every VQ method was evaluated under different situations.

The databases used for this evaluation were Albayzin [11] and the Spanish part of Speechdat-Car [10]. Albayzin is a Spanish corpus that contains phonetically balanced sentences uttered under noise-free conditions. On the other hand, Speechdat-Car contains noise-free signals recorded using a close talk microphone and noisy speech recorded using a Hands-Free microphone placed on the ceiling of the car in front of the speaker in different driving conditions. In Speechdat-Car isolated words, navigation commands, isolated and connected digits, phonetically rich words and sentences, etc. can be found.

As mentioned before, the codebooks of DVQ, VQ and two-class VQ were trained using different numbers of codewords under three different conditions:

- WELL MATCHED CONDITIONS

  Under these conditions, VQs are adapted to both the task and the environment. The codebooks were trained with the training set of the isolated and connected digits task of Speechdat-Car using the hands-free microphone when hands-free signals are used in the recognition step and using the close-talk microphone when close-talk signals are used for recognition.

- MEDIUM MISMATCHED CONDITIONS

  In this case, VQs are adapted to the acoustic environment but not to the task, hence, codebooks are well suited for working in a specific acoustic environment, but they can be used for any task. For the experiments, the codebooks were trained using the parts of the training set of Speechdat-Car that do not contain digits. Again close-talk microphone signals are used for training if clean speech is going to be used in the decoding stage and hands-free mic signals when noisy speech is going to be recognized.

- HIGHLY MISMATCHED CONDITIONS

  This is the case where the generated VQs are adapted neither to the task nor to the acoustic environment. They can be exported to any other task and environment. This is the case of the ETSI standard VQ. For this purpose, the codebooks of the VQs were trained using the training-set of the phonetically balanced databased Albayzin recorded under laboratory conditions.

Quantizers trained under highly matched conditions are expected to perform better, in terms of quantization error and recognition accuracy, than VQs trained under mismatched conditions. However, the former has the disadvantage of being less portable to any other environment or task. Moreover, quantizers with shorter codebook lengths would be desirable without degrading recognition performance since larger codebooks will need higher bandwith.

4.1. Quantization Mean Square Error

The quantization mean square error (MSE) obtained with a specific VQ approach can give an idea of its behaviour regarding recognition accuracy. Fig. 3 shows the MSE obtained for the three different VQ methods along with the MSE obtained with the ETSI Front-end VQ. Phonetically balanced sentences of the Albayzin database were used to obtain the codebooks and Speechdat-Car connected digit sentences were used to estimate the MSE for every VQ approach. It can be seen that DVQ obtains better performance in terms of quantization error than the rest of methods which indicates that a higher recognition accuracy is expected to obtain. As shown in Fig. 3 the MSE decreases when the length of the codebook increases, as expected, but DVQ achieves low levels of MSE even for very small codebooks.

4.2. Recognition Accuracy Evaluation

Recognition experiments were carried out using the Spanish part of Speechdat-Car database and the connected digits task. For all the recognition experiments, the Back-end of the automatic speech recognition engine is based on continuous HMMs with 16 component Gaussian Mixture Models as observation pdf. The acoustic feature set is the ETSI standard Front-end for all the experiments.

Acoustic units consist of three state word models for digits and cepstral mean subtraction (CMS) was applied right after

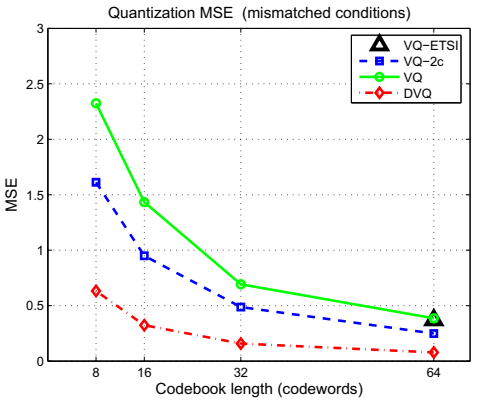

Figure 3: *MSE for the different proposed methods.*

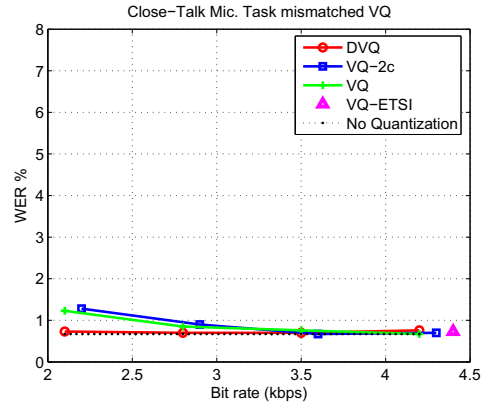

Figure 4: *WER with Close-Talk Mic.for different VQ methods and codebook lengths. Medium mismatched conditions.*

vector de-quantization in the back-end. Acoustic models are always adapted to the previously trained VQ and to the environmental acoustic conditions, therefore, with respect to acoustic models, matched conditions are always considered.

In Fig. 4 the experimental results for medium mismatched conditions and close-talk signal are displayed. It can be seen that under these conditions the performance of all the methods is very high even for small codebooks. Nevertheless, DVQ outperforms the rest of the systems obtaining as good results as performing the recognition without quantizing the MFCC. The results for the matched conditions and close-talk signals are presented in Fig. 5, obtaining lower WER as expected. Again, DVQ achieves the best results even for very low bit rates.

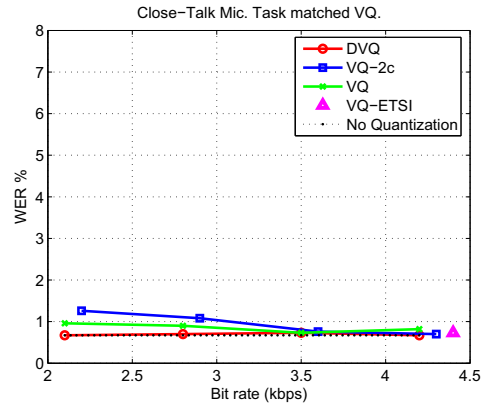

Figure 5: *WER with Close-Talk Mic.for different VQ methods and codebook lengths. Well matched conditions.*

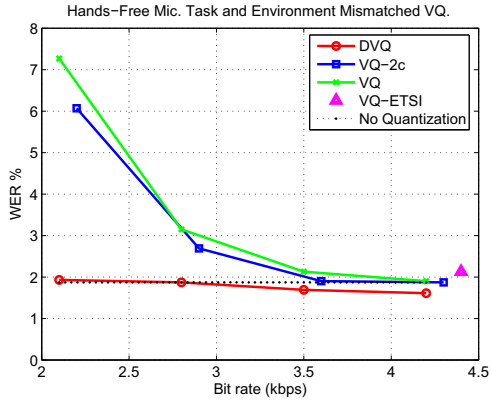

Figure 6: *WER with Hands-Free Mic.for different VQ methods and codebook lengths. Highly mismatched conditions.*

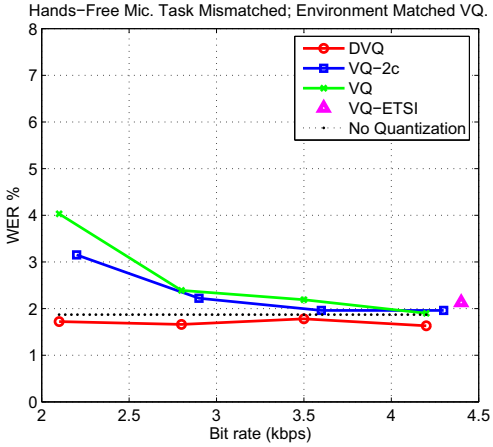

Figure 7: *WER with Hands-Free Mic. for different VQ methods and codebook lengths. Medium mismatched conditions.*

When dealing with noisy signal higher error rates are expected to be obtained. Figs. 6, 7 and 8 show the results for hands-free microphone recognition using highly mismatched, medium mismatched and well matched VQs respectively.

As can be seen in Figs. from 4 to 8, the DVQ method has a consistent behaviour along all bandwidths both for clean and for noisy conditions. The WER obtained is similar to that obtained both by the ETSI standard VQ and by non-quantized MFCC.

Traditional VQ and two-class VQ obtain higher WER when very low bandwidth is used, especially for noisy conditions. However, for bandwidths higher that 3 kbps VQ and two-class VQ have a similar performance to that obtained without using VQ. It can be seen also that matched condition codebooks obtain especial importance in low bit rates for VQ and two-class VQ, improving WER under highly mismatched conditions.

5. Conclusions

In this work a method for compressing acoustic feature vectors based on Differential Vector Quantization has been proposed and evaluated. Several experiments have been carried out showing that Differential Vector Quantization outperforms traditional Vector Quantization and the ETSI Standard Front-End Vector Quantization when low bit rate conditions are desired. Several situations have been considered regarding the degree of matching between the data used to train the codebooks and the data that will be used in the recognition stage. Differential Vector Quantization obtains similar performance to that obtained without using traditional Vector Quantization even when the data used for training the codebooks are adapted neither to the acoustic environment nor to the recognition task for bit rates as low as 2.1 Kbps. Recognition performance for the rest of the Vector Quantization techniques degrades when available bandwith is reduced but Differential Vector Quantization can offer similar performance to that obtained using the standard ETSI Vector Quantization (which has a transmission rate of 4.4 kbps) and that obtained using MFCC without any kind of quantization, both for clean signal and noisy signal.

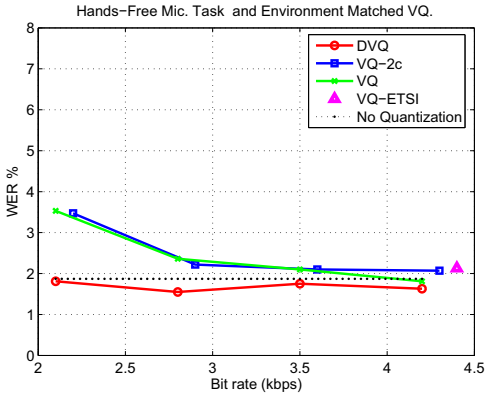

Figure 8: *WER with Hands-Free Mic. for different VQ methods and codebook lengths. Well matched conditions.*

6. References

[1] Y. Linde, A. Buzo, R. Gray, "An Algorithm for Vector Quantizer Design", IEEE Trans on Comm., v.28, 1980.

[2] S. P. Lloyd, "Least Squares Quantization in PCM", IEEE Trans. on Inf. Theory, vol. 28(2), 129-137, 1982.

[3] X. Huang, X., A. Acero, H. Hon, "Spoken Language Processing: a guide to theory, algorithm, and system development", Prentice-Hall, 2001.

[4] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", J. R. Statistic Society, vol. 39(1), 1-21, 1977.

[5] D. Qiu, and C. Ajit Tamhane, "A comparative study of the K-means algorithm and the normal mixture model for clustering: Univariate case" Journal of Statistical Planning and Inference, vol. 137(11), 3722-3740,2007.

[6] ETSI standard document, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end Feature extraction algorithm; Compression algorithms", ETSI ES 201 108 Ver. 1.1.2.(2000-04).

[7] J. E. Fowler. M. R. Carbonara. and S. C. Ahalt. "Image coding using differential vector quantization." IEEE Trans Circuits Syst. Video Tech., vol. 3., 350-367, 1993.

[8] N. Srinivasamurthy, A. Ortega and Shrikanth Narayanan, "Efficient scalable encoding for distributed speech recognition", Speech Communication, vol. 48, 888-902, 2006.

[9] G. N. Ramaswamy, and P. S. Gopalakrishnan, "Compression of Acoustic Features for Speech Recognition in Network Environments" in Proc. of ICASSP, 1998.

[10] A. Moreno, B. Lindberg, C. Draxler, G. Richard, K. Choukri, S. Euler, and J. Allen, "Speechdat-car: A large speech database for automotive environments", in Proc. of LREC, vol. 2, 895-900, 2000.

[11] F. Casacuberta, R. Garcia, J. Llisterri, C. Nadeu, J.M. Pardo, A. Rubio, "Development of a Spanish Corpora for Speech Research (Albayzin)" Workshop on Standarization of Speech Databases and Speech Assessment Methods, 1991.