ROBUST SPEECH RECOGNITION USING FUZZY MATRIX QUANTISATION, NEURAL NETWORKS AND HIDDEN MARKOV MODELS

Professor C S Xydeas and Lin Cong

Speech Processing Research Laboratory,
Electrical Engineering Division, School of Engineering, University of Manchester,
Dover Street, Manchester, M13 9PL, UK, Tel/Fax: +44[161]2754511/2754528, E-Mail: c.xydeas@man.ac.uk

Abstract

In this paper a new approach to robust speech recognition using Fuzzy Matrix Quantisation, Hidden Markov Models and Neural Networks is presented and tested when speech is corrupted by car noise. Thus two new robust isolated word speech recognition (IWSR) systems called FMQ/HMM and FMQ/MLP, are proposed and designed optimally for operation in a variety of input SNR conditions. The schemes and associated system training methodologies result into a particularly high recognition performance at input SNR levels as low as 5 and 0 dBs.

1. Introduction

Next generation voice communication and information systems require efficient interaction mechanisms between users and terminals or remote database systems, and speaker dependent (SD)/independent (SI) isolated word speech recognition (IWSR) can be employed for this purpose. For example SI/SD, IWSR is an enabling technology for hands free dialling and interaction with voice store and forward systems, when the user is otherwise occupied driving a car. Of course IWSR has received considerable attention in the last two decades but, there is still a challenge in designing robust IWSR systems capable of operating successfully at relatively low Signal to Noise Ratio (SNR) input conditions, when speech is corrupted by acoustic noise. In general, the performance of existing medium to small vocabulary size IWSR schemes tends to deteriorate rapidly when SNR is lower than 20 dBs.

Our previous work [1] to [4] in robust IWSR systems excluded the use of acoustic noise reduction preprocessing and involved training the system using "clean" speech, during the IWSR design phase of the process. Improved performance under noisy input conditions is mainly obtained in this case by employing system components which are intrinsically robust enough to acoustic noise.

This paper considers the case where an IWSR system is designed and optimised, during training, using "clean" as well as "noise corrupted" speech signals. In particular, two new IWSR systems are proposed, which employ FMQ as the spectral labelling process, followed by a Hidden Markov Model (FMQ-HMM) or a Neural Network (FMQ-MLP) classification technique.

Both systems provide significant benefits in recognition accuracy, at low SNR input signal conditions, when compared to conventional previously reported methods ([1] to [4]). This robust performance ensures that a typical recognition accuracy of the order of 93% and 85% is obtained at input SNR values of 5 dB and 0 dB respectively. In contrast, under the same conditions, conventional systems can manage only a poor 55% and 37%. The theory and structure of the proposed relatively small vocabulary, SD-IWSR schemes is presented in the paper together with recognition performance computer simulation results based on extensive tests.

The paper is divided into six sections. Section 2 discuss the input speech database used in the design and training of the two systems while section 3 considers the FMQ/MLP system. FMQ/HMM is discussed in section 4, whereas computer simulation results and conclusions are presented in sections 5 and 6 respectively.

2. Input Speech Data

Input speech signals are band limited to 3.6 kHz, sampled at 8 ksamples/sec and quantised with 16 bits per sample. A 16th order LSP analysis is performed every 20 msecs allowing for a 10 msecs overlap between analysis frames. The input speech database, employed in system training and recognition experiments, is configured using the 26 English letters. In particular, 130 versions of each letter were obtained. These were produced by five male and five female speakers, each providing 13 versions. 100 versions of a word were used for training and the remaining 30 were employed in testing system recognition performance. Furthermore the above 130 versions of each vocabulary word were provided for each of the following input SNR conditions: clean speech, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. The total number of training and testing utterances contained in the input database is therefore 20280.

3. FMQ/MLP System Description

Figure 1 shows the proposed FMQ/MLP system and highlights its two modes of operation: i) training mode when switches SW and $\overline{\text{SW}}$ are set to 1 and ii) speech recognition mode with SW = 2 and $\overline{\text{SW}}$ = 2. In this figure the input $\{s(k)\}$ signal is processed via a voice activity detector (VAD) that effectively defines the end points of input words. Word sequences of samples segmented into 20 msecs frames with a 10 msecs overlap are then pre-emphasised. An LSP analysis performed on each frame provides the speech short term spectral envelope information to the system whose spectrum labelling process is performed via Fuzzy Matrix Quantisation. During the training mode, the process of Matrix quantisation is divided into two parts: i) the design of MQ codebooks at different input SNR conditions, i.e., with clean speech and speech corrupted by car noise at different SNR values. These different sets of codebooks are then employed in the MLP design process. ii) the design of a robust MQ codebook that is designed using both clean and noisy input data. This codebook is employed during the recognition mode of system operation.

The training and recognition modes of operation are described in the following sub-sections.

FMQ Codebook Design Process

The process of designing six sets of Matrix codebooks involves the following three steps:

1. The training part of the input database is sub-divided into six sections D1 to D6 based on signals obtained at six different SNRs levels (∞ dB, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB). Each section consists of 2600 words, since each of the 26 vocabulary words is represented by 100 versions. Thus section D1 consists of 2600 clean speech words, section D2 contains 2600 words at 20 dB SNR and so on.
2. Each database section D_i, $i = 1, 2, \cdots, 6$, provides an assembly LSP_i of vectors containing 16 LSP spectral coefficients and each assembly is then used to design a set of Matrix codebooks MCB_j^i, $j = 1, 2, \cdots, u$, where u = 26 is the number of vocabulary words in the system.
3. Thus six sets of Matrix codebooks are generated. Each set contains 26 codebooks and each codebook has C = 128 entries. Notice that each entry is a P by N matrix [4]. These codebooks are designed optimally for minimum quantisation distortion using the Matrix LBG algorithm [4].

The final set of **u** matrix codebooks to be used during recognition, is designed via the Matrix LBG algorithm, using the whole training part of the input database.

MLP Training Process

The training database formed from 26 vocabulary words, each repeated 100 times for 6 different input SNR conditions, can be presented diagrammatically as in Figure 2. Each "x-y word plane" contains 600 versions and each x-column of the word plane contains the same word version at different SNR values, i.e., ∞, 20, 15, 10, 5 and 0 dB. Recall that a set of MCB_j^i codebooks, $j = 1, 2, \cdots, 26$ for each input noise condition $i = 1, 2, \cdots, 6$, is already available. The MLP training process is organised so that a given version of a vocabulary word, that has been produced at the kth SNR input condition, is Matrix quantised by the kth set of 26 MCB_j^k codebooks. This quantisation process produces FD_j^k, $j = 1, 2, \cdots, u$ distance measures [3] which form the MLP input, see Figure 1. Thus the MLP network is trained for the nth vocabulary word, using the back propagation algorithm, by quantising the nth "x-y word plane" one column at the time using the appropriate set of MCB_j^i Matrix codebooks. Each of the six SNR values in a column

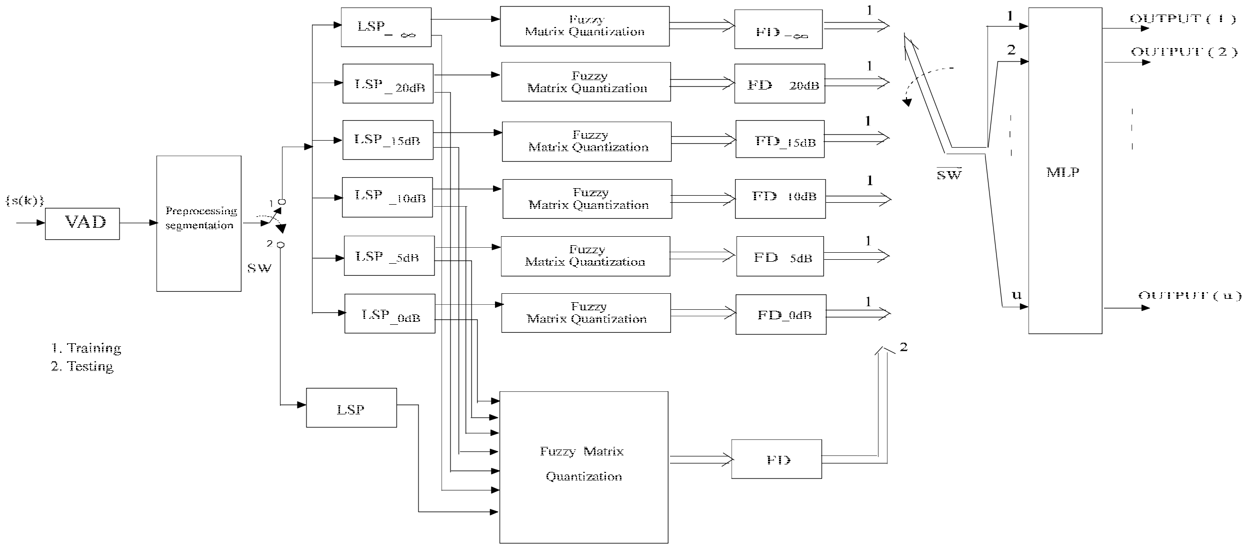

Fig. 1 FMQ/MLP System

provides a set of FD_j^i distance measures which are presented as input to the MLP training process.

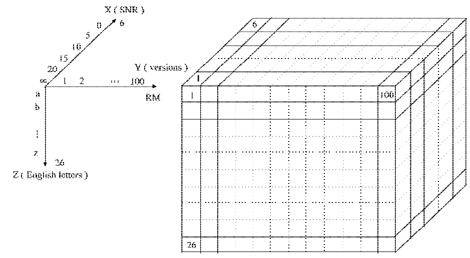

Fig. 2 Training word database arrangement

FMQ/MLP Recognition Process

When the system operates in a recognition mode, an input word W_j represented by a series $\{x_1, x_2, \cdots, x_{T_j}\}$ of T_j LSP vectors, is Fuzzy Matrix quantised in parallel by **u** different matrix codebooks each designed using all the training data available for a given vocabulary word. The fuzzy, codebook-dependent, **u**-dimensional distortion measure vector

$$\overline{FD} = [FD_1, FD_2, \cdots, FD_u]$$

is presented to the MLP classification process: whose outputs {OUT(1), OUT(2), ⋯, OUT(u)}, assume values in the region $0 \le OUT(j) \le 1$. The system classifies the input word W_j to be the ith vocabulary word if:

$$OUT(i) = \max\{OUT(1), OUT(2), \cdots, OUT(u)\}$$

4. Robust FMQ/HMM System

The FMQ/MLP improved recognition performance characteristics at low input SNR values can be attributed to the particular methodology used to expose the FMQ and MLP design processes to different input signal conditions. This powerful and general system training approach can be also applied to other IWSR systems. Thus a further robust HMM based IWSR structure is discussed in this following section.

System Description

The system shown in Figure 3, uses FMQ as the front end of a HMM classifier and involves training FMQ codebooks and HMMs with signals at different SNR conditions. Training is performed when SW is set in position 1 whereas during recognition SW = 2. The system has been trained using the speech database discussed in section 2. Training involves the processes of LSP calculation, MQ codebooks design and HMM design.

Words in the training database are organised again as in Figure 2. Each plane has 600 versions of a given word with 100 versions allocated to one of six input SNR conditions. All the 15600 words are LPC analysed to produce vectors of LSP coefficients.

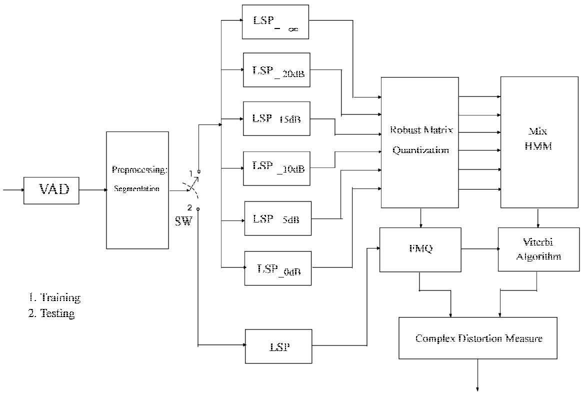

Fig. 3 FMQ/HMM System

Each robust codebook MCB_j, j = 1, 2, ..., u, is formed using the matrix LBG algorithm [4] on LSP vectors produced by the jth word plane. In this way, **u**, C-entries codebooks are designed for minimum quantisation distortion.

Furthermore, the method described in [1] is used to set up an HMM λ_j for each word vocabulary. However, in this case, the observation sequences $O = \{O_1, O_2, \cdots, O_{T_i}\}$ are now obtained from the 600 utterances of a given word plane, which are matrix quantised by the corresponding and previously designed robust codebook.

The recognition process is similar to that of the FVQ/HMM system discussed in [1]. The only difference is that the single frame (N = 1) Fuzzy Vector Quantisation (FVQ) scheme is replaced with multiframe (N > 1) Fuzzy Matrix Quantization (FMQ). The FMQ output is the Fuzzy distortion measure vector, $\overline{FD} = [FD_1, FD_2, \cdots, FD_u]$ whereas the HMM part provides the maximum likelihood probability $Pr(O/\lambda_j)$, j = 1, 2, ..., u. These two measures are combined and the jth vocabulary word is recognised by the minimum:

$$D_h(j) = \min_{1 \le j \le u} (FD_j - \alpha\, Pr(\mathbf{O}/\lambda_j))$$

where α is a scaling constant.

5. Experimental Results

The "test" part of the input database discussed in section 2 was used in computer simulation experiments, in order to determine system performance at different SNR input conditions.

FMQ/MLP System Performance

In these experiments, the matrix quantisation length is chosen as N = 2 and the number of MLP hidden nodes P is 26. Furthermore it takes about 15000 iterations to train the MLP part of the system and thus FMQ/MLP converges faster than the FVQ/MLP [2] or FVQ/HMM/MLP [3] systems.

FMQ/MLP system performance is shown in Figure 4 where it is compared with that of FMQ/FVQ-HMM and FMQ. The last two systems are discussed in [4] and operate under "mismatched" input noise conditions, i.e., they were trained using clean speech.

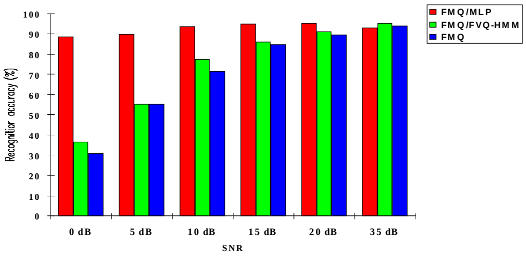

Fig. 4 FMQ/MLP, FMQ/FVQ-HMM and FMQ system performance, C = 128

From Figure 4, it is clear that FMQ/FVQ-HMM and FMQ outperform FMQ/MLP when SNR is set at 35 dB. However, at SNR values of 20 dB or lower, FMQ/MLP achieves progressively significantly better performance. Its recognition rate at 20 dB is 95.13%, compared to 91.03% and 89.61% obtained from FMQ/FVQ-HMM and FMQ respectively. Furthermore at the SNR value of 0 dB, FMQ/MLP system achieves an impressive of 88.46%, which is more than 50% higher than the rate of the other two systems. This shows that the method of gradually contaminating during training the input speech signal with noise, gives the MLP network a significant "noise immunity" capability.

FMQ/HMM System Performance

The FMQ/HMM system has been tested for different matrix lengths N and different numbers of codewords C in each Matrix codebook, i.e., N varies from 2 to 3, and C is set to 128 or 256. The FMQ degree of fuzziness F is 1.2 whereas the number of HMM states is equal to 5.

Figure 5 shows FMQ/HMM performance for N = 2, N = 3 and C = 128, together with that of the FMQ part of the system operating as an independent IWSR system. Results of the same system but with C = 256 are shown in Figure 6. Notice that FMQ/HMM with C = 256 provides consistently the best overall performance. FMQ/HMM has an 1 to 3% advantage over the equivalent FMQ scheme.

When N = 2 and C = 128, and at input SNR values of 0 dB, 15 dB and 20 dB, a comparison between FMQ, FMQ/HMM and FMQ/MLP reveals that FMQ/MLP outperforms the other two systems. However, this can not be stated for higher SNR values although in this case the difference in recognition performance between all the systems is rather small. However, what can be clearly stated from Figure 4, 5 and 6 is that both FMQ/HMM and FMQ/MLP are particularly robust at low SNR input conditions.

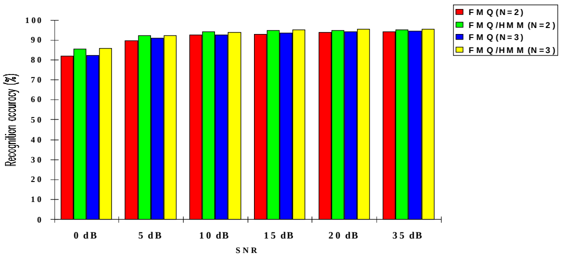

Fig. 5 FMQ, FMQ/HMM systems, C = 128

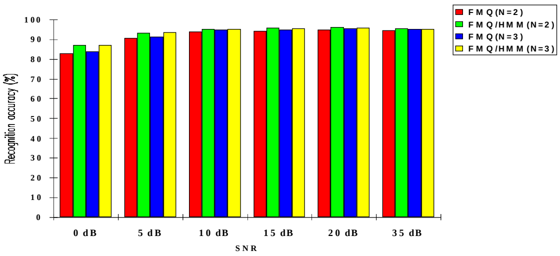

Fig. 6 FMQ, FMQ/HMM systems, C = 256

6 Conclusions

This paper considers the case where an IWSR system is designed and optimised, during training, using clean as well as noise corrupted speech signals. In particular, two new IWSR systems are proposed, which employ FMQ as the spectral labelling process, followed by a Hidden Markov Model (FMQ/HMM) or a Neural Network (FMQ/MLP) classification technique. Both systems provide significant benefits in recognition accuracy, at low SNR input signal conditions by using a new and successful system training process, when compared to conventional previously reported methods [1] to [4]. FMQ/HMM achieves a recognition rate of 87.05% at 0 dB input SNR whereas at 20 dB SNR performance increases to 94.74%. The corresponding FMQ/MLP rates are 88.46% and 95.13%.

REFERENCES

[1] **L. Cong, C S Xydeas and A F Erwood**: "A Study of Robust Isolated Word Recognition Based on Fuzzy Methods", EUSIPCO-94, Vol. 1, pp.99-102, Sept., 1994, UK

[2] **L. Cong, C S Xydeas and A F Erwood**: "Combining Fuzzy Vector Quantisation and Neural Network Classification for Robust Isolated Word Speech Recognition", ICCS'94, Vol. 3, pp.884-887, Nov., 1994, Singapore

[3] **C S Xydeas and L. Cong:** "Combining Neural Network Classification with Fuzzy Vector Quantization and HMMs for Robust Isolated Word Speech Recognition", IEEE Intern. Symposium on Information Theory, pp.117, Sept., 1995, Canada

[4] **C S Xydeas and L. Cong:** "Robust Speech Recognition in A Car Environment", Intern. Conference On Digital Signal Processing, Vol. 1, pp.84-89, June, 1995, Cyprus