

# ROBUST CONNECTED DIGIT RECOGNITION IN A CAR NOISE ENVIRONMENT

*Lin Cong and Saf Asghar*

Advanced Micro Devices, Peterson Way,  
Santa Clara, CA, USA

## ABSTRACT

This paper proposes a robust speaker-independent, connected digit recognition system for mobile applications. The system requires a small amount of ROM and low computational cost with high recognition accuracy. In addition, the system can be efficiently implemented on most currently available 32-bit fixed-point DSP chips. To reach these goals, we combined robust speech parameter processing technologies with dual MQ and VQ pairs, which supply discrete gender-dependent HMM to increase the performance of HMMs. The dual MQ/VQ pairs exploit the “evolution” of the speech short-term spectral envelopes with one pair providing error compensation using LSP mean compensated coefficients. Correspondingly, we proposed the dual MQ/HMM and VQ/HMM decoding pair algorithm. In a car noise environment, the system attains an 80% average connected digit recognition accuracy at around 10 dB SNR. A digit accuracy of 93% is obtained at 5 dB SNR.

## 1. INTRODUCTION

Our goal is to develop a voice recognition system for hand free dialing purposes. Since this application requires a small vocabulary size, it was possible to build our system based on discrete Hidden Markov Models (HMMs) with low cost and good performance, as discrete HMMs have much less computational requirements. But the key is to deal with the quantization error which is a well-known problem for speech recognition systems based on discrete HMMs. Therefore, we correspondingly proposed several methodologies discussed in Section 2 and 3 to compensate the errors caused by VQ in order to realize our goals.

The quantization errors were compensated by using dual quantization/decoding technology and compensated speech parameters. The Matrix Quantization (MQ) which can capture spectral envelope “dynamics”, is used to compensate quantization error caused by the conventional Vector Quantization (VQ). Correspondingly, we proposed the dual MQ/HMM and VQ/HMM decoding pair algorithm in this paper. I.e., instead of only using the conventional Viterbi decoding algorithm based on VQ/HMM, we also used MQ/HMM to compensate the decoding errors from VQ/HMM in the recognition process of the system. Furthermore, the LSP mean compensated coefficients were added to form two data streams to reduce noise effects. The

combination of these methodologies can provide the highest recognition accuracy for speaker-independent (SI) connected digit recognition in a car noise environment.

We have developed the algorithms based on whole digit models. One of the most important parts of this design was how to build the corresponding background noise discrete HMM model. We have built the back ground noise and short pause discrete HMMs based on proposed dual MQ/VQ pair, which is discussed in Section 2.3. Another most important part of this design is the proposed dual MQ/HMM and VQ/HMM decoding pair algorithm which is described in Section 3.

Furthermore, the grammar free/constraint decoding algorithm can be selected used in the system in order to increase the performance of the system. In this way, we can build a recognition system with the following advantages: i) the ROM size and the computational complexity are greatly reduced for HMM modeling and Viterbi decoding in both of train and recognition processes, so that our recognition algorithms can be implemented at a very low cost on a 32-bit fixed-point DSP chip, ii) good performance in noise environment can be achieved by using the proposed compensation methodologies and grammar constraint decoding algorithm.

The speech database TIDIGITS is used in this paper. The car noise database in our system was from NOISEX\_92. The speech parameter analysis is performed every 20 msecs, allowing for a 10 msecs overlap between analysis frames.

This paper is structured as follows. Section 2 describes the proposed effective enhanced modeling methodologies. The proposed Viterbi decoding pair algorithm is discussed in Section 3. The complexity issue is in the following Section. The experiment results are given in Section 5 and conclusions are presented in Section 6.

## 2. EFFICIENTLY ENHANCED PROCESSING

We proposed Dual MQ/VQ quantization pair designing method with LSP mean compensated coefficients for both of speech and noise. Correspondingly, the integrated noise digit methodology was used to build the digit HMM models. The noise HMM model can be built based on three deferent kinds of codebook designing methodologies discussed in Section 2.3. These processing can efficiently enhance the robustness of the proposed system.

## 2.1 LSP Mean Compensated Coefficients

Considering the distortion caused by using different microphones among the training speech, noise databases and the testing databases, the LSP coefficients generated in LSP and energy parameters were mean compensated by LSP mean compensated parameters in order to normalize the noise and channel distortion effects. LSP mean compensated parameters operation mean compensates each  $j^{th}$  order LSP coefficient for each of the TO frames of speech input signal in accordance with the following Equation:

$$\hat{f}_t(j) = f_t(j) - (\sum_{n=1}^{TO} f_n(j)) / TO \quad (2-1)$$

where  $j = 1, 2, \dots, P$  and  $t = 1, 2, \dots, TO$ .  $\hat{f}_t(j)$  represents the  $j^{th}$  order compensated LSP coefficient for the  $t^{th}$  frame of speech input signal, and the subtrahend of Equation (2-1) is the mean value of the  $j^{th}$  LSP coefficient over all TO frames of speech signal. In this simple way, we found that there was a 2% improvement when adding the LSP mean compensated parameters data stream into the original system based on 12-dimension LSP spectral coefficients with the normalized log frame energy, their 1<sup>st</sup> and 2<sup>nd</sup> order time derivative of the frame energy.

## 2.2 Integrated Noise Word MQ/VQ

First, the car noise database used in our system was from NOISEX\_92, for a car travelling at 120 km/h, which was recycled through the complete training database TIDIGIT (which is sub-divided into six sections, based on signals obtained at six different car noise SNRs levels ( $\infty$  dB, 25dB, 20dB, 15dB, 10dB, 5dB)).

The processes of designing the MQ/VQ parts involve the following 2 steps:

- i) each database section  $D_i, i = 1, 2, \dots, 6$  based on signals obtained at six different car noise SNRs levels) provides an assembly of vectors data stream 1 and data stream 2 (the 12-LSP spectral coefficients with the normalized log frame energy, their 1<sup>st</sup> and 2<sup>nd</sup> order time derivative of the frame energy ( $e_t, \Delta e_t$  and  $\Delta^2 e_t$ ) was called data stream 1, and 12-dimension LSP mean compensated coefficients was called data stream 2).
- ii) the two data streams are used to design the four codebooks (MQ/VQ) separately. The split quantization algorithm can be used to design the MQ/VQ in order to target noise-affected input signal parameters and minimize noise influence when we know the spectral distribution of the background noise. For example, in our application training MQ/VQ can be simply divided into two parts, according to the distribution of the car noise spectrum. Each part is sub-matrix quantized separately using MQ/VQ.

## 2.3 Dual MQ/VQ Pair Design for Noise

Considering the background noise and short pause HMM modeling, we have designed the codebooks in three different ways: i) the separate noise codebooks were designed based on noise database only, ii) the system shared the digit codebooks described in Section 2.3 with noise codebooks, iii) designing one codebook for each digit and one codebook for the noise first, then combining these individual codebooks to produce one codebook.

## 2.4 Integrated Noise Word HMM Modeling

Following the completion of designing the robust codebooks, we built gender-dependent HMMs [2][4]  $\lambda_j, j = 1, 2, \dots, u$  with 7 or 8 states (notice that the first state and the last state are non-emitting.  $u$  is the size of the input vocabulary) for each input digit based on the two data streams. In this case, the observation sequences  $O = \{o_1, o_2, \dots, o_T\}$  are now obtained from a given word

at different car SNR levels, which are quantized with Dual MQ and VQ pairs separately, by the corresponding and previously designed robust codebooks.

## 2.5 Background Noise HMM Modeling

Correspondingly, we have to build two set of noise HMM parameters in the system by using the two noise data streams as well. One is short pause HMM models with 3-state added at the each digit HMM models, and the other one is background noise HMM models with 5-states (notice that the first state and the last state are non-emitting).

Building the noise HMMs based on the three different codebooks described in Section 2.3, had been tested in this paper. The computer simulation illustrated that the system gave much better performance by using last two noise codebooks to design noise HMMs than by using the first one, as it is difficult for the system to find the end points of the input speech signals by using the noise HMMs based on the first noise codebook design method.

## 3. VITERBI DECODING PAIR ALGORITHM

Following the robust training system design, we proposed the dual MQ/HMM and VQ/HMM pair viterbi decoding algorithm in the recognition system to compensate the quantization error caused by the algorithm based on the system VQ/HMM. In this algorithm there were two decoding paths in the recognition system shown in Figure 1: path MQ/HMM and path VQ/HMM. In the Viterbi decoding processes, the Word Link Record (WLR) was built in each of them, separately.

Meanwhile, for our applications, task-specific information can be used so that we used the N-best algorithm [5][6] to tracing back the best digit string in our connected digit recognition system. In the process of the system tracing back the N-best digit string, there are the two

probabilities at each digit boundary of each path in the WLR:  $Prob_{VQ}(arc, n)$  and  $Prob_{MQ}(arc, n)$ , where  $arc = 1, 2, \dots, number$  ( $number$  is the length of the input digit string) and  $n = 1, 2, \dots, N$  [6]. After  $Prob_{VQ}(arc, n)$  was scaled and compared with  $Prob_{MQ}(arc, n)$ , the system classifies the each input digit to the output digit according to which path has bigger probability as shown in Fig. 1. In this way, the system classifies the final N-best digit strings in a very simple and efficient way.

Also the VQ error can be compensated by the MQ/HMM decoding process. Meanwhile, a speaker is generally not able to speak at precisely the same rate for different repetitions of the words, which can cause difficulty for Matrix Quantization if a speaker speaks words very quickly. The solution to this difficulty has been to use VQ to compensate for MQ error by obtaining the best knowledge for the training system and high recognition accuracy in the recognition process as well.

In more detail, the MQ/HMM and VQ/HMM decoding pair

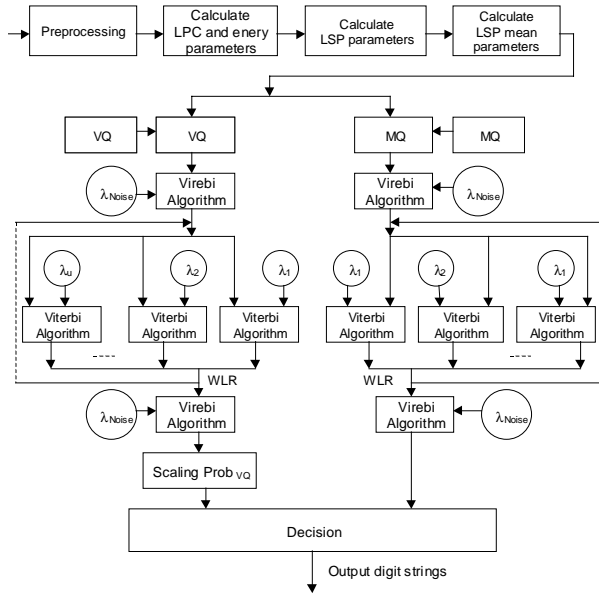


Fig.1 The recognition system

algorithm is simply as follows:

Algorithm:

```

for n=1 to N do
  for arc=1 to number
    Scaling  $Prob_{VQ}(arc, n)$  at each WLR;
    Compare the two path probabilities in the WLR;
    Judge the output digit according to the probabilities;
  end;
end;

```

#### 4. EFFICIENCY RESOURCE IMPLEMENTATION

We will briefly discuss the storage and computational complexity of the recognition part of our system.  $C_N * F_n$  is

the computational complexity of the algorithm, where  $C_N$  is a constant and  $F_n$  is the frame number of input speech signals. The requirements for the storage and the computational complexity of the recognition system can be divided into three parts: the pre-processing of the signal (the computation included by using Burg algorithm to get 12-LPC speech parameters and then from LPC to LSP), MQ/VQ and Viterbi decoding.

For the requirement of ROM size to storage a codebook, we have,

$$Number\ of\ kernel = C = 320 \quad (4-1)$$

where C is the number of the codewords. For the requirement of ROM size to storage the HMMs, we have:

$$Size\ of\ ROM = u * M * M + u * C * M \\ = 10 * 6 * 6 + 10 * 320 * 6 = 19560 \quad (4-2)$$

where u the number of the input vocabulary, and M is the number of states per model (notice that we added the short pause model to the end of each of the words model). The ROM size increases correspondingly when we use Dual MQ/VQ pair quantization and gender-dependent HMM.

For the requirement of computation in each speech frame, briefly, in the MQ part, we have :

$$Number\ of\ kernels = C * P1 + C * P2 \\ = 320 * 15 + 320 * 12 = 8640 \quad (4-3)$$

where P1 is number of the dimension of data stream 1, P2 is the number of the dimension of the LSP mean compensated coefficients. In the grammar constraint Viterbi decoding part, we have about:

$$Number\ of\ kernels \approx W * u * M = W * 10 * 6 = 60 * W \quad (4-4)$$

and in the grammar free Viterbi decoding part, we have about:

$$Number\ of\ kernels \approx u * M = 10 * 6 = 60 \quad (4-5)$$

where W is the number of the digit string. The total number of kernels is 8700. Considering we use the gender HMMs, the computation for MQ and Viterbi decoding parts should be double. Therefore, it is clear that the computation requirement is much less than using continuous density HMM in the system.

#### 5. EXPERIMENTS AND RESULTS

The system MQ/HMM\_VQ/HMM was tested by using about 2000 connected digit strings from the database TIDIGITS added with the car noise (NOISEX\_92) at about 10 dB. In the experiments, the MQ length is chosen as 2. The scaling number  $\alpha$  is set to 1.8, experimentally.

Meanwhile, the system has been tested by using deferent codebook size  $C = 320, 640, 740$  and 1280. The state number  $S_N$  of the HMMs were varied from 7 to 8 in order to find the best C and  $S_N$  values for the proposed system, although the Isolated Word Speech Recognition system gets

the best results when  $C = 320$  and  $S_N = 7$  through the computer simulations.

When the system operates in a recognition mode, an input digit string  $W_j$  represented by a series  $\{x_1, x_2, \dots, x_{T_j}\}$  of  $T_j$  LSP, energies and LSP mean compensated parameters, is computed by the two path Viterbi decoders as shown in Fig. 1. Both grammar free (GF) and grammar constraint (GC) methods can be selectively used in the Viterbi decoding algorithm, which depends on known input signal length.

The length of the input strings can be known, or unknown. For an unknown input string, we use the GF decoding algorithm and for known input string, we use the GC decoding algorithm. For the telephone dialing task, the useful string number of the digits can be: 1, 3, 4, 5, 7 and 10. The results of comparing GF/GC decoding algorithm based on 1-best is shown in Figure 2. In this case, the system is set  $C = 320$  and  $S_N = 7$ . We also compared the system based on 1-best and 2-best decoding algorithm and the results are shown in Figure 3 with  $C = 1280$  and  $S_N = 8$ . In order to see the difference of the systems MQ/HMM\_VQ/HMM and VQ/HMM, they were compared and the results shown in Figure 3 as well. The performance results with different  $C$  show in Figure 4.

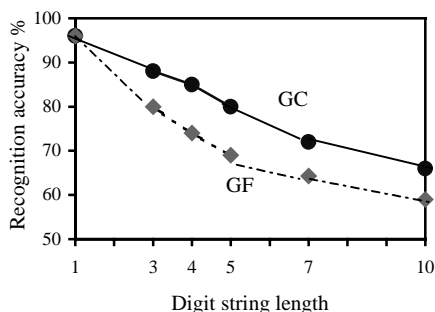


Fig. 2 The performances of the system using GC decoding vs. GF decoding

It is clear that dual MQ/HMM and VQ/HMM decoding pair algorithm can enhance the performance of the system VQ/HMM about 2-4%. Also it is better when using a grammar constraint algorithm, which provides a more constrained grammar composed of a sequence to more accurately trace back digit strings, especially in the case when the length of input string is long. Meanwhile, by using 2-best decoding algorithm, the recognition accuracy of the digit string recognition can be increased about 3%, compared with 1-best decoding algorithm. Furthermore, we can see that increasing the codebook size can improve the performance of the recognition system and the best numbers of  $C$  and  $S_N$  for the proposed connected digit recognition system is  $C = 1280$  and  $S_N = 8$ . We plan to investigate the compensation methodologies such as [1][3].

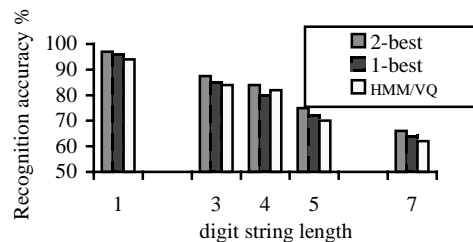


Fig. 3: The performances of the system using GC decoding

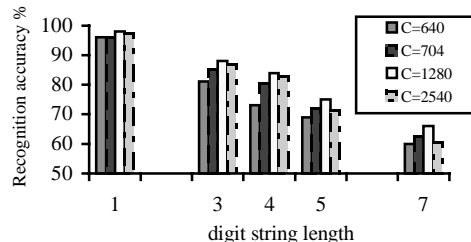


Fig. 4: The performances of the system using GC decoding

## 6. CONCLUSIONS

The robust speaker-independent connected digit speech recognizer proposed in this paper combined the dual MQ/VQ pair with discrete HMMs in a very simple and elegant way, which requires a small amount of ROM and low computational expenditures. Meanwhile, both of the grammar free and constraint decoders can be used in the system. The methodologies we present provide high accuracy and robustness of the system in a car noise environment and efficient use of processing resources.

## REFERENCES

- [1] C. H. Lee, "Adaptive Compensation for Robust Speech Recognition", Proc. IEEE workshop on Automatic Speech Recognition and Understanding (ARSU), p357-364, 1997.
- [2] S. Furui and M. M. Sondhi, "Advances in Speech Signal Processing", Marcel Dekker, Inc. 1992.
- [3] M. Rahim and L. Saul, "Minimum Classification Error Factor Analysis for Automatic Speech Recognition", Proc. IEEE workshop on ASRU pp.172-178, 1997.
- [4] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Application Speech Recognition", Proc. IEEE Vol. 77, pp257-286, 1989.
- [5] C. Myers and L. R. Rabiner, "Connected digit recognition using a level building DTW algorithm", IEEE Trans. ASSP, vol. 29, no. 3, pp.351-363, 1981.
- [6] R. Schwartz and S. Austin, "Efficient, High-Performance Algorithms for N-Best Search", Proc. DARPA Speech and Natural Language Workshop, pp.6-11, 1990.