A ROBUST SPEAKER INDEPENDENT ISOLATED WORD RECOGNIZER OVER THE TELEPHONE NETWORK BASED ON A MODIFIED HMM APPROACH

J. M. Song and A. Samouelian*

Speech Technology Research Group
Department of Electrical Engineering
The University of Sydney

*Speech Technology
Technology Development Group
International Business Unit, OTC Australia

ABSTRACT - This paper presents an accurate and robust, speaker independent isolated word recognition system based on continuous hidden Markov modelling. By using the state of the art techniques, we are able to achieve speech recognition performance of 97.9% testing on 20 males speakers over public switched telephone network (PSTN). In order to enhance the robustness of the system under noisy conditions, we explore a modified Gaussian pdf by using vector projection approach and improve the recognition performance from 88.8% to 92.1% at 10dB SNR (Gaussian white noise).

INTRODUCTION

In recent years, speech recognition technology has advanced to such a point where real time implementation of speech recognition systems for public use over PSTN is now possible.This paper presents results of research and development on a real-time speaker-independent isolated word speech recognition system operating over PSTN. The database used in this task is introduced. A front-end based on MFCC and delta MFCC is discussed. The model training and related problems are then developed. The recognition system and different schools of thought on system optimization are presented. A modified mixture of Gaussian pdf implemented in the recognition system and the experiment results are also reported.

DATABASE DESCRIPTION

The vocabulary under investigation consists of 19 words, i.e. twelve digits from *zero* through *nine* plus *oh* and *nought*, along with 7 command words, i.e. *cancel, help, next, start, stop, yes* and *no*. 105 speakers participated in speech data collection, each one pronounced the whole vocabulary just once. These speakers are from different nationalities, about 65% of them are Australian origin, and the rest are of different language backgrounds. A commercial telephone interface card on a PC is used to collect speech data using 8-bit μ-law encoding protocol. We used data from 85 speakers for training, and remaining 20 speakers for testing. In order to test the robustness of the recognizer under noisy environment, we added zero mean white Gaussian noise onto the 20-speaker testing database and create two additional databases with SNR levels of 15dB, 10dB respectively.

ACOUSTIC FEATURE EXTRACTION

A pertinent scheme of acoustic feature extraction is the first important stage in any speech processing system. The important considerations of choosing an adequate front end for speech recognition within the HMM framework can be summarized as follows:

- capturing the message-related spectral/temporal information, while reducing the irrelevant information, such as environmental and personal characteristics resided in the original speech.
- making use of human auditory characteristics to enhance the relevant message-related information.

- transforming the extracted acoustic feature vector to better satisfy the assumptions required by the HMM approach.

In our speech recognition system, incoming speech data is µ-law coded and we convert it back to linear representation with 14 bits. We then use first order pre-emphasis filter of $1\text{-}0.95z^{-1}$ and apply a 256 point hamming window to speech signal. Frame size is 256 points (32 ms), and frame shift is 80 points (10ms). After hamming windowing, FFT operation is applied to transform speech data into power spectrum. Most of the computation required by the front end is in the FFT, therefore we split the 256 real points FFT to 128 complex points and reduce the FFT computation by half. A set of 19 mel scale distributed triangular filters are imposed to the FFT power spectrum. The output of 19 spectral values is further transformed by discrete cosine transform to 12 mel frequency cepstral coefficients (MFCC), i.e

$$\hat{c}(i) = \sum_{j=0}^{N-1} \log S_j \cos\left(i\frac{(2j+1)\pi}{2N}\right) \qquad 1 \le i \le 12$$

where S_j is the energy output from j^{th} mel filter imposed on power spectrum.The energy term is discarded.

In order to enhance the distortion measure used in the recognition system, a raised sine function is used to weight the acoustic features, i.e.

$$W(i) = 1 + \frac{D}{2}\sin\left(\frac{\pi i}{D}\right) \qquad c(i) = \hat{c}(i)W(i) \qquad 1 \le i \le 12$$

We use delta MFCC to capture the transitional spectrum pattern which is important, particularly for speaker independent speech recognition. The delta MFCC is calculated in the following way.

$$\Delta c_t(i) = G\sum_{k=-2}^{2} k c_{t+k}(i) \qquad 1 \le i \le 12$$

The value of G is chosen to make the variances of MFCC and delta MFCC equal.

Augmented MFCC plus delta MFCC are used as acoustic feature vector, i.e. $O_t = \{c_t(i) \dots \Delta c_t(i)\}$

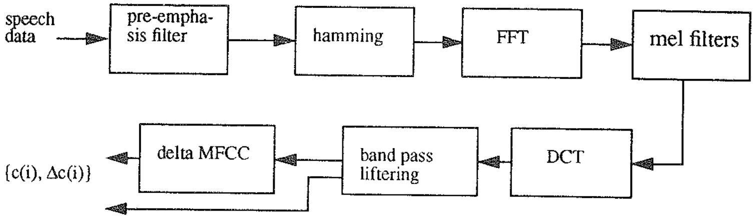

Figure 1. The block diagram of the front end used in the recognition system

HMM TRAINING

Initial HMM Training

The HMM topology is of five states left-to-right model without skip transitions, the type of pdf is of 3-mixtures of Gaussian pdf with diagonal covariance matrix. The choice of the HMM specifications is actually a trade-off between recognition performance, the amount of training data available and the real-time processing requirement. In order to compensate the insufficient training database to some degree, a flooring technique is imposed on all covariance matrices.

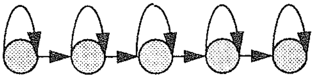

Figure 2. Five state left-to-right Markov chain

Since all training speech files have some silence at the beginning and end of each spoken word, we select a small number of speech files in the training database and manually segment speech signal using *xwaves* facilities. Then segmental-k means algorithm (Rabiner, 1989) is used to produce an initial set of word models for this small portion of data. Each sequence of acoustic feature vectors is linearly distributed onto 5 states. The optimality criterion of the segmental-k means approach is to maximize the conditional joint probability of state sequence and acoustic events conditioned on an initial model, i.e.

$$\lambda = argmax_{\lambda} \{max_S P(O, S| \hat{\lambda})\}$$

Automatic Segmentation

Once we have produced an initial set of models, we perform an automatic segmentation approach onto the whole training database based on this set of models. The mechanism of this automatic segmentation is to force the recognition results to the desired one for each training speech input, which usually has silence at beginning and end. By doing this the actual word boundaries can be located by back tracking procedure. The following network depicted in the Figure 3 is used in the automatic word segmentation.

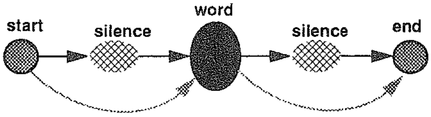

Figure 3. A composed HMM topology embeds silence model with word model

Since in general we do not know for sure whether there are silence portions in each speech file or not, we should allow a transition from starting node to both the first state of silence/noise model and the first state of the word model. Similarly, we allow the transition from the last state of word to the ending node of the network. Both starting node and ending node are dummy states, the dashed line transitions are null transitions.

Model Refinement

When the word boundaries have been decided through automatic segmentation, we use the Baum-Welch algorithm to re-estimate each word model based on the segmented word tokens and initial set of models. This method gives a best model estimation in the sense of maximum likelihood with respect to the initial model and segmented word tokens, i.e. the model estimated maximizes a prior probability,

$$\lambda_{new} = argmax_{\lambda_{old}} P(O| \lambda_{old})$$

Once the updated new set of models have been generated, we iterate the automatic segmentation and Baum-Welch re-estimation procedure again. This procedure is shown in the Figure 4. Only 3 iterations are needed to produce a good set of models during training. In fact, we find that over running the iterations will make the word models over-tuned towards to the particular training database, and therefore reduce the generalization for unseen speech data in real speech recognition.

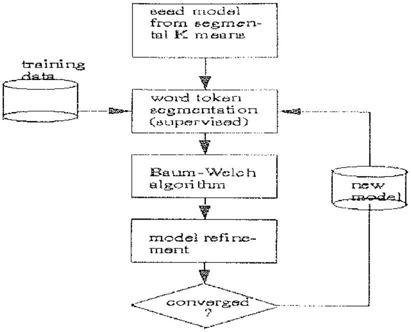
seed model from segmental K means
training data
word token segmentation (supervised)
Baum-Welch algorithm
new model
model refinement
converged ?

Figure 4. The block diagram of iterative automatic word model training procedure

SPEECH RECOGNITION ALGORITHM AND EXPERIMENT RESULTS

Embedded Silence/noise Absorbing

The accuracy of end-point detection is important in the isolated word recognizer. The approaches based on the information such as energy and duration are not robust when SNR becomes less than 30 dB. In our speech recognition system, two steps of noise/silence detection have been implemented. The first step gives a simple end-point detection for incoming speech (Hunt *et al*, 1992). The second step is an implicit end-point detection mechanism created by a composed HMM network in which the silence/noise model is concatenated at the beginning and ending of each word model. The topology of this composed model is similar with the structure shown in the Figure 3. When the silence/noise is at the beginning or end of an utterance, it will be located at the initial or final states of the composed HMM. This composed HMM network assimilates the silence/noise part of the incoming signal in an efficient way, and it is more robust and accurate than ad hoc end-point detection algorithms. Considering that the characteristics of silence/noise are typically stationary, we only use 2 states to represent its corresponding HMM model.

Recognition Algorithm

We apply a variant of synchronous beam searching Viterbi algorithm for isolated speech recognition. The database used for testing consists of 20 speakers collected through PSTN in an office environment, where each speaker speaks the 19-word vocabulary once. By using the optimal set of refined models, recognition performance of 97.9% is achieved. Further analysis on the mis-recognized speech files reveals that some speaker pronounce words with a strong foreign accent, for example, word *nought* is sometimes pronounced like word *no*. How to enhance the recognition performance is still an open question from acoustic pattern matching point of view. However, most of these errors should be avoided by incorporating a simple language model into recognition system.

In order to accelerate the recognition process while still keeping the performance at the same level, two techniques have been implemented in our system.

The first one is a beam searching technique, i.e. for each frame of speech, we select the maximum likelihood among all active states of all models, denoted as δ_{max}, if the likelihood associated with a particular state of any HMM is δ_{state} meets the condition: $\delta_{state} < \delta_{max}$ - threshold, this state will be deactivated and will not be dealt with at the next frame. The threshold value can be either fixed or computed at each frame by measuring the value of δ_{max}.

The second technique implemented in the recognizer is to select the pre-dominant pdf component from the mixtures of Gaussian pdf at each time for each state and each model, i.e. for the pdf representation

$$b_j(O) = \sum_{k=1}^{M} c_{jk} b_{jk}(O_t)$$

we approximate $b_j(O_t)$ by the principal value in the mixture, i.e. $\hat{b}_j(O) = max_k c_{jk} b_{jk}(O_t)$

In fact, this approximation is reasonable from semi-continuous HMM point of view, which partitions the whole acoustic space by a set of overlapped Gaussian distributions (Huang 1992). The experimental result shows that the recognition performance only degrades slightly compared with using the summation of mixtures.

Tied Covariance Matrices for HMM

We conducted a series of tests on the tied covariance matrices of HMM models. The first investigation used a single covariance matrix for every state and every model. The second investigation used a single covariance matrix for every state of a model, and a different covariance matrix for a different model. The tied covariance matrices are obtained from the corresponding pooled speech database. Table 1 shows the recognition results of using standard HMM and different tied covariance matrices. This shows there is some performance degradation by using tied covariance for this training database. However, the storage and computation requirement can be reduced significantly by using tied covariance.

standard models	single Covar for all models	single Covar for each model
97.9	96.4	96.5

TABLE 1. Recognition performance with and without tied covariance matrices.

A Robust Likelihood Measure

It is well known that speech recognition performance degrades rapidly when the environment become very noisy. Typically, recognition rate could drop from 98% when testing condition is matched with training condition down to 30% percent when these two conditions are not matched. This problem can be tackled from different aspects. We adopted a projected distance measure in the pdf calculation (Carlson and Clements, 1992), instead of using standard pdf, i.e. we replace the following calculation

$$b_j(O_t) = \sum_{k=1}^{M} c_{jk} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} exp[-0.5 \times (O_t - \mu_{jk})^T \Sigma_{jk}^{-1} (O_t - \mu_{jk})]$$

by a projected version

$$b_j(O_t) = \sum_{k=1}^{M} c_{jk} \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} exp[-0.5 \times (O_t - \lambda_{jk}\mu_{jk})^T \Sigma_{jk}^{-1} (O_t - \lambda_{jk}\mu_{jk})]$$

Where O_t is the acoustic vector and D is its dimension, μ_{jk} and Σ_{jk} are mean vector and diagonal covariance matrix of k^{th} components at state j respectively, T denotes transpose operation.The value of λ_{jk} is determined based on the orthogonal principle in the MFCC space shaped by Σ_{jk}, i.e.

$$\lambda_{jk} = \frac{O_t^T \Sigma_{jk}^{-1} O_t}{\mu_{jk}^T \Sigma_{jk}^{-1} \mu_{jk}}$$

To justify the use of this new likelihood measure, we conducted a set of tests on the 20 male speaker database discussed above. The noisy speech is simulated by adding zero mean white Gaussian noise to the original test database. Two noisy conditions have been tested, i.e. 15 and 10dB global SNR, together with the original test speech. The experimental results obtained from these four different conditions are shown in the Table 2. Only the original test database is matched with the training data condition.

SNR (dB)	original	15	10
Standard	97.9	93.5	88.8
Projection	97.9	94.7	92.1

TABLE 2. Open test using standard pdf and projected pdf for different levels of SNR

CONCLUSION

We have discussed different aspects in the design of an accurate, robust and fast speaker independent isolated word recognition system in this paper. We have optimized the performance of the speech recognition system and implemented a modified Gaussian pdf calculation by using vector projection. By using the techniques presented in this paper, we achieved 97.9% performance on test database of 20 male speakers, and 94.7%, 92.1% on the white Gaussian noise contaminated test data for SNR of 15dB, 10dB respectively. The results from a series of experiments show that a good performance of recognition system over telephone network is achievable, although many challenging problems remain to be overcome.

ACKNOWLEDGEMENT

This work was carried out under contract to the Speech Technology, Technology Development Group, International Business Unit, OTC Australia, 255 Elizabeth Street, Sydney, NSW, 2001, Australia. The permission of the Managing Director of OTC Australia to publish this work is gratefully acknowledged. The views and opinions expressed in this work are those of the authors and do not necessarily imply OTC policy or further service offerings.

REFERENCES:

B. Carlson, M. Clements, (1992), *Speech recognition in noise using a projection-based likelihood measure for mixture density HMM's*, Proc. ICASSP 92, pp. 237-240.

A. J. Hunt, P. C. B. Henderson, A. Samoulien, J.M.Song, and R.W.King, (1992), *Engineering a speech controlled voice-mail demonstration system operating on the telephone network*, elsewhere in this Proceedings.

X. D. Huang, (1992), *Phoneme Classification using semicontinuous hidden Markov models*, IEEE, Trans. Signal Processing, Vol. 40, No. 5, May 1992, pp. 1062-1067.

L. R. Rabiner, (1989), *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceeding of the IEEE, Vol. 77, No. 2, Feb. 1989, pp. 257-285.