Proceedings of the Institute of Acoustics

SPEAKER VERIFICATION USING CONNECTED WORDS

Michael J Carey and Eluned S Parris

Ensigma Ltd., Chepstow, Gwent

1. INTRODUCTION

This paper describes the improvements made to a speaker verifier [1] for use in a commercial system. Speaker verification is performed using competing Hidden Markov Models (HMMs) in which one model represents the population in general and the other represents the speaker to be verified. The system was designed to be used over the telephone network and therefore had to be insensitive to level, line distortions and independent of the handset used. The storage for the user date, ie. HMMs, was required to be less than 1 kbyte. The system needed to be 'user-friendly' and therefore connected words and a fast training algorithm were used.

2. SYSTEM DESCRIPTION

The previous system [1] operated on isolated utterances. For each utterance spoken, the output of two Markov models was compared. One model was derived from utterances of the speaker to be verified (the personal model) and the other model was derived from utterances of a sample of the population in general (the world model). The verification test sore for each model was output as a log-likelihood measure and the difference in scores compared with a threshold. Verification was carried out over five isolated words and if the difference in scores exceeded the threshold for the majority of utterances, the user was accepted.

In the improved system, speaker verification is performed on connected five digit strings by comparing the output probability of two concatonated Markov models of the digit string (see Figure 1).

The first Markov model is speaker specific and referred to as the personal model. The model is built from utterances obtained during training from the speaker whose identity is to be verified. The second Markov model is speaker independent and is built from embedded utterances from a large population of speakers. The operation of the system uses the fact that the likelihood of the personal model given the utterance will be greater than that of the world model when a true speaker verifies, and vice versa when an impostor verifies.

Each digit is represented by a left to right seven state HMM in which only transitions to the same or next state are allowed. The observation probability for each state is described by a continuous density probability distribution. The covariance matrix of the probability distribution is assumed to be diagonal. The personal models are represented by a single mixture density and the world models by three mixture densities.

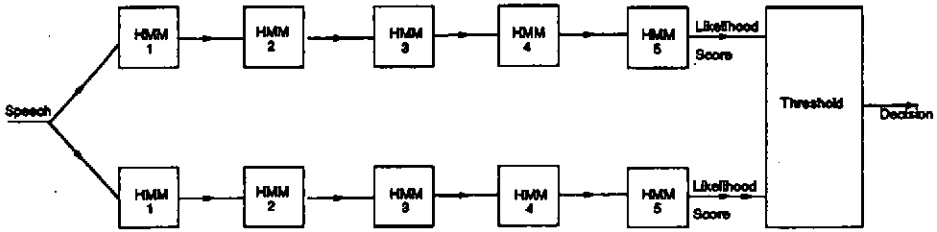

Figure 1 Verification System

The input speech to the system is sampled at a rate of 8 kHz and analysed using an eleven filter version of the RSRE Filterbank, SRUBANK [2]. In this configuration the centre frequencies of the first five filters are linearly spaced up to 1 kHz and the remaining six are logarithmically spaced. The log-power outputs of the filterbank are transformed using a discrete cosine transform to give the mel cepstrum of the speech at a frame rate of 20 ms.

Mel cepstrum and their time derivatives have been successively used in speaker independent word recognition in noisy conditions [3]. It was found in our verification system that the first and second derivatives of cepstrum were invariant to the handset and line and enhanced the system performance. The static mel cepstrum were not invariant, degraded performance and therefore were not used for telephone applications.

The feature vector consists of fourteen parameters :

1. six estimates of the time derivative of cepstrum
2. time derivative of the energy

3. six estimates of the second derivative of cepstrum

4. second derivative of the energy.

The second derivatives 1) and 2) are calculated using

$$\Delta\Delta(t) = \Delta(t+1) - \Delta(t-1)$$

where $\Delta(t+1)$ and $\Delta(t-1)$ are the time derivatives of cepstrum at time $(t+1)$ and $(t-1)$ respectively.

The system runs in real-time on a DSP32C on a PC plug-in card connected via a PABX. Both training and verification are performed on-line either over the telephone network or locally.

3. TRAINING

During training the user is required to speak twelve connected five digit strings. The strings are designed to capture the coarticulation effects that can occur when digits are strung together, eg. 'zero one'.

Each training string is automatically end pointed into digits by matching the speech to a speaker independent model of the string, the word model, and finding the best fitting position of each digit. If the probability of any digit is below a predetermined threshold, the user is asked to repeat the string. At the end of training a Viterbi re-estimation process is used to build an HMM for each digit. The total time taken for training is usually less than one minute.

User specific thresholds are set by examining the variability of the HMMs produced. Users who have shown little variability during training would have more difficulty verifying as any variation in voice at a later date would produce poor matches to the models. For this reason, more lenient thresholds are given for verification. Impostors to the system also tend to match more easily to users who have shown greater variability in training, justifying tighter thresholds for more variable users.

The total storage required for a user's models is reduced to less than 1 kbyte by scalar quantization. The mean and variance of each of the fourteen parameters is scaled differently to optimize performance.

4. VERIFICATION

During verification, the user is asked for their claimed identity. The user is then required to speak a five digit string chosen at random to reduce the possibility of fraud. The utterance spoken is matched to both the claimed user's models and a set of speaker independent models. Likelihood scores are calculated for the utterance being generated by the personal model and world model from a frame synchronous Viterbi search.

The user is accepted as the true speaker if

$$L_p - L_w > T$$

where L_p is the likelihood score for the personal model,
L_w is the likelihood score for the world model,
T is the user specific threshold set during training.

If the test fails then the user is required to speak a different string. If the test is passed then the user is accepted to the system, otherwise rejected.

The level of security of the system can be changed by adjusting the value of the threshold. A high level of security can be set by increasing the threshold. This makes it more difficult for a true user to be accepted but far more unlikely that an impostor would be accepted. For similar reasons, a low level of security can be set by reducing the threshold.

5. PERFORMANCE

An in-house trial was set up with ten speakers (five male, five female) using the system via a PABX connection over several weeks. Each user trained at the start of the trial and the same models were used for all future verifications. Each user verified one hundred times against their own models and ten times against every other user's models.

Table 1 shows the number of false rejections and false acceptances for each speaker when a single string was used in verification. An equal error rate of 1.4% was achieved. Table 2 shows the number of false rejections and false acceptances for each speaker when two strings were used in verification. An equal error rate of 0.4% was achieved.

Speaker	False Acceptances (%)	False Rejections (%)
1	0.0	1.0
2	0.0	0.0
3	0.0	2.0
4	0.0	0.0
5	2.2	0.0
6	0.0	2.0
7	1.1	0.0
8	5.6	3.0
9	1.1	4.0
10	4.4	3.0
Average	1.4	1.5

Table 1: Single String Results

Speaker	False Acceptances (%)	False Rejections (%)
1	0.0	0.0
2	0.0	0.0
3	0.0	3.0
4	0.0	0.0
5	0.0	0.0
6	0.0	0.0
7	0.0	0.0
8	2.2	1.0
9	0.0	0.0
10	2.2	1.0
Average	0.4	0.5

Table 2: Double String Results

6. CONCLUSION

This paper has described a speaker verification system which is robust to telephone line and hand set variations for remote or local verification. The HMM parameters have been efficiently quantized to less then 1 kbyte. Verification equal error rates of 1.4% for a single string and 0.4% for a double string have been achieved.

7. REFERENCES

[1] M J Carey, E S Parris and J S Bridle, 'A Speaker Verification System Using Alpha-Nets', ICASSP 1991, Toronto, pp. 397-400.

[2] D W Knox, SRUBANK Users Manual', Ensigma, Chepstow, 1987.

[3] B A Hanson and T H Applebaum, 'Robust Speaker-Independent Word Recognition Using Static, Dynamic and Acceleration Features: Experiments with Lombard and Noisy Speech', ICASSP 1990, Alhuquerque, pp. 857-860.