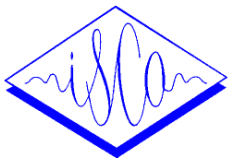

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Experiments in Automatic Talker Verification Using Sub-Word Unit Hidden Markov Models

Aaron E. Rosenberg, Chin-Hui Lee, Frank K. Soong, and Maureen A. McGee

Speech Research Department
AT&T Bell Laboratories
Murray Hill, NJ 07974 USA

ABSTRACT A talker verification system based on characterizing talker utterances as sequences of sub-word units represented by Hidden Markov Models (HMM's) has been implemented and tested. Two types of sub-word units have been studied, phone-like units (PLU's) and acoustic segment units (ASU's). PLU's are based on phonetic transcriptions of spoken utterances and ASU's are extracted directly from the acoustic signal without use of any linguistic knowledge. The ASU representation has the advantage of not requiring transcriptions of training utterances. Verification performance has been evaluated on a 20-talker database of isolated digit utterances and a 20-talker database of continously spoken sentences drawn from a 1000-word vocabulary. In the isolated digit experiments the verification equal-error rate is approximately 7 to 8% for 1-digit test utterances (approximately 0.5 sec in duration) and 1% or less for 7-digit test utterances (approximately 3.5 sec in duration) with only small differences in performance between PLU- and ASU-based representations. In the continuously spoken sentences experiments using ASU's the best verification performance is 1.7% equal-error rate for 5 second test trials. This is obtained using 64 ASU models trained from 90 seconds of speech. In addition, a technique for updating models, using data from current test utterances, has been devised and implemented. Using this adaptation technique for isolated digits, the error rate falls to 6% for 1-digit utterances and less than 0.5% for 7-digit utterances. The experiments show that excellent verification performance can be obtained with sub-word units represented by HMM's. The techniques can be readily extended from small vocabularies and isolated words to large vocabularies and connected sentences.

1. Introduction

Experiments have been carried out to evaluate the performance of a Hidden Markov Model- (HMM-) based talker verification system which models sub-word units. There are significant potential benefits associated with using sub-word unit models to represent talkers for talker verification. With a complete set of sub-word unit models for each speaker, verification can be carried out using utterances drawn from arbitrary, unlimited vocabularies. In a text dependent mode, random test utterances can be prescribed and compared with models of the utterances which consist of concatenated sub-word units composed according to the specifications of a lexicon. Enhanced security is obtained by prescribing utterances which may never have been used before in verification tests. In a text independent mode, where test utterances are not prompted or known in advance, utterances can be considered to be composed of sequences of sub-word units occurring with equal probabilities or with probabilities reflecting general statistics of the language. Thus, with this approach, text dependent and text independent verification can be carried out by the same system based on representing talker utterances as sequences of sub-word units. Although most text dependent verification systems have been based on templates or models of utterances, usually words or phrases, text independent verification systems have generally operated in a fundamentally different manner. Such systems have been based on prototype measurements of long-term statistics of spectral and other signal features [1,2]. This approach lacks the detailed representations that are possible by modelling individual instantaneous speech events, such as sub-word segments, where temporal information is preserved.

Two types of sub-word units have been investigated, phone-like units (PLU's), based on phonetic transcription of utterances, and acoustic segment units (ASU's) based directly on signal properties of the utterances [3]. The PLU representation requires phonetic transcriptions of training utterances in order to to provide an initial segmentation and labelling of the training utterances for the calculation of HMM's. Also, a sufficient number of tokens of each PLU must be available in the training data in order to create an adequate set of models. The ASU representation does not require transcriptions of the training data nor is it necessary to pay special attention to the linguistic content of the data provided a reasonably large and phonetically balanced set of utterances is included. However, to create an ASU lexicon of test utterances, one or more tokens of each utterance is required for each talker since there is no simple way to obtain the ASU representation of an utterance without actual tokens.

The use of HMM's to model subword units offers the same advantage for talker verification that it does for speech recognition. This is the ability to design remarkably robust models for these subwords units without the necessity of providing detailed labelling and segmentation information of training utterances.

2. Overview of Experiments and Talker Verification System

Two sets of experiments have been carried out. In the first set the database consists of isolated digit utterances recorded over dialed-up telephone lines by 20 talkers, 10 male and 10 female. Each talker recorded 200 digit utterances, 20 tokens of each digit, in 5 recording sessions held over a period of up to two months. The recordings were made in a sound booth using an ordinary carbon button telephone handset. Both PLU's and ASU's have been used in these experiments. Preliminary results of these experiments have been previously reported [4].

In the second set of experiments the database consists of approximately 40 read sentences recorded by 20 talkers, 9 female and 11 male. The recordings were made in a sound booth using a noise cancelling microphone. These sentences are part of the "speaker independent training" section of the Naval Resource Management database recorded for DARPA [5]. The sentences are composed of words drawn from a vocabulary of 991 words designed to make queries in the Resource Management task. This database is referred to as the RM sentences. The amount of speech recorded by each talker ranges from approximately 110 sec to 190 sec. Only ASU's have been used in experiments with this database. PLU experiments cannot be carried out because there is not enough training data to obtain a complete set of models for each talker.

2.1 Model Parameters

Each sub-word unit is represented by a left-to-right HMM, containing either 2 or 3 states. Fig. 1 illustrates a typical 3-state model.

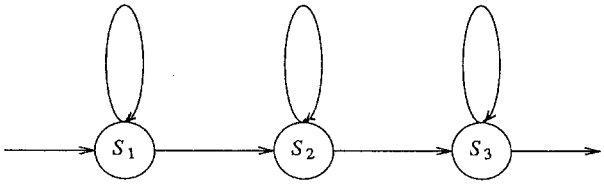

Figure 1. 3-state left-to-right HMM

The spectral observation probability for each state is characterized by a continuous probability density function specified as a mixture of Gaussian densities. Thus, the probability of observing spectral vector $\mathbf{O}_t$ at frame t in state j is given by

$$b_j(\mathbf{O}_t) = \text{Prob}(\mathbf{O}_t \mid \text{state } j) = \sum_{m=1}^{M} c_{jm}\mathbf{N}(\mathbf{O}_t;\mu_{jm},\mathbf{U}_{jm})$$

where $\mathbf{N}$ represents a multivariate Gaussian probability density function with mean vector μ_{jm} and (diagonal) covariance matrix $\mathbf{U}_{jm}$ for the m-th component of the M-component mixture, and c_{jm}, $m=1,...,M$ are the set of mixture weights. The maximum number of mixture components, M, varies from 1 to 3 in these experiments. In addition, HMM's are characterized by a set of transition probabilities from state i to state j, a_{ij}, where $j-i$ can be 0 or 1.

2.2 Training

The model parameters are estimated by means of a modified segmental k-means training procedure [6] using a set of training utterances. The procedure is given in the block diagram in Fig. 2.

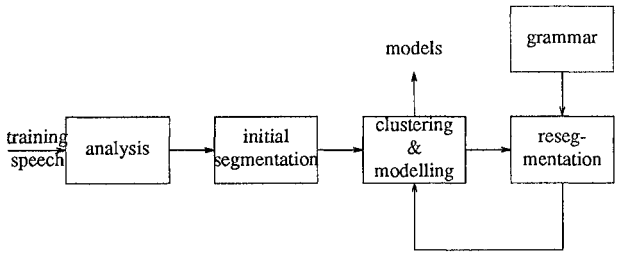

Figure 2. HMM training procedure

Following an initial segmentation and labelling of training utterances into units and states, all frames corresponding to state j of a particular unit are partitioned into M clusters using a standard vector quantization (VQ) clustering technique. Estimates of μ_{jm}, $\mathbf{U}_{jm}$, and c_{jm} are obtained. The training utterances are then resegmented and labelled using a Viterbi decoding technique based on the current models. Resegmentation may be guided by text dependent constraints contained in a "grammar". Clustering and model reestimation, followed by resegmentation, are iterated until the average model likelihood stops increasing. The way in which initial segmentations are made and the choices of text dependent constraints are described in the following sections. The number of training iterations generally varies from 4 to 8.

2.3 Initial Segmentations

Initial segmentations for training are obtained as follows. For the PLU-based system used in the experiments with isolated digits the initial segmentation and labelling is simply a linear partitioning of each training utterance into units and states. The unit label sequences are obtained from phonetic transcriptions of each vocabulary word. A total of 20 PLU's are needed to transcribe the vocabulary of 10 digits.

The initial segmentation and labelling for ASU's is obtained using the following maximum likelihood segmentation technique [7]. Assume it is desired to segment an utterance of T frames into L segments. A set of segment boundaries, t_l, $l=1,2,..,L$, is obtained such that the global distortion measure

$$D = \sum_{l=1}^{L} \sum_{t=t_l}^{t_{l+1}-1} d(y_t, c_l)$$

is minimized, where y_t is the spectral vector for the t-th frame, c_l is the centroid of vectors in the l-th segment, and d is the local spectral distance or distortion. Alternatively, instead of specifying the number of segments desired, a distortion threshold, $D_{\max}$, can be specified to obtain the smallest number of segment boundaries such that D does not exceed $D_{\max}$. After segment boundaries are obtained, segment labels are found by clustering all segments into a VQ codebook of a desired size. Labels are assigned to each segment associated with the best matching codebook entry. For the isolated digit experiments a codebook size of 16 has been chosen. This size is commensurate with the number of phonetic units for this vocabulary. For the RM sentences experiments, 2 codebook sizes have been used, 32 and 64. For the isolated digit experiments both segmentation specifications have been tried. In the first, the number of segments per word is specified to approximate the number of phones per word. In the second, a distortion threshold is selected for which 4 or 5 segments per word are obtained. Experimental results are essentially the same for these two specifications. For the experiments with RM sentences, a distortion threshold has been chosen so that approximately 12.5 segments per second of speech are obtained, corresponding to an average segment duration of 80 msec.

2.4 Testing

Verification tests are carried out by comparing test utterances with reference talker models. A verification test score is output as a likelihood score from a frame-synchronous Viterbi search [8]. For text dependent experiments the search takes place under constraints provided by a "grammar" using specifications provided by a lexicon. For text independent experiments the search treats all units as occurring equally likely. For each reference talker, a set of "true" talker test scores is obtained by comparing the talker's test utterances with his or her models. "False" talker test scores are obtained by comparing other talkers' test utterances with the reference talker's models. No *a priori* verification thresholds are used. Instead, equal-error rate estimates are calculated from the set of "true" talker and "false" talker test scores obtained for each reference talker. In both sets of experiments the performance figures are estimated equal-error rates averaged over 20 talkers.

3. Isolated Digit Experiments

3.1 Analysis

The database utterances are bandpass filtered from 200 to 3200 Hz and sampled at a rate of 6.67 kHz. The digitized speech signal is preemphasized using a first order digital network and an 8-th order autocorrelation analysis is carried out over blocks of 300 samples (45 msec), applying a Hamming window, shifted every 100 samples (15 msec). Each frame of autocorrelation coefficients is converted to a linear predictive coding (LPC) derived set of 12 cepstral coefficients. In addition, a set of delta cepstrum coefficients [9], which are estimates of the time derivatives of the cepstral coefficients, are used to augment the analysis vector to 24 coefficients per frame. The delta cepstrum coefficients are calculated by fitting a regression line to a sequence of 5 cepstral coefficient vectors centered around the current vector. Both the cepstral coefficients and delta cepstrum coefficients are weighted using a sinusoidal "lifter" [10].

3.2 Experimental Setup

The first 80 utterances for each talker, that is, 8 occurrences of each digit, (approximately 40 sec in duration), are designated as training utterances. The remaining 120 utterances are used as test data. Test trials consist of from 1 to 10 distinct digit utterances varying in duration, on average, from approximately 0.5 to 5 sec. The number of "true" talker test trials varies from 120 for 1-digit trials to 10 for 10-digit trials. Test utterances are used in the same order in which they were recorded. The number of "false" talker test scores varies from 19*120 for 1-digit trials to 19*12 for 10-digit trials.

3.3 Experimental Results

Fig. 3 shows results for the PLU-based system for both text dependent (solid lines) and text independent (dotted lines) modes for 4 combinations of model parameters. Results are shown as plots of estimated equal-error rate averaged over 20 talkers as a function of the length of test trials. Data points are plotted for 1 to 5, 7, and 10-digit long trials.

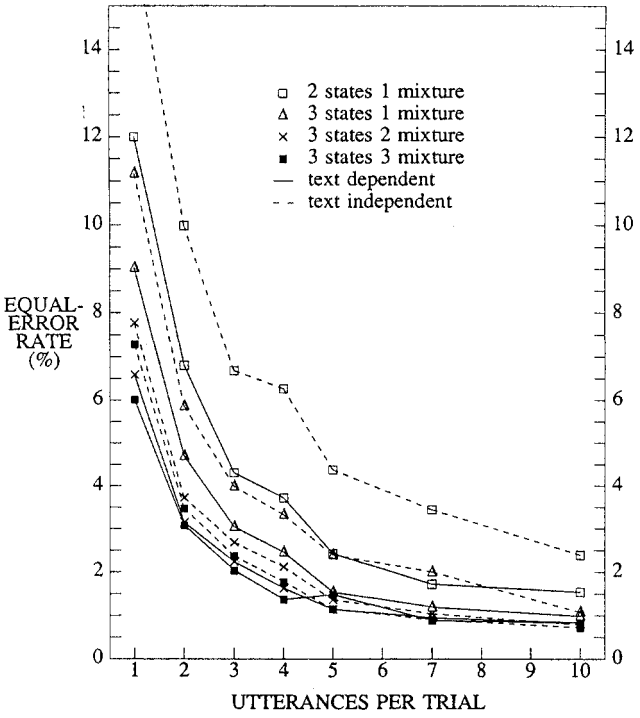

Figure 3. PLU verification results for isolated digits

The model parameter specifications relate to the amount of "spectral resolution" used to represent each PLU HMM. The amount of spectral resolution, as defined here, is the number of spectral vector centroids allotted to each unit. For example, if a model is specified by 3 states and 2 mixture components per state the spectral resolution is said to be $3 \times 2 = 6$ vectors per unit. Results for 4 values of spectral resolution, 2, 3, 6, and 9 vectors per unit, are shown in the figure.

In each plot it can be seen that equal-error rate decreases monotonically with the length of the test trial. This is a universal observation in talker verification experiments indicating that as the test trial length increases, the variance of true-talker test trial scores decreases providing sharper discrimination between the distribution of true-talker and false-talker scores. However, the performance levels off for longer test trials, indicating a residual variance attributable to correlation among test utterance samples. Generally, it can be seen that there is very little performance improvement for test trials longer than 4 or 5 digits (approximately 2.5 sec).

For each value of spectral resolution, performance with text dependent test trials is better than text independent performance. This is a natural and expected result, since constraining the sequence of units to be decoded to match the phonetic sequence of the input should provide higher discriminability between talkers. This is because the scores obtained from false-talker utterances matched with constrained models are poorer than those obtained for matches with unconstrained models.

Increasing spectral resolution implies providing more detailed and accurate models of sub-word units of talkers. In turn, this provides improved talker verification. However, increasing spectral resolution beyond approximately 6 vectors per unit does not provide any significant further improvement. Presumably, this is because the fixed amount of training data used cannot provide any additional useful spectral detail. It can also be seen that as spectral resolution increases the difference in performance between text dependent and text independent test modes becomes quite small.

Results for the ASU-based system are quite similar to results for the PLU's. The PLU and ASU results are summarized and compared in Table I.

units	training mode	test mode			
		TD		TI	
		number of utterances per trial			
		1	7	1	7
PLU	TD	6.02	0.88	7.29	0.90
ASU	TI	7.57	0.93	8.76	1.29
ASU/adaptive	TI	5.84	0.29	6.43	0.40

Table I. Summary of verification equal-error rates (%) for isolated digits.

Although the ASU performance is consistently slightly worse than the PLU performance, the difference can be interpreted in terms of the smaller amount of spectral resolution obtained for ASU's since there are 16 ASU's and 20 PLU's. Included in the summary are the results of ASU-based experiments in which talker models are adapted in a supervisory manner using data from talker test utterances. The error rates fall to approximately 6% and less than 0.5% for 1- and 7-digit test utterances, respectively.

4. RM Sentence Experiments

4.1 Analysis

The database utterances are bandpass filtered from 100 to 3800 Hz and sampled at a rate of 8 kHz. The digitized speech is preemphasized with a first order digital network and a 10th order autocorrelation analysis is carried out over 30 msec windows shifted every 10 msec. 12 LPC-derived cepstral coefficients are calculated as well as 12 delta cepstral coefficients in the same way described earlier for the isolated digits database.

4.2 Experimental Setup

The ASU HMM's are trained using either the first 60 or 90 seconds of each talker's utterances. The speech remaining for each talker after the first 90 seconds is used for testing. Although it would be useful to examine the effect of using more than 90 sec of data training, there would not be enough data remaining for each speaker for adequate testing. Test trials vary from 1 to 5 sec in duration. The number of "true" talker test trials varies from approximately 20 to 100 for 1-second test trials and from approximately 4 to 25 for 5-second test trials. The 20 talkers are divided into 2 groups of 10 talkers each. Each reference talker is compared with 9 "false" talkers from one of the groups.

4.3 Results

The results for the RM sentences are shown in Fig. 4. Four plots are shown in the figure giving the average estimated equal-error rate over the 20 talkers as a function of test trial duration for each combination of training conditions. As expected, the results show that improved performance is obtained with greater amounts of training. Also, for each of the 2 amounts of training data, better performance is obtained with 64 ASU's than for 32 ASU's. The plots suggest that the relative improvement in performance with 64 ASU's is greater with 90 sec of training than with 60 sec of training. This can be interpreted to mean that larger amounts of training data is more critical for providing good 64-ASU models than for 32-ASU models. In addition, the plots suggest that performance at longer test trials levels off faster for 60 sec of training than for 90 sec. The results are summarized in Table II. The best result is obtained with 90 sec training, 64 ASU's, and 5 sec test trials for which the equal-error rate is 1.7%.

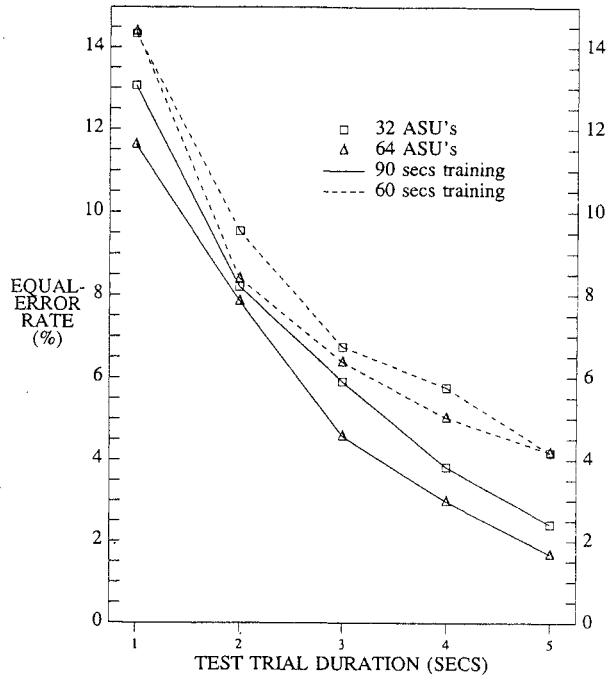

Figure 4. Text independent verification results for RM sentences

training utterance duration (sec)	number of ASU's			
	32		64	
	test trial duration (sec)			
	1	5	1	5
60	14.3	4.2	14.4	4.2
90	13.1	2.4	11.6	1.7

Table II. Summary of verification equal-error rates (%) for RM sentences

5. Discussion

Only a few experiments have been reported in the literature on the use of HMM's in talker recognition. The earliest reported experiment was described in a paper by Poritz [11] in 1982. Poritz represented each talker with a 5-state ergodic autoregressive model (that is, all transitions between states are allowed). He achieved good discrimination among a population of 5 talkers with 40 sec of training data per talker in a text independent mode. Tishby [12] expanded on Poritz's idea using 8-state ergodic autoregressive mixture models with 2 to 8 mixture components per state. Tishby carried out his experiments using the same speech database used in the current experiments, except that 100 utterances were used for training rather than 80. In order to compare Tishby's results with the current results, an ASU-based verification experiment was carried out using 100 training utterances. When the number of spectral vectors used to represent talker utterances is the same for both systems (set to 64) the equal-error rate for the ASU-based system is 1.1% for 7-digit long utterances compared to 2.0% for Tishby's system. The improvement obtained in the ASU-based system can be attributed to the introduction of more temporal detail and structure in the current talker models. With the ergodic model a global model of a talker's utterances is constructed with weak, general, temporal constraints. With sub-word models talker utterances are constructed as concatenations of sub-word unit models each with strict left-to-right temporal constraints.

In conclusion, the results indicate that talker verification based on HMM representation of sub-word units performs well and looks quite promising for extended applications of talker verification. For the isolated digits database, it has been shown that, adjusting for differences in spectral resolution, the results using PLU's and ASU's are quite similar. This indicates that segmentation and labelling based on different but reasonable criteria can produce equally good models and verification performance. It has also been seen that text dependent and text independent performance is about the same. This is probably a consequence of the small vocabulary and inventory of units for this database. Generally it can be expected that performance in the text independent mode will be worse than in the text dependent mode. This is because there are fewer constraints in the text independent mode allowing false talkers to obtain relatively better scores since they have more models to choose from. For the vocabulary of isolated digits, however, there are relatively fewer alternate models available and thus less opportunities for false talkers to obtain good scores in the text independent mode.

The text independent results with the RM sentences database show that the sub-word HMM approach can be extended to large vocabularies and continuous speech. Good results are obtained with relatively small amounts of training and test data. The results suggest that still better performance might be obtained with larger amounts of training and testing data, although the limitations of the current database prevented examination of those possibilities. In any case, the amounts of training and test data needed to obtain good verification performance is significantly less than what is generally required in the approaches to text independent verification based on long-term statistics. For example, in the Markel and Davis study [2], the best verification results, 4.25% equal-error rate, are obtained with approximately 1 hour of training data and 40 sec of test data. The current experiments should be extended to encompass conversational speech with unrestricted vocabularies and, in addition, text dependent input.

REFERENCES

[1] P.D. Bricker et al, "Statistical Techniques for Talker identification," *Bell System Technical Journal*, v. 50, pp. 1427-1454, 1971.

[2] J.D. Markel and S.B. Davis, "Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base," *IEEE Trans. on ASSP*, v. 27, pp. 74-82, Feb. 1979.

[3] C-H. Lee, F.K. Soong, and B-H. Juang, "A Segment Model Based Approach to Speech Recognition," *Proc ICASSP88*, v. 1, pp. 501-504, Apr 1988.

[4] A.E. Rosenberg, C-H. Lee, and F.K. Soong, "Sub-Word Unit Talker Verification Using Hidden Markov Models," *Proc. ICASSP 90*, v. 1, pp. 269-272, April 1990.

[5] P. Price, W. Fisher, J. Bernstein, and D. Pallet, "The DARPA 1000-Word Resource Management Database for Continous Speech Recognition," *Proc. ICASSP88*, v. 1, pp. 651-654, April 1988.

[6] L.R. Rabiner, J.G. Wilpon and B-H. Juang, "A Segmental K-Means Training Procedure for Connected Word Recognition," *AT&T Technical Journal*, v. 65, pp. 21-31, May/June 1986.

[7] T. Svendsen and F.K. Soong, "On the Automatic Segmentation of Speech Signals," *Proc. ICASSP87*, v. 1, pp. 77-80, Apr. 1987

[8] C-H. Lee and L.R. Rabiner, "A frame-synchronous network search algorithm for connected word recognition," *IEEE Trans. on ASSP*, v.37, pp. 1649-1658, Nov. 1989.

[9] S. Furui, "Cepstrum Analysis Technique for Automatic Speaker Verification," *IEEE Trans. on ASSP*, v. 29, pp.254-272, April 1981.

[10] B-H. Juang, L.R. Rabiner and J.G. Wilpon, "On the Use of Bandpass Liftering in Speech Recognition," *IEEE Trans. on ASSP*, v. 35, pp. 947-954, July 1987.

[11] A.B. Poritz, "Linear Predictive Hidden Markov Models and the Speech Signal," *Proc. ICASSP82*, v. 2, pp. 1291-1294, May 1982.

[12] N. Tishby, "On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition," (submitted for publication).