SPEAKER IDENTIFICATION WITH USER-SELECTED PASSWORD PHRASES

Aaron E. Rosenberg *S. Parthasarathy*

Speech and Image Processing Services Research Lab
AT&T Labs
Florham Park, NJ 07932 USA

ABSTRACT

An open-set speaker identification system is described in which general-text, sentence-long phrases are used as passwords. Customers are allowed to select their own password phrases and the system has no knowledge of the text. Passwords are represented by phone transcriptions and whole-phrase Hidden Markov Models (HMM's). Phrase identification, carried out using both speaker dependent and speaker independent models, constitutes an identity claim. Verification of the claim uses likelihood ratio scoring with speaker independent phone HMM's providing the background model score. An evaluation has been carried out over a database of password phrases spoken by 250 speakers. 100 of the speakers are test speakers. In an experimental trial, each test speaker is designated as a customer or an imposter and speaks the phrase associated with the customer. The imposter set for each customer consists of same-gender test speakers excluding the customer. At a 5% reject level, the rate of imposter identification is approximately 4%. The misidentification rate for both customers and imposters is less than 0.1%. The closed-set identification error rate is less than 1%, while the average verification equal-error rate is approximately 3%.

1. INTRODUCTION

In this paper, we describe an (open-set) speaker identification system in which each customer's password phrase is fixed, arbitrary text, and sentence-long. Both speaker independent and speaker dependent recognition are used first to validate the password text and then to verify the speaker's identity. An earlier system combining speaker independent phrase recognition and speaker dependent identity verification uses digit-string passwords[1]. Other approaches for validating password texts for speaker recognition have been described by Matsui and Furui [2, 3].

In the current system, it is assumed that customers select their password phrases and that the texts are unknown to the system. Two kinds of reference models are constructed from training utterances for each customer. These are automatically derived phone transcriptions of the selected password phrase, stored in a lexicon with other customers' password phrases, and a whole-phrase, speaker dependent hidden Markov model (HMM) of the phrase. In an identification trial an unknown speaker records a password phrase. Identification proceeds in three phases. In the first phase the input phrase is processed and scored against the lexicon of customer password phrases using speaker independent phone models to obtain a small set of highest scoring items. In the second phase, the input phrase is scored against HMM's for the highest scoring items to select the one that scores best. In the third phase, the verification score for this item is obtained and compared with a decision threshold to determine whether to accept or reject the proposed identification. There is no explicit assumption that the customer-selected phrases are unique. But successful operation of the system requires that the correct customer's phrase should be included in the small set of highest scoring items in the lexicon. This would be compromised by duplicated or confusable password phrases.

2. DETAILS OF OPERATION

2.1. Front-end processing

Each utterance is recorded over the long distance telephone network and subjected to the following front-end processes. The signal is digitized with a 3200 Hz low-pass anti-aliasing filter. The digitized recording is high-pass filtered at 300 Hz to minimize the effects of variable low frequency spectral shaping in the telephone network, preemphasized using a first order difference digital network, and converted to 10th order linear predictive coding (LPC) coefficients every 10 ms over 30 ms windows. The LPC coefficients are converted to 12th order cepstral coefficients and augmented by 12th order delta cepstral coefficients calculated over 5-frame windows. Each utterance is endpointed using an energy-based endpointing algorithm. Channel normalization is carried out with cepstral bias removal by calculating the averages of the cepstral coefficients over the speech portions of each utterance and subtracting them from the instantaneous cepstral coefficients for each frame.

2.2. Training

Each customer enrolls in the system by recording samples of his or her selected password phrase in a single session. Each endpointed training utterance token is compared with preceding tokens using a dynamic programming time warp (DTW) procedure until 3 utterance tokens are obtained whose distance scores with preceding tokens fall below a threshold. A speaker independent phone recognizer is used to obtain a phone transcription for each of the selected training utterances. The recognizer operates in an "automatic" or "free running" grammar mode. That is, any phone (or silence) unit can follow any other unit. The phone models are a set of 43 context independent phone models, representing 41 phone units and 2 silence units, trained from a large telephone database These are 3-state models, each with (nominally) 64 mixture components. The set of 3 phone transcriptions jointly serves as the entry for the customer's password phrase in a lexicon of customer phrases.

In addition to the set of phone transcriptions in the customer password phrase lexicon, each customer's password is also represented by a speaker dependent, whole-phrase, continuous density, Gaussian mixture, hidden Markov model (HMM). The selected training utterances are input to a segmental k-means HMM training process [4] in which the initial segmentations are provided by the endpoints previously obtained in the phone transcription process. A "best" phone transcription (described below) for the password phrase is also obtained. This transcription is used to specify the sequence of phone units for a speaker background model representing the password phrase. The background model is used to obtain scores to help select the best candidate phrase in identification and to normalize verification scores. The number of phones in the "best" transcription is also used to specify the number of states in the whole-phrase speaker dependent HMM. Currently, the number of model states is taken to be 1.5 times the number of phones in the "best" transcription. The nominal number of mixture components for each state is 4. The "best" phone transcription is obtained by scoring each selected training utterance against each of the 3 phone transcriptions contained in the customer phrase lexicon. The scoring is carried out using a speaker independent phone recognizer in a "forced string" grammar mode. The transcription for which the highest overall score is obtained is taken to be the "best" transcription.

2.3. Identification

In an identification trial, an unknown speaker records his or her password phrase. The utterance is input to a speaker independent phone recognizer which scores the input utterance jointly against the alternate transcriptions for each entry in the phrase password lexicon. The recognizer operates in an n-best candidate mode and outputs the 5 highest scoring entries in the lexicon. The speaker independent recognizer also reports scores for the "best" phone transcription for each of the highest scoring items. These scores are referred to as background or "B" scores. The utterance is then scored against the whole-phrase HMM's associated with the highest scoring lexical items. The scores output by this process are referred to as reference or "R" scores. The B and R scores are summed. The item associated with the highest sum identifies the putative customer. The difference of the R and B scores, referred to as the normalized verification score, is then compared with a customer dependent decision threshold to determine whether to accept or reject the putative identity.

Block diagrams of an identification trial are shown in Figs. 1 and 2.

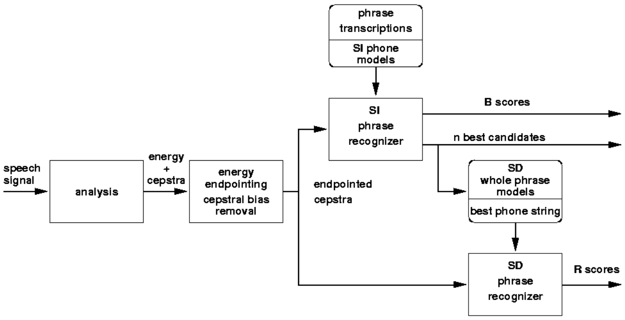

Figure 1. Identification process, part 1, phrase identification

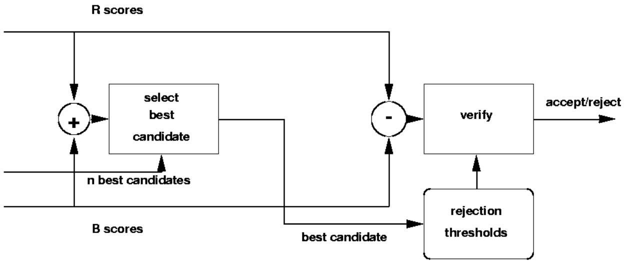

Figure 2. Identification process, part 2, verification

3. EXPERIMENTAL EVALUATION

3.1. Database Description

The speech database used in the experiments is excerpted from a large database of spoken phrases recorded digitally over the telephone network. Each subject provided 27 recording sessions. Subjects were asked to use their normal home or business telephones for the first and last recording sessions but were encouraged to use a variety of telephones, excluding speakerphones, for the remaining sessions. The first and last sessions were designated training and the remainder, testing. Each recording session was managed by a script which provided prompts to the subjects. The recordings took place over a period of approximately 4 weeks for each subject. The data was recorded in 8-bit mu-law formant and checked and labelled by trained listeners. The data excerpted for these experiments consist of a phrase common to all subjects, "I pledge allegiance to the flag", and a subject-selected personal phrase which is different for all speakers. In the instructions, it was suggested that subjects might choose as their personal phrase, the phrase "My name is <name>", where <name> is their own name. As a result of this suggestion, approximately 75% of the subjects chose this example. There are no exact duplications in text over the database of personal phrase selections. The number of words in the personal phrases ranges from 2 to 12, but over 60% of the phrases contain 5 words.

Six tokens of each phrase were recorded in the training sessions while 2 tokens were recorded in each test session. Missing, wrong, botched, or significantly truncated utterances were judged to be faulty and deleted from training and test lists. Training utterances for these experiments are drawn from the first recording session unless all the utterances in that session are marked faulty. Excluding utterances marked faulty, 50 test utterance tokens per phrase are available from each subject. Approximately 250 subjects are used in the experiments.

3.2. Description of Experiments

In an (open set) speaker identification experiment performance figures are calculated for two classes of test speakers: speakers included in the set, designated customers, and speakers outside the set, designated imposters. A test speaker speaks a password utterance assigned to a customer, referred to as a target speaker. The outcome of an experimental trial can be an identification with the target speaker (correct for a customer utterance, an error for an imposter utterance), a misidentification (identification with a speaker other than the target speaker which is an error for both customer and imposter utterances), or a reject (an error for a customer utterance, correct for an imposter utterance).

The intended operation of the system presents some difficulties for evaluation. Since each customer selects his/her own password phrase, it would be difficult to provide a reasonably sized set of imposter speaker tokens for each customer's personal phrase password utterances. To resolve this problem we take advantage of the fact that the speakers in the experimental database provide both personal phrase and common phrase tokens in each test session. We can design experiments to use personal phrase models to evaluate the phrase identification aspects of the system operation and common phrase models to evaluate the verification aspects.

100 of the 250 available database speakers, 50 male, 50 female, are designated as test speakers, used as both customers and imposters. The test utterances are the common phrase test utterances recorded by these speakers. For each test speaker, both common and personal phrase models are constructed (phone transcriptions and HMM's) using each speaker's training utterance tokens. For the remaining 150 speakers, only personal phrase

models are constructed.

The evaluation is carried out in subexperiments, speaker by speaker in the test speaker set. In each such subexperiment one test speaker is designated the target speaker. The set of models for each subexperiment consists of the common phrase models (transcriptions and HMM) for the target speaker and personal phrase models for all other speakers (including the 150 non-test speakers). In a customer test, the common phrase test utterances for the target speaker are compared with all the models. There are nominally 50 test utterances per customer in each test but there may be fewer if faulty utterances are omitted.

The entire set of imposter utterances for the evaluation consists of 400 test utterances, 4 common phrase test utterances from each test speaker. These are taken from the third and fourth test sessions. In an imposter test subexperiment, the target speaker's test utterances are omitted and only test utterances from speakers whose gender is the same as the target speaker are included in the list of test utterances. Thus the imposter comparisons are same gender for the target speaker but both genders for all other speakers in the database. There are 196 imposter test utterances from 49 speakers for each subexperiment.

For the purpose of these experiments, fixed, speaker dependent rejection thresholds are calculated directly from customer test data. In this way, all results are obtained with reference to a calibrated level of customer utterance rejection. A customer rejection level, say 5%, is selected and the score for which 5% of the customer test utterances are rejected is assigned as the threshold. Thus, the results will show imposter identification rates (the rates for which imposters are accepted as the target speaker) and misidentification rates (the rates for which customers and imposters are accepted as speakers other than the target) at thresholds calibrated to reject customers at a 5% rate.

An important ingredient in all our previous speaker verification systems is a model adaptation technique which updates model parameters (mixture means and weights) using current verification data[5]. For the purposes of the evaluation, model adaptation is carried out in a supervised manner, updating the models with a specified number of customer test utterances. Imposter utterances are compared with fully updated target speaker models.

4. RESULTS AND DISCUSSION

Overall results are summarized in Fig. 3. The error rates shown are averages of individual error rates over the 100 target speaker population.

Error rates are shown for three model update conditions with rejection thresholds set to reject 5% of the customer test utterances. The model update conditions are: no model updating, models updated with 4 customer test utterances from the first 4 test sessions, and models updated with 8 customer test utterances from the first 8 test sessions. It can be seen that the imposter accept target speaker rate drops by about 40% with 4 model updates from 6.8% to 4.0%. This trend is consistent with what has been observed in previous experiments for updated models. An additional 4 updates provides only a small additional improvement, to 3.7%. The dashed line plots show that the rates for accepting speakers other than the target speakers are all well below 1%. Model updating improves the imposter accept other speaker rate from 0.48% to 0.01%, while the customer accept other speaker rate decreases from 0.16% to 0.06%. Over a range of rejection thresholds from 2% to 8%, the average error rate remains approximately constant.

A more detailed examination of system performance is obtained by separate examinations of each of the three identification phases described earlier. The output of the first phase is a list of the n best scoring candidate phrases, where n is set to 5 in this experiment.

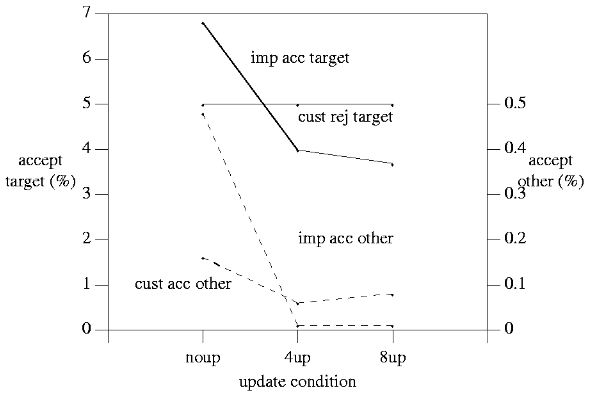

Figure 3. Accept target speaker and other speaker rates for customer and imposter utterances when customer reject thresholds are set at a 5% rate. The solid line plots are associated with the left-hand axis; the dashed line plots with the right-hand axis. Three model update conditions are shown: no updating, models updated with 4 customer test utterances, and models updated with 8 customer test utterances.

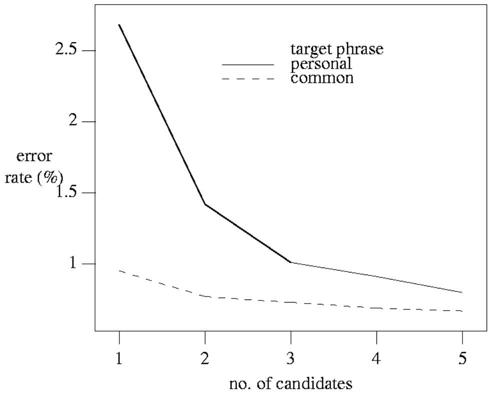

Figure 4. Average phrase identification error rate as a function of the number of best recognizer candidates. The solid line plot refers to personal phrase target speaker phrases while the dashed line plot refers to common phrase target speaker phrases.

The performance of this phase is shown in Fig. 4 where error rate is plotted as a function of number of best candidates for customer test utterances. An error is defined as not finding the correct matching phrase among the specified number of best matching candidate phrases. Error rates are the average of individual speaker error rates over the 100-speaker set of test speakers. Two plots are shown in the figure. The dashed line plot represents the performance associated with the experimental results previously shown. Here the target speaker phrase is the common phrase while all other speakers are represented by personal phrases. The solid line plot shows performance when the target phrase is the personal phrase and all speakers are represented by personal phrases. It can be seen that the personal target phrase error rates are consistently higher than the common target phrase error rates demonstrating that the personal phrase database is more confusable than a database in which the common phrase is substituted for the target speaker's personal phrase. However, the performance gap narrows considerably when all 5 candidate phrases are allowed. This performance is 0.80% for the personal target phrase lexicon versus 0.67% for the common target phrase lexicon. Note that the 5-candidate error rates represent the best possible identification error rate for the whole system. That is, if the correct candidate is not found among the 5 best matching candidate phrases, an identification error is guaranteed.

The second phase of the identification system selects the putative phrase from among the 5 candidates selected in the first phase. The putative phrase is associated with the candidate with the greatest sum of R and B scores. Since the speaker dependent model and background model comprise two different representations of the phrase, it seems reasonable to select the candidate based on the product of their likelihoods (equivalently, the sum of their log likelihood scores) to take advantage of any statistical independence. Experimentally we have found that the use of the sum of R and B scores for identification provides a small, 5% to 10%, improvement over using the R score alone.

The third phase is verification. Basic speaker verification performance can be examined by comparing customer and imposter common phrase utterances with common phrase models. Equal-error rates averaged over the test speakers are shown in Table 1.

transcription	equal-error rates (%)			
	without adaptation		with adaptation	
	unnorm.	norm.	unnorm.	norm.
free phone	6.76	5.34	5.02	2.94
dictionary	6.83	3.46	4.96	2.08

Table 1. Equal-error rate performance for common phrase utterances and models, without and with adaptation, for free phone (unknown), and dictionary (known) transcriptions

Performance is shown without and with (4-update) model adaptation for both unnormalized and normalized scores. There are two experimental conditions. The first row shows "free phone" results which is the experimental condition associated with the results shown in Fig. 3. Here segmentations for both training utterances and test utterances and the phone specifications for background models are obtained using the "best" free phone transcription, as described earlier. For comparison, the second row shows results for dictionary-based segmentations and background models. In this case the phone transcription is the same for each speaker.

Examining the free phone results, we see that the normalized equal-error rates are 5.3% without adaptation and 2.9% with adaptation. The 2.9% equal-error rate is much smaller than the average, 4.5%, of customer reject and imposter accept rates shown in Fig. 3. This is primarily because individual equal-error rates are associated with optimum thresholds while the error rates in Fig. 1 are obtained with thresholds which maintain a fixed rate of rejection for all speakers.

With dictionary based segmentations and background models the equal-error rates are approximately 3.5% without adaptation and 2.1% with adaptation. This is an improvement of about 50% over the error rate with free-phone transcriptions. The degradation must be attributable in some way to the effects of free phone transcriptions. If there were significant differences in the quality of the segmentations between free phone and dictionary transcriptions, we would expect to see differences in the error rates associated with unnormalized scores. However, a comparison of the unnormalized error rates indicates essentially no difference in performance between free phone and dictionary transcriptions. There are, however, significant differences in equal-error rate performance associated with normalized scores. Since normalized scores are dependent on the composition of background models, this suggests that background models based on free phone transcriptions are inferior to background models based on dictionary transcriptions.

5. CONCLUSION

It is useful to compare the performance and operation of the general phrase speaker identification system described here with an earlier digit-string password system[1]. In the digit-string system, customer are assigned unique 14-digit account number passwords. Speaker models are composed of concatenated, speaker dependent, digit HMM's and background models are concatenated, speaker independent HMM's. Phrase identification, providing the identity claim for verification, is carried out by means of a speaker independent digit recognizer. Digit recognizer accuracy is high enough to carry out this operation in an "open loop" mode, with no grammatical constraints other than the digit string length. In contrast, because of poor phone recognizer capability, general phrase recognition must use the entire lexicon of customer phrases as a grammar to drive the recognizer to maintain performance comparable to the digit-string password system. This imposes practical limits on the size of the lexicon for reasonable operation in real time.

The use of phone rather than digit models also degrades the verification component of performance. Average individual equal-error rates of 1% or less are obtainable for digit-string passwords with model adaptation[1]. For general phrase passwords we have obtained 2.9% average individual equal-error rate for "free phone" transcriptions and 2.1% for dictionary transcriptions. The dictionary transcription performance is a fairer comparison since the digit-string passwords are known to the system. Even so, this reduced error rate is twice the error rate obtainable for digit string passwords.

To sum up, it is attractive to allow customers to use general text passwords in a speaker identification system, but the generalization is accompanied by some performance penalties. The degradations might be lessened by supplying the system with a transcription of text, by supervising the selection of password phrases to control acoustic confusability, and/or by specifying hybrid password phrases consisting of digits (or other words known to the system) and general texts. The phrase identification performance of the system is quite high as demonstrated by the very small misidentification error rates. However, the entire lexicon of customer phrases must be used to drive the phrase recognition implying practical limits on system operation.

REFERENCES

[1] A.E. Rosenberg, S. Parthsarathy, "Speaker Background Models for Connected Digit Password Speaker Verification," *Proc. ICASSP 96*, vol. 1, pp. 81-84, International Conference on Acoustics, Speech, and Signal Processing, Atlanta, May, 1996.

[2] T. Matsui and S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition," *Proc. ICASSP 93*, vol. II, pp. 391-394, International Conference on Acoustics, Speech, and Signal Processing, Minneapolis, April, 1993.

[3] T. Matsui and S. Furui, "Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text-Prompted Speaker Recognition," *Proc. ICASSP 94*, vol. I, pp. 125-128, International Conference on Acoustics, Speech, and Signal Processing, Adelaide, May, 1994.

[4] L.R. Rabiner, J.G. Wilpon, and B-H. Juang, "A Segmental K-Means Training Procedure for Connected Word Recognition," *AT&T Tech. J.,* vol. 65, pp. 21-31, 1986.

[5] A.E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang, and F.K. Soong, "The Use of Cohort Normalized Scores for Speaker Verification," *Proc. ICSLP 92*, vol. 2, pp. 599-602, International Conference on Spoken Language Processing, Banff, 1992.