SPEAKER VERIFICATION USING VERBAL INFORMATION VERIFICATION FOR AUTOMATIC ENROLLMENT

Qi Li and Biing-Hwang Juang

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974, USA
{qli,bhj}@research.bell-labs.com

ABSTRACT

A conventional speaker verification (SV) system needs an enrollment session to collect the training data. In [1], we introduced a speaker authentication method called *verbal information verification* (VIV) which verifies a speaker by verbal contents instead of speech characteristics. Such a system does not need an enrollment session. In this paper, VIV is combined with SV. We propose a system which uses VIV to collect training data during the first few accesses automatically, which are often from different acoustic environments. Then, a speaker dependent model is trained and speaker authentication can be performed by SV. This approach not only avoid formal enrollment session which brings convenience to the user, but mitigates the mismatch problem causing by different acoustic environments between training and test sessions. Our experiments show that the proposed system improved the SV performance over 40% compared to the conventional SV system.

1. INTRODUCTION

In [2], we introduced the concept of *speaker authentication*. It is the process of verifying or associating a speaker with an identity using pre-stored information. There are two major approaches to speaker authentication: by the speech characteristics and by the verbal content. The first approach includes *speaker verification* (SV), which has been studied for several decades. We named the second approach as *verbal information verification* (VIV) [1]. It is the process of verifying the spoken information against the content of a given (pre-stored) data profile. This paper is to combine the advantages of these two approaches, and propose a new system for speaker authentication.

A typical speaker verification system is shown in Fig. 1. It involves two kinds of sessions, enrollment and test. In an enrollment session, an identity, such as an account number, is assigned to the speaker, and the speaker is asked to select a spoken pass-phrase, e.g. a connected digit string or a phrase. The system then prompts the speaker to repeat the pass-phrase for several times, and a speaker dependent (SD) hidden Markov model (HMM) is built based on the enrollment utterances in the session. In a test session. The speaker's test utterance is compared against the pre-trained, SD HMM model. A speaker is accepted if the matching score exceeds a preset threshold; otherwise the speaker is rejected.

An example of the VIV system is the current telephone banking procedures: after an account number is provided, an operator verifies a user by asking some personal information, such as mother's maiden name, other personal information or some user-selected and pre-stored pass-phrases. A user has to answer the questions correctly in order to gain access to his or her account. To automate the whole procedure, the questions can be prompted by a text-to-speech system (TTS), and the spoken responses can be verified automatically by ASR or an utterance verification technique.

The major difference between SV and VIV in speaker authentication is that SV inspects speakers' speech characteristics while VIV inspects speakers' verbal content. The difference can be further discussed in three aspects. First, SV needs to train speaker dependent (SD) models or classifiers while VIV just needs speaker independent (SI) acoustic phone models. Second, SV needs an enrollment session to record speech and to train SD models while VIV does not. The profile in VIV is created when the user's account is set up. Third, in SV, different users may use the same pass-phrase while in VIV, it is the speakers' responsibility to protect their own personal information.

In this paper, we combine the advantages of both VIV and SV to propose a new speaker authentication method. Using the method, a speaker is first verified by VIV. After the speaker accesses the account for a few times, usually 4 to 5 times, a SD HMM is trained using the recorded pass-phrases of previous accesses. Then, the authentication process can

be switched from VIV to SV. Thus, impostors can not get into a user's account even he or she knows the user's pass-phrase. Since the training data are from different sessions, i.e. different handsets and channels, the mismatch problem can be mitigated. The concept of this new approach is shown in Fig. 2.

As we know, the accuracy of collected training data is very important to SV performances. Even a true speaker may make a mistake when repeating the training utterances. When VIV is used in the data collection, it can avoid involving incorrect utterances in training.

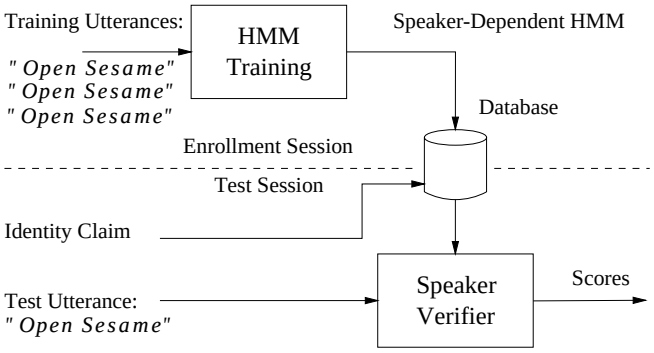

Figure 1: Conventional speaker verification system

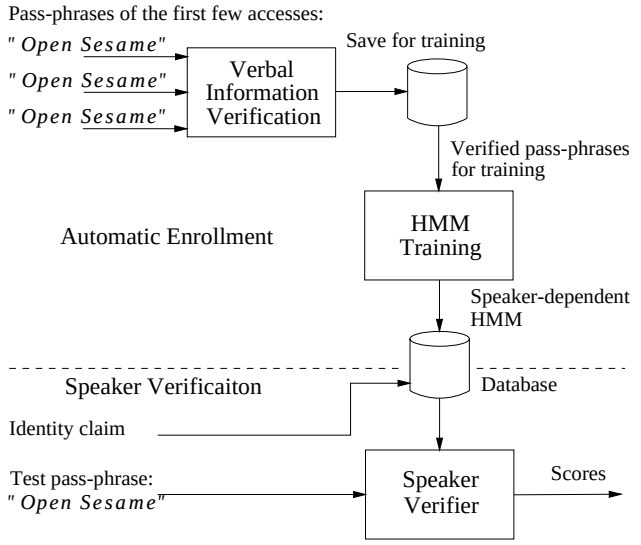

Figure 2: Proposed speaker verification system

In Section 2 and 3, we will review the VIV and SV algorithms respectively. In Sections 4 and 5, we will introduce the database and report the performances of the proposed system.

2. VERBAL INFORMATION VERIFICATION

There are two methods to verify the verbal content, automatic speech recognition (ASR) and utterance verification. They are shown in Fig. 3 and 4 respectively. The utterance verification approach is to verify whether the answer utterance matches the phone sequence generated from the expected answer phrase. As we have reported in [1], the utterance verification approach can give us much better performance than the ASR approach.

In the utterance verification approach (Fig. 4), input speech is first aligned with a sequence of transcribed phones of the correct answer via SI HMM's. Then, for each phone, the likelihood scores from the SI HMM's and a set of anti-HMM's of the corresponding phrases are calculated for hypothesis testing. A confidence measure is then calculated for verification decision [1]. Here, the anti-HMM for a target phone is trained using the data of the neighboring phones [3].

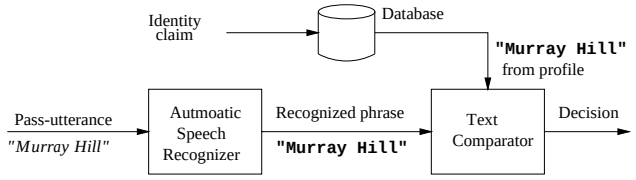

Figure 3: Verbal information verification by automatic speech recognition on a pass-utterance

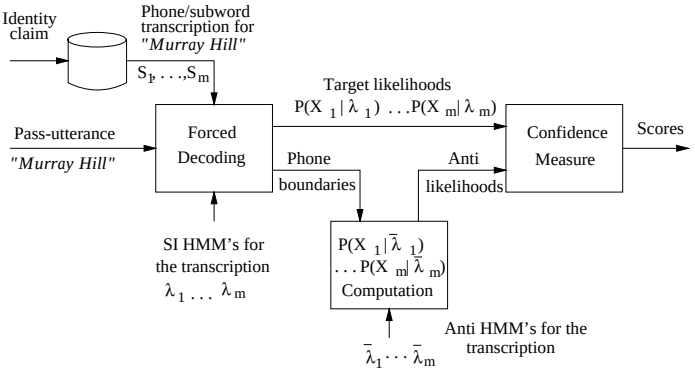

Figure 4: Utterance verification for VIV

The VIV system has been tested on a database of 100 English speakers [1]. Each speaker has 3 utterances as the answers to three questions asked for verification. The answer to each of the three questions is verified by utterance verification. If all three answers are correct, the speaker is accepted. If any answer is different from the information registered in the corresponding personal profile, the speaker is then rejected and no further questions are asked. For the above task, when a speaker dependent threshold is set for

each key information field for that speaker, we achieved 0% average individual equal-error rate.

3. SPEAKER VERIFICATION

A block diagram of the SV system used in this paper is shown in Fig. 5 [4, 5]. After the speaker claims the identity, the system expects the same phrase obtained in the training session. First, a speaker independent (SI) phone recognizer segments the input utterance into a sequence of phones and silence by forced decoding using the transcription saved in the user's profile. The segmentations are for cepstral mean subtraction (CMS) and for the background models [4]. The input utterance is also decoded by a speaker dependent (SD) target model. The target model is trained by the training utterances verified by VIV. In the verifier, a log-likelihood-ratio score is calculated based on the log-likelihood scores from the target and the background models [4].

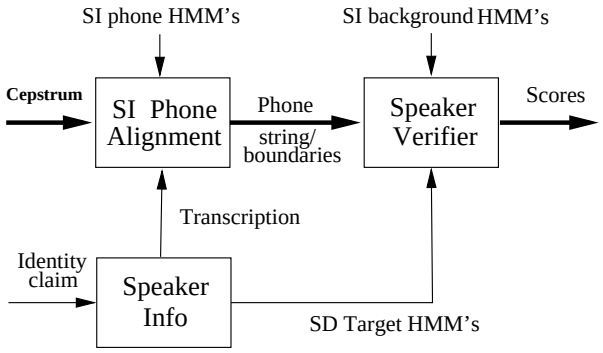

Figure 5: A phrase-based speaker verification system

$$L_R(\mathbf{O};\Lambda_t;\Lambda_b) = L(\mathbf{O},\Lambda_t) - L(\mathbf{O},\Lambda_b) \qquad (1)$$

where $\mathbf{O}$ is the observation sequence over the whole phrase, and Λ_t and Λ_b are the target and background models respectively. The background model is a set of HMM's for phones. The target model is an SD HMM with multiple states for the whole phrase and is trained for the particular speaker. As reported in [4], this configuration provides the best results in experiments. In (1),

$$L(\mathbf{O},\Lambda_t) = \frac{1}{N_f} P(\mathbf{O}|\Lambda_t), \qquad (2)$$

where $P(\mathbf{O}|\Lambda_t)$ is the log-likelihood of the phrase evaluated by the SD target HMM, Λ_t, using Viterbi decoding, and N_f is the total number of non-silence frames in the phrase, and

$$L(\mathbf{O},\Lambda_b) = \frac{1}{N_f} \sum_{i=1}^{N_p} P(\mathbf{O}_i|\Lambda_{b_i}) \qquad (3)$$

where $P(\mathbf{O}_i|\Lambda_{b_i})$ is the log-likelihood of the ith phone, $\mathbf{O}_i$ is the segmented observation sequence corresponding to the ith phone in the phrase, Λ_{b_i} is an SI background HMM for the ith phone, N_p is the total number of the decoded non-silence phones, and N_f is the same as above.

A finial decision on rejection or acceptance is made based on the L_R score with a threshold. If a significantly different phrase is given, the phrase could be rejected by the SI phone recognizer before using the verifier.

4. FEATURES AND DATABASE

The feature vector in this paper is composed of 12 cepstrum and 12 delta cepstrum coefficients. The cepstrum is derived from a 10th order LPC analysis over a 30 ms window. The feature vectors are updated at 10 ms intervals.

The experimental database consists of fixed phrase utterances recorded over the long distance telephone network by 100 speakers, 51 male and 49 female. The fixed phrase, common to all speakers, is "I pledge allegiance to the flag" with an average length of 2 seconds. Five utterances of each speaker recorded from five separate VIV sessions are used to train the SD HMM. For testing, we used 40 utterances recorded from a true speaker in different sessions (different telephone channels at different times), and 192 utterances recorded from 50 impostors of the same gender in different sessions. For model adaptation, the second, fourth, sixth, and eighth test utterances from the tested true speaker are used to update the associated HMM for verifying succeeding test utterances.

The target models for the phrases are left-to-right HMM's. The number of the states are 1.5 times the total number of phones in the phrases. There are 4 Gaussian components associated with each state [4]. The background models are concatenated SI phone HMM's trained on a telephone speech database from different speakers and texts [6]. Each phone HMM has 3 states with 32 Gaussian components associated with each state.

Due to unreliable variance estimates from limited amount of training data, a global variance estimate is used as a common variance to all Gaussian components [4] in the target models.

5. EXPERIMENTAL RESULTS

In [1], we have reported the experimental results of VIV on 100 speakers. The system has 0% error rates when three questions were asked, and answers were verified by utterance verification. So, we assume that all the training utterances collected by VIV are corrected.

The SV experimental results without and with adaptation are listed in Table 1 and Table 2 for the 100 speakers. The numbers are in the average percentages of individual

equal-error rates (EER's). The first data column is the EER using individual thresholds and the second data column is the EER using common (pooled) thresholds. All the scores are obtained with log-likelihood-ratio scores using phrase-based target model and phone-based speaker background models.

The baseline system are the conventional SV system in which a single enrollment session is used. The proposed system are the combined system in which VIV is used for the automatic enrollment of SV. After the VIV system is used for 5 times, it then switches over to the SV system. The test utterances for both the baseline and the proposed systems are the same.

Without adaptation, the baseline system has an EER of 3.03% and 4.96% for individual and pooled thresholds respectively, while the proposed system has an EER of 1.59% and 2.89% respectively. With adaptation, the baseline system has an EER of 2.15% and 3.12%, while the proposed system has an EER of 1.20% and 1.83% respectively. The proposed system without adaptation has even lower EER than the baseline system with adaptation.

Table 1: **Experimental Results in Average Equal-Error Rates (%)**

Algorithms	Individual Th.	Pooled Th.
SV (Baseline)	3.03	4.96
VIV+SV(proposed)	1.59	2.89

Table 2: **Experimental Results with Adaptation in Average Equal-Error Rates (%)**

Algorithms	Individual Th.	Pooled Th.
SV (Baseline)	2.15	3.12
VIV+SV(proposed)	1.20	1.83

6. CONCLUSIONS

In this paper, verbal information verification and speaker verification are combined to construct a convenient speaker authentication system with improved error rates. This is based on the assumption that the VIV system can be almost error free in collecting training data from different sessions. The combined system is convenient to users since they can start to use the authentication system without going through enrollment sessions and waiting for model training. However, it is still the user's responsibility to protect his or her pass-phrases from impostors during the first few accesses before an SD HMM is trained. After the SD model is trained, the system can function as an SV system, i.e. impostors can not get into an account although the impostors may know the pass-phrases. Since the training data could be collected from different phone calls of different VIV sessions, the acoustic mismatch problem is mitigated which leads to the better system performances in test sessions. The SD HMM's can be further updated to cover different acoustic environments and improve the system performance.

7. ACKNOWLEDGMENT

The authors wish to thank S. Parthasarathy and Aaron E. Rosenberg for their original contributions to the general phrase speaker verification system.

8. REFERENCES

[1] Q. Li, B.-H. Juang, Q. Zhou, and C.-H. Lee, "Verbal information verification," in *Proceedings of EUROSPEECH*, (Ghode, Greece), Sept. 1997.

[2] Q. Li, B.-H. Juang, C.-H. Lee, Q. Zhou, and F. Soong, "On speaker authentication," in *Proceedings of IEEE Workshop on Automatic Identification Advanced Technologies*, (Stony Brook, NY), Nov. 1997.

[3] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," *IEEE Trans. Speech and Audio Process.*, vol. 4, pp. 420–429, November 1996.

[4] S. Parthasarathy and A. E. Rosenberg, "General phrase speaker verification using sub-word background models and likelihood-ratio scoring," in *Proceedings of ICSLP-96*, (Philadelphia), October 1996.

[5] Q. Li, S. Parthasarathy, and A. E. Rosenberg, "A fast algorithm for stochastic matching with application to robust speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Munich), pp. 1543–1547, April 1997.

[6] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Atlanta), pp. 81–84, May 1996.