ON SPEAKER AUTHENTICATION

Qi Li, Biing-Hwang Juang, Chin-Hui Lee, Qiru Zhou, and Frank K. Soong

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974
{qli,bhj,chl,qzhou,fks}@bell-labs.com

ABSTRACT

The concept of speaker authentication is reviewed in this paper. It includes: *speaker recognition* and *verbal information verification* (VIV). While speaker recognition has been studied for more than two decades, VIV is a rather new concept. After a brief overview of both techniques, we present two advanced systems: a speaker verification system and a VIV system, respectively. Their good performances are reported.

1. INTRODUCTION

Speaker authentication is the process of verifying or associating a speaker using pre-saved information. Applications of speaker authentication include access control for telephones, computer networks, databases, bank accounts, credit-card funds, and automatic teller machines, etc. Automatic authentication by voice is convenient for users. It needs less on additional devices compared with other biometric methods. Speaker authentication is important to special terminals, such as telephone handsets, and becomes more and more important in mobile and wireless applications.

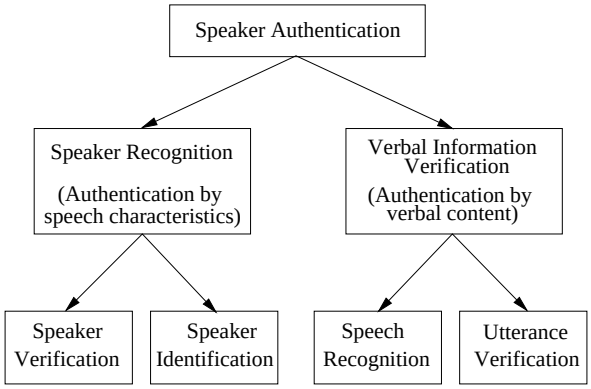

Figure 1: Speaker authentication approaches

There are two major approaches to speaker authentication: by speech characteristics and by verbal content, as shown in Fig. 1. The first approach is called *speaker recognition* including speaker verification and speaker identification. By definitions, *speaker verification* (SV) is the process of verifying whether an unknown speaker is the same as the speaker whose identity is being claimed; on the other hand, *speaker identification* is the process of associating an unknown speaker with a member of a known population of speakers. We named the second approach as *verbal information verification* (VIV) [1]. It is the process of verifying spoken information against the content of a given (pre-stored) data profile. There are two methods to verify verbal content, automatic speech recognition (ASR) and utterance verification. Since SV and VIV systems have more applications in telecommunication, we focus our discussions on these two kinds of systems in this paper.

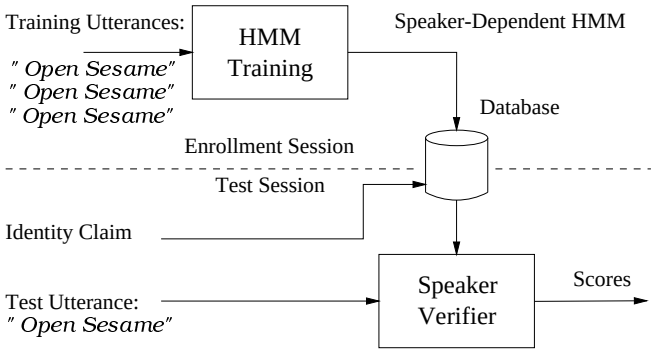

Figure 2: A speaker verification system

A typical speaker verification system is shown in Fig. 2, and there are two kinds of sessions, enrollment and test. In an enrollment session, an identity, such as an account number, is assigned to a 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 hidden Markov model (HMM) is built based on the enrollment utterances. In a test session, first an identity claim is made by the speaker, the system then prompts the speaker to utter

the pass-phrase. The speaker's test utterance is compared against the pre-trained, speaker dependent HMM model. A speaker is accepted if the matching score exceeds a preset threshold, otherwise rejected.

A simple VIV system based on automatic speech recognition (ASR) is shown in Fig. 3 and 4. It is similar to 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, birth date, address, home telephone number, etc. 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.

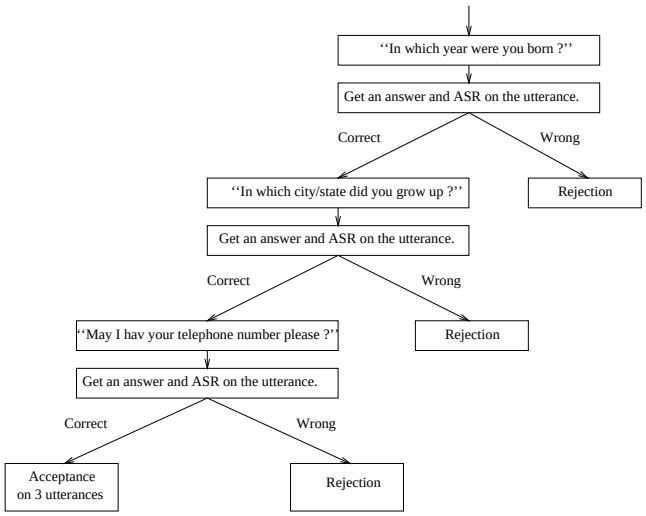

Figure 3: Verbal information verification by automatic speech recognition

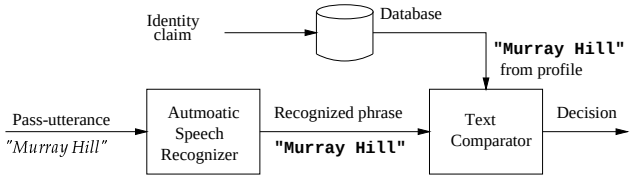

Figure 4: Verbal information verification by automatic speech recognition on one pass-utterance

A major difference between speaker recognition and VIV in speaker authentication is that speaker recognition inspects speakers' speech characteristics while VIV inspects speakers' verbal content. The difference can be further discussed in three aspects. First, both speaker identification and speaker verification need to train speaker dependent (SD) models or classifiers while VIV just needs speaker independent (SI) acoustic phone models. Second, speaker recognition needs an enrollment session to record SD speech data and to train SD models while VIV does not. The profiles are created when users' accounts are set up. Third, in speaker verification, the system can reject an imposter who uses a true speaker's spoken password while, in VIV, it is the speakers' responsibility to protect their own personal information. VIV, however, can be used for automatic enrollment of speaker recognition systems, or can be used in conjunction with speaker verification to further enhance the security.

2. ADVANCED SPEAKER VERIFICATION SYSTEM

An advanced speaker verification system is shown in Fig. 5 [2, 3]. The block diagram illustrates how the system verifies a spoken pass-phrase. After a speaker claims his or her identity (ID), the system expects the same phrase obtained in the associated training session. First, a speaker independent (SI) phone recognizer segments the input utterance into a sequence of phones by aligning the input utterance with transcription obtained from the enrollment session. Since the SD models are trained on a small amount of data from a single session and they can't be used to provide a reliable and consistent phone segmentations, SI phone models are used instead. Also, a so-called "stochastic matching" procedure is used to compensate the noise and channel difference between training and test utterances. Then, the compensated cepstral coefficients, decoded phone sequence, and associated phone boundaries are fed to a verifier. In the verifier, a log-likelihood-ratio (LLR) test is performed by computing the log-likelihood scores of SD-target and SI-background models. The LLR score is then compared with a SD threshold to make a decision of acceptance or rejection.

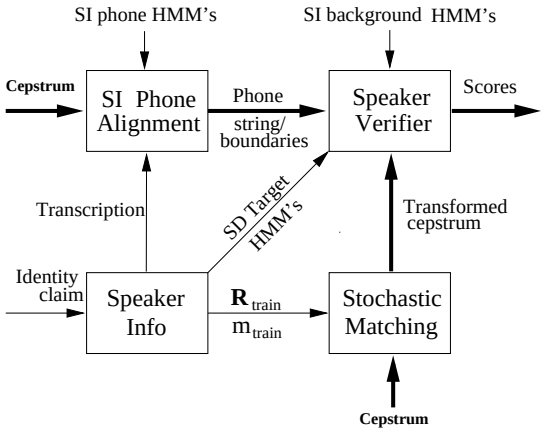

Figure 5: A phrase-based speaker verification system

The system has been tested on a database consisting of fixed phrase utterances recorded over the long distance tele-

phone networks 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 in one enrollment session are used to train a SD HMM. For testing, we used 50 utterances recorded from a true speaker in different sessions (different telephone channels and handsets at different times), and 200 utterances recorded from 50 impostors of the same gender in different sessions. In order to further improve the SD HMM, an adaptation procedure is performed. 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. For the above database, the average individual equal-error rates over 100 speakers are 2.6% without adaptation and 1.8% with adaptation [2]. If the system is tested on the impostors using different pass-phrases than the common one, the performance can be much better.

3. ADVANCED VERBAL INFORMATION VERIFICATION SYSTEM

We have developed an advanced VIV system based on the techniques of speaker verification [4, 2] and utterance verifications [5, 6, 7, 8, 9]. A block diagram is shown in Fig. 6, and more details on utterance verification is shown in Fig. 7. Similar to speaker verification discussed in the last session, the voice response to a question is first aligned with a sequence of transcribed phones of the correct answer via speaker-independent HMM's. Then, for each phone, the likelihood scores from the SI HMM's and a set of anti-HMM's are calculated for hypothesis testing. The confidence measure is to combine the hypothesis test score on each phone into an utterance-level score for verification decision.

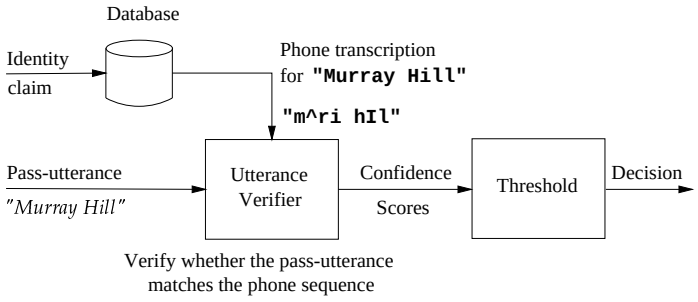

Figure 6: Verification approach for VIV

The VIV system has been tested on a database of 100 English speakers. Each speaker record 3 utterances as the answers to three questions:

- "In which year were you born?"

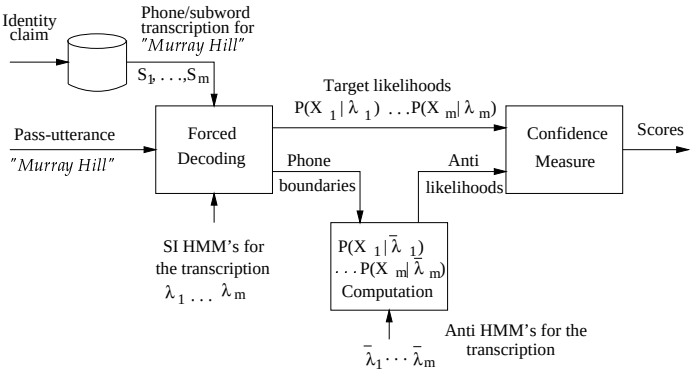

Figure 7: Utterance verification for VIV

- "In which city and state did you grow up?" and
- "May I have your telephone number, please?"

The answer to each of the three questions is verified by our VIV system. 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 in our current implementation.

It is a biased database since 26% of the speakers were born in 1950's, and 24% were in 1960's. We note that there is only one or two digit difference in those years. In the city and state names, 39% are "..., New Jersey", and 5% of the speakers use exactly the same address "Murray Hill, New Jersey", which means the verification system will give the same results on these addresses. Thirty eight percent (38%) of telephone numbers start with the same area code "908" and prefix "582", that means at least 60% of the context of the telephone numbers are exactly the same for this piece of information. Moreover, some of the speakers have accents, and some cities and states are in foreign countries. In our experiments, each speaker is tested as a true speaker against his or her data profile. Other speakers are tested as impostors against the true speaker's profile. Thus, for each speaker, we have three utterances from the true speaker and 99×3 utterances from the impostors.

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 with 6% tolerance interval, which means a true speaker can still be accepted if verified phones in the verbal responses are 6% lower.

In real speaker authentication applications, to avoid impostors using a speaker's personal information which is just uttered, a VIV system can randomly ask a subset of personal information for each access. For example, based on registered 6 items, each time system can randomly pick 3 to verify the speaker. Furthermore, the system can ask some

dynamic information recorded in the past transactions, such as the date or the amount of the last deposit.

4. CONCLUSIONS

The common techniques used in the above advanced systems are based upon the hypothesis testing procedure in statistics. They are derived from the Neyman-Pearson's lemma [10, 11]. By applying the background models [4] (Fig. 5) and the anti-models [5] (Fig. 7), the verification problem becomes a two-class hypothesis testing problem and system performance is improved significantly by applying the technique over the raw scoring based method. Essentially, the performance which we achieved on the databases shows that both systems are ready to be deployed in real applications.

5. ACKNOWLEDGMENT

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

6. 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, 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.

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

[4] 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.

[5] 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.

[6] R. A. Sukkar, A. R. Setlur, M. G. Rahim, and C.-H. Lee, "Utterance verification of keyword string using word-based minimum verification error (WB-MVE) training," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Atlanta), pp. 518–521, May 1996.

[7] A. R. Setlur, R. A. Sukkar, and J. Jacob, "Correcting recognition errors via discriminative utterance verification," in *Proc. Int. Conf. on Spoken Language Processing*, (Philadelphia), pp. 602–605, Oct. 1996.

[8] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Combining key-phrase detection and subword-based verification for flexiable speech understanding," in *Proceedings of ICASSP*, (Munich), May 1997.

[9] M. G. Rahim, C.-H. Lee, and B.-H. Juang, "Robust utterance verification for connected digits recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Detroit), pp. 285–288, May 1995.

[10] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purpose of statistical inference," *Biometrika*, vol. 20A, pp. Pt I, 175–240; Pt II, 1928.

[11] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. Roy. Soc. A*, vol. 231, pp. 289–337, 1933.