International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

CANTONESE VERBAL INFORMATION VERIFICATION SYSTEM USING GMM-BASED ANTI-MODEL

Chao Qin and Tan Lee

Department of Electronic Engineering,
The Chinese University of Hong Kong,
Shatin, Hong Kong.
Email: {cqin, tanlee}@ee.cuhk.edu.hk

ABSTRACT

Verbal information verification (VIV) is one of the approaches for speaker authentication [1]. It is a process in which the spoken utterance of a claimed speaker is verified against the key information in speaker's registered profile. VIV in English has been extensively studied and there has also been some work on Mandarin VIV. In the paper, we study the VIV for users who speak Cantonese, the most commonly used dialect in Southern China and Hong Kong. We propose a new technique for anti-modeling. It uses context-independent Gaussian Mixture Model (GMM) instead of the conventional Hidden Markov Model (HMM). Experiments on 50 Cantonese native speakers show that the proposed method provides better separation of verification scores of claimant utterances from that of impostor utterances than the HMM based method. An equal error rate of 0.00% is attained with robust interval up to 15%, which manifests an excellent performance.

1. INTRODUCTION

To ensure the proper access to private information or security systems, automatic user authentication is necessary. Traditional methods using password, personal identification number (PIN) and signature to verify the identity of users have been used for a long time. On the other hand, biometrics based techniques that recognize a person using his/her physiological or behavior characteristics is becoming the foundation of an extensive array of highly secure identification and personal verification solution. Useful biometric features include face, fingerprints, voice, etc. Among them, voice is the most convenient one in that it is easy to produce, capture and transmit for remote processing.

The process of confirming a speaker's identity is referred to as speaker authentication (SA). If the decision is based on voice characteristics, the task is named speaker verification (SV). Research on SV technology has made significant progress in the past few years. There are however still some problems left to be resolved for real-world applications, e.g. acoustic environment mismatch and the inconvenience caused by the requirement of enrollment [3].

Speaker verification doesn't utilize the verbal content of a pass-phrase explicitly. However, verbal information is obviously a useful knowledge for authentication. Verbal information verification was proposed as a complementary technique for speaker authentication [1]. It requires the user to speak out his/her personal information such as name, birth date and residence address in order to verify the identity. Different kinds of questions can be asked to implement different security levels, which could be used for users with different level of authorities when accessing a security system. Meanwhile, VIV doesn't require the enrollment process and hence suffer less from the acoustic mismatch problem.

There are two approaches for VIV, which are based on the techniques of automatic speech recognition (ASR) and utterance verification (UV) respectively. UV is preferred to ASR since the ASR approach doesn't effectively utilize the registered information in the user's profile [1]. In a UV system, there are three key modules that determine the system's performance [1]. They are the target models, anti-models and confidence measure. In this paper, a Cantonese VIV system is implemented and evaluated. We propose to use Gaussian Mixture Models (GMM) to realize the anti-models. Experimental results show that GMM based anti-models perform better than the conventional Hidden Markov Models.

The UV approach for VIV will be described in the next section. Then the design of target models and anti-models for Cantonese VIV is discussed in Section 3. Experimental results for Cantonese VIV are presented and analyzed in Section 4. Finally in Section 5, a conclusion will be drawn.

2. UTTERANCE VERIFICATION FOR VIV

Figure 1 gives the details of the utterance verification approach for VIV. The input utterance is first aligned with a sequence of subword units transcribed from the expected answer. This is done with speaker-independent (SI) acoustic models. Afterwards, for each subword, the likelihood scores from the corresponding target model and the anti-model are computed for hypothesis test [1][2]. Decisions will be made on both the subword and the utterance levels. In this study, the confidence measure for subword unit n in an observed speech segment O_n is defined as [1],

$$C_n = \frac{\log P(O_n \mid \lambda_n) - \log P(O_n \mid \overline{\lambda}_n)}{-\log P(O_n \mid \overline{\lambda}_n)} \qquad (1)$$

where λ_n and $\overline{\lambda}_n$ denote respectively the corresponding target model and anti-model for the subword unit *n*. For utterance-level decision, the results of subword tests need to be combined. In this work, we adopt the normalized confidence measure as defined in [1]. If the utterance contains *N* subword units, the normalized confidence measure is given by,

$$M = \frac{1}{N}\sum_{n=1}^{N} f(C_n), \qquad (2)$$

where,

$$f(C_n) = \begin{cases} 1, & if\, C_n \geq \theta; \\ 0, & otherwise; \end{cases} \qquad (3)$$

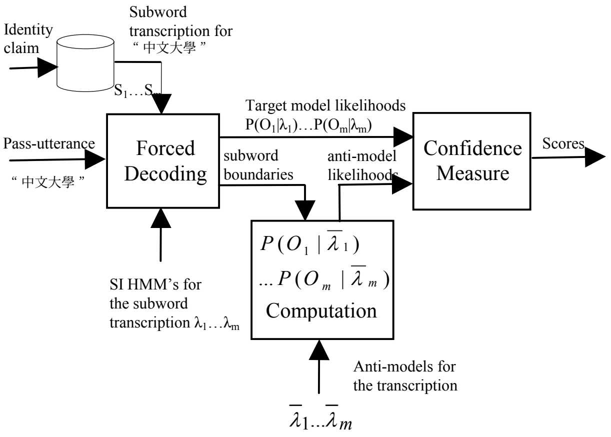
Identity claim
Subword transcription for “ 中文大學 ”
S1…Sm
Target model likelihoods P(O1|λ1)…P(Om|λm)
Pass-utterance
“ 中文大學 ”
Forced Decoding
subword boundaries
anti-model likelihoods
Confidence Measure
Scores
$P(O_1 \mid \overline{\lambda}_1)$
$\ldots P(O_m \mid \overline{\lambda}_m)$
Computation
SI HMM’s for the subword transcription λ1…λm
Anti-models for the transcription
$\overline{\lambda}_1 \ldots \overline{\lambda}_m$

Fig. 1 Utterance Verification Approach for VIV

In Eqn. (1), the subword-level confidence score is normalized using the negative likelihood score from the anti-model. By doing so, it is expected that the confidence measures for different subword units would have similar dynamic range of values. As a result, the threshold θ becomes subword independent.

The utterance-level confidence score M has the value between 0 and 1. It can be roughly interpreted as the percentage of acceptable subwords in the test utterance. When M is larger than a threshold, the utterance will be accepted and the user’s identity is verified. Otherwise it will be rejected. In order to enhance the security level, sequential utterance verification is commonly adopted for VIV. It means that more than one test utterances are used. The user would be accepted only if all of the utterances pass the test.

Due to variability of speech signals, the computed confidence measure may not have the same value if the same person speaks the same utterance twice. The concept of robust interval was introduced to provide certain flexibility in system design. It allows the system to accept an utterance even if the confidence measure is lower than that of previous trials. The robust interval is defined as the maximum percentage of threshold relaxation so that false rejection (FR) can be avoided. Besides, the SI target models and anti-models are both trained on-the-fly. This is important for real-time application.

3. MODEL DESIGN FOR CANTONESE VIV

3.1. General considerations

In VIV, the primary role of the target models is to provide accurate time alignment of the sub-word units for the subsequent processing. Inaccurate sub-word alignment will cause bad effect on subsequent processing. Context-dependent acoustic modeling at phoneme level is considered to be most appropriate for the target model. On the other hand, anti-models should be designed to separate the data (speech containing the expected subwords) from the non-data (speech not containing the expected subwords) as much as possible [8].

3.2. Cantonese

As a prominent Chinese dialect, Cantonese is the mother tongue of the 60 million population in Southern China and Hong Kong. Cantonese is a monosyllabic and tonal language. Each Chinese character is pronounced as a single syllable carrying a specific tone. Each can be divided into an Initial and a Final. The Initial is typically a consonant while the Final consists of a vowel nucleus and an optional nasal or stop coda [9].

3.3. Target models

Cantonese has 20 Initials and 53 Finals, constituting 660 base syllables [5]. In continuous speech recognition for Cantonese, context-dependent (CD) Initial-Final (IF) models, namely Bi-IF and Tri-IF models, have been commonly used. Decision-tree tying method was adopted to tackle the problem of sparse training data [4].

In this research, the target models used for Cantonese VIV are speaker-independent and context-dependent Initial-Final models. Each Initial HMM has 3 states and each Final HMM has 5 states. The HMMs were trained on 20 hours of continuous speech database in the CUSENT corpus collected at the Chinese University of Hong Kong [5]. For base syllable recognition in Cantonese, Tri-IF gave the best accuracy of 80.54% and Bi-IF attained 79.11%.

3.4. Anti-models

To facilitate the hypothesis test in UV, anti-models for different INITIALs and FINALs need to be established. Anti-models can be either context-dependent or context-independent. In Mandarin VIV [6], it was reported that the performance with CD anti-models is worse than that by CI anti-models.

Let a denote a subword unit being modeled and XaY be one of its contextual variation, where X and Y denote the left and right context respectively. In CD anti-modeling, the anti-model for *XaY* is trained by the data that are labeled as *XâY,* where *â ≠ a.* In other words, a training token for the anti-model must have different subword identity, i.e. *â ≠ a* but the same context, i.e. *X* and *Y*. Those segments in which both subword identity and context are different from *XaY* are not used for training. If an imposter utterance contains such an "unseen" segment, the anti-model would generate a relatively low likelihood and a false acceptance (FA) error may occur. Therefore, we decide to use CI anti-model in this research.

In our implementation, the subword acoustic models are first grouped by decision-tree tying method. As a result, six and nine clusters are generated for Initials and Finals respectively. All models belonging to the same cluster share an anti-model, which is trained by the data from all the other clusters. Intuitively, the topology of anti-models should be the same as that of target models. In other words, the anti-models for Initial and Final would be left-to-right HMMs that have 3 and 5 states respectively. In this study, we propose to use GMM instead of HMM for the anti-models. The simplified notations of anti-HMMs and anti-GMMs will be used henceforth. Initial and Final anti-models are treated separately in both clustering and training. The estimation of anti-model parameters is done by the Baum-Welsh re-estimation algorithm. The number of mixtures per state is determined empirically. In all experiments, the anti- models have 16 mixtures per state.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1. Acoustic Features

The acoustic features used in experiments are the mel-frequency cepstral coefficients (MFCC) which includes 12 cepstral coefficients together with the energy. The features are energy normalized and cepstral mean normalized based on each short-time segment. By including the first and second derivatives of the parameters, the final feature vectors have 39 components.

4.2. Speech Databases

The speech database used to train the SI target model is CUSENT [5], a database of Cantonese read sentences. The testing database contains a population of 50 Cantonese native speakers. Each of them is asked to record nine utterances corresponding to their registered profiles through microphone channel.

4.3. Experimental I

We first conduct a VIV experiment in which up to three questions (K=3) are asked to each user. The questions are: "What is your mother's name?" and "What is your birth date?" and "Which high school did you graduate from?" In this experiment, we compare the performance of the proposed anti-GMMs with the conventional anti-HMMs. Here the target model is fixed as Tri-IF. The question sets are the same as in Experiment I. The results are given in Table 1. It is observed that in terms of EER the anti-GMMs outperform anti-HMMs slightly in the case of one prompted question. The anti-GMM performs much better in terms of robust interval which represents the system's robustness. It is found that the optimal CI subword thresholds based on two anti-models are different.

In Chinese, the same sentence or word may be spoken in several different ways. For instance, "Day" in Chinese could be pronounced as "日" or "號". Errors in time alignment will occur when the real content of an utterance doesn't exactly match the corresponding item in the user's profile. Intuitively, we can perform forced alignment for each possible. The one with the highest likelihood score is used for subsequent processing.

Anti-model Number of questions	Anti-HMMs	Anti-GMMs
1	0.200% (0%)	0.157% (0%)
2	0 (1%)	0 (8%)
3	0 (10%)	0 (15%)

Table 1. Statistics of EER on experimental results with Optimal CI subword threshold (Robust Interval in parentheses)

In order to further evaluate and analyze the performance of the two anti-models, anther experiment on single utterance verification has been carried out.

4.4. Experimental II

In this experiment, it is assumed that only a single utterance is available for VIV decision. As said in Section 4.2, each of the 50 test speakers has nine utterances. In other words, for a particular claimant, there are nine claimant utterances and 49×9 impostor utterances. In total, there are 50×9 true utterances for the evaluation of false rejection errors and $50 \times 49 \times 9$ wrong utterances for the evaluation of false acceptance errors. The target model is Tri-IF. The resulted detection error tradeoff curve is plotted as in Fig.2. The use of CI anti-HMMs yields an EER of 1.69% (dotted line) while the system based on CI anti-GMMs leads to an EER of 1.22% (solid line). Again, the anti-GMMs perform better than anti-HMMs.

For the CI anti-model of a particular subword cluster, the training data are subword units or phonemes from all the other clusters. These phonemes could be in any context. However, the sequencing of phonemes found in training data is not related to that found in the test data. Explicitly adding state transition probability may not be appropriate. Unlike HMMs, the GMMs don't attempt to capture the temporal aspects of the training data. In other words, the anti-GMMs are less constrained by the temporal structure of the subword units than the anti-HMMs. They model only the underlying distribution of acoustic observations [7] and each mixture of an anti-model could be viewed as representing one of other clusters, it is therefore more suitable for anti-modeling. Figure 3 shows the distribution of verification scores from FA and FR experiments with anti-HMM and anti-GMM approach. Clearly, the anti-GMM approach has provided better separation of the two distributions. This also explains the superiority of anti-GMMs over anti-HMMs.

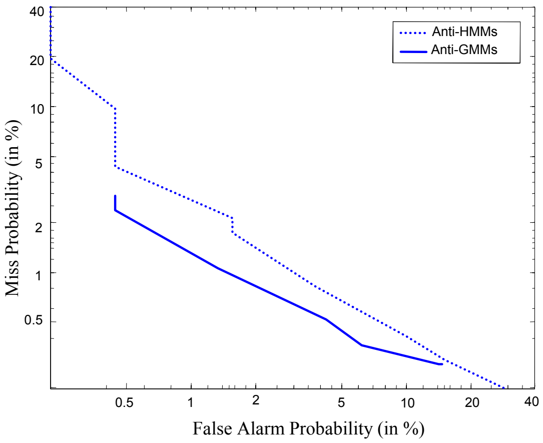

Fig.2 Detection Error Tradeoff Curve based on two anti-modeling approaches

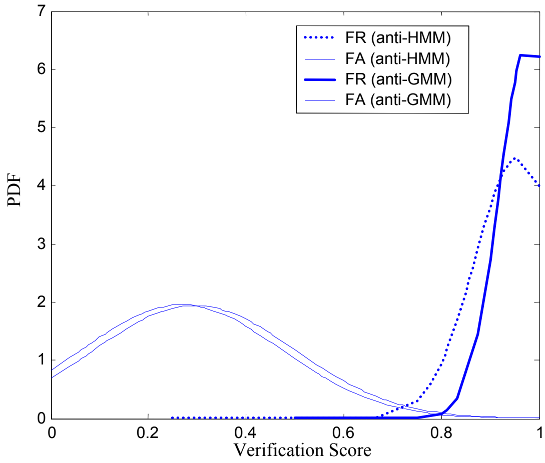

Fig.3 Distribution of verification scores from FA and FR experiment

5. CONCLUSION

In this paper, we have presented a Cantonese VIV system for automatic speaker authentication. Simulation results show that our system yields comparable performance with that of other VIV systems [1][6]. When sequential utterance verification strategy is adopted, error-free performance could be achieved with a robust interval up to 15%. We propose to use anti-GMMs which show the positive effect on experimental results. Preliminary analysis and explanation have been given. Our study indicates that Cantonese VIV technology is ready for real-world application.

6. ACKNOWLEGE

This research was partially supported by a Central Allocation Grant from the Hong Kong Research Grants Council. The first author wishes to acknowledge Ms. C. Yang for instructive discussions and suggestions during this work.

7. REFERENCE

[1] Q. Li, B. Juang, Q. Zhou, and C. Lee, "Verbal Information Verification", *Proc. EUROSPEECH*, 1997, pp. 839-842.

[2] Q. Li, B. Juang, "Speaker Verification using verbal information verification for automatic enrollment", *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Seattle, WA, May 1998.

[3] Q. Li, B. Juang, "Automatic Verbal Information Verification for user authentication," *IEEE Trans. Speech Audio Processing* 8(5), pp. 585-596, 2000

[4] W. Reichl and W. Chou, "Decision Tree State Tying based on segmental clustering for acoustic modeling", *Proc. ICASSP'98*, pp. 801-804, 1998.

[5] Yiu Wing Wong, Large Vocabulary Continuous Speech Recognition for Cantonese, MPhil Thesis, The Chinese University of Hong Kong, 2000.

[6] Xiaolong Li, Ke Chen, "Mandarin Verbal Information Verification", *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Orlando, Florida, May 2002.

[7] Doulas, A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing* 10, 19-41 (2000).

[8] P. Rameshm, C. Lee and B. Juang, "Context Dependent Anti Subword Modeling for Utterance Verification", Proc. ICSLP' 96.

[9] Tan Lee, W.K. Lo, P.C.Ching and Helen Meng, "S Spoken Language Resource for Cantonese Speech Processing", in *Speech Communication*, vol.36, No.3-4, pp. 327-342, March 2002.