Speaker Verification Using Speaker-Specific Prompts

Yongxin Zhang, Adel Iskander Fahmy, Michael S. Scordilis

Department of Electrical and Computer Engineering, University of Miami, Coral Gables, Florida 33124
{yxzhang, fahmy}@umsis.miami.edu, m.scordilis@miami.edu

Abstract

Intra- and inter-speaker information, which include acoustical, speaker style, speech rate and temporal variation, despite their critical importance for the verification of claims, still have not been captured effectively. As a result of such modeling deficiency, existing speaker verification systems generally test claimed utterances with interfacing procedures that are common to all speakers. In this paper, a novel method is introduced in which speaker-specific attributes are expressed with reliable, first and second order intra-speaker and inter-speaker statistical information on the output space of speaker models in an explicit way. This is achieved through the computation of the Speech Unit Confusion Matrix (SUCM) that is employed in the scoring phase. An online updating procedure of SUCM is also presented. Experimental results with spoken alphabetic characters used as the basic speech unit indicate that the new method can improve system performance significantly. The method can also be directly extended to the use of other speech units (phonemes, sub-words, digits).

1. Introduction

In Speaker Verification (SV) systems, speaker models are constructed for each registered speaker in the enrollment phase. The likelihood of input speech being that of the claimed speaker is calculated using the speaker utterance, for both speaker and anti-speaker models, and it is used to confirm or deny the claimed identity (Bimbot et al. 2000; Reynolds and Rose, 1995; Higgins, Bahler and Porter, 1991). In general, when making a decision, the likelihood function is defined as $\vartheta(\mu \mid X)$, which is the probability of utterance μ belonging to speaker model X (correspondingly $\overline{X}$ for the anti-speaker model). On the other hand, the likelihood function also can be defined as $\vartheta(\mu \mid X, W)$, which is the probability of utterance μ belonging to speaker model X for spoken utterance W. This approach includes the expected linguistic content W.

In the approaches reported to date the setting of the likelihood ratio threshold may be speaker-dependent (SV) or speaker-independent (SI), i.e., $\Theta(\mu, X)$, but not text-related, $\Theta(\mu, X, W)$ (for example, Bimbot et al. 2000; Matsui and Furui, 1999). No consideration of the correlation between the text W and speaker X has been included in the above likelihood terms.

Matsui and Furui (1994) used tied-mixture HMMs to model speaker phonemes during training. Positive verification was indicated when it was decided that the true speaker correctly spoke the prompted text. Because SV usually relies on small amounts of training data, phonemes with little or no presence in the training speech were modeled by adapting universal phoneme models to registered speaker voices.

In the work of de Veth and Bourlard (1993) text-dependent and text-independent digit-based SV was tested using 32 context-dependent phoneme models. They demonstrated that for real-world SV applications with limited training data single Gaussian HMM is preferable to tied multi-Gaussian HMM, both in terms of model size as well as in achieved performance.

In the CAVE project (Bimbot et al. 2000), SV was carried out with more a practical objective, which was to provide the relevant framework for an actual application system through research and evaluation of the different technologies such as HMM, GMM, VQ, and DTW. That common framework in which telephone-based prototypes of SV were whole word HMM was investigated at great length. Good performance resulted by employing alternative solutions to particular problems such as the use of a common non-speech model, handling of untrained models, likelihood normalization by utterance length, and an efficient technique for adaptive variance flooring when the data in the enrollment sessions were insufficient.

Another contribution of CAVE was the formalization and solution of the issue of a priori threshold setting. The detailed discussion of a priori threshold setting was presented and exposed by comparing several different setting approaches. The insights provided not only in CAVE but also in the work of Matsui and Furui (1994) as well are of great interest. However, further investigation in this field is required is required in order to increase the efficiency of the underlying methods.

Statistical intra-speaker and inter-speaker information has always been used in speaker recognition systems for prior threshold setting and model adaptation (Bimbot et al. 2000; Reynolds and Rose, 1995; Matsui and Furui, 1999). Generally, the expectation and variance of model score are used for threshold setting (Chen, 2003). Another approach (Wang, Chen and Chi, 2002) employs inter-speaker information, such as physiological differences and manner of speech, for speaker identification.

In this paper, we present a new approach to improve the performance of SV system. We focus on how to select the spoken vocabulary required so that it best conveys user's acoustic characteristics. Thus is achieved by introducing (a) the so-called Speech Unit Confusion

Matrix (SUCM) and (b) the ranked verification Speech Unit (RVSU). The aim of this novel approach is to improve the scoring algorithm by increasing the correlation between the overall likelihood score and the utilized speech units with reliable first and second order speaker-specific statistical information. In this method, the likelihood function is a function of speaker X and prompted text W, that is, $\vartheta(\mu \mid X, W)$.

Section 2 introduces the new models, methods and the training procedure in our work. In Section 3, the testing experiments are described and results reported. Discussion and conclusion are in Section 4.

2. Motivations and methods

Our proposed approach is based on the fact that some speech units provide little or no discrimination between speakers and in fact they may lead to degradation of system performance, while at the same time other more discriminative units may be more effective. Thus, we will present a procedure, which identifies and such units and allow them to play a more important role in the decision while diminishing the effect of the less discriminative ones.

Observations from others studies and as well our earlier experiments show that reliable and useful statistical information for speaker recognition system are the means of the speaker and anti-speaker scores, and the variance of anti-speaker score. Typically, variance of speaker score is not included (Chen, 2003). Therefore, from a statistical viewpoint, we make an assumption that the best candidate speech unit for speaker verification should have statistically the best of speech features for the task at hand. We checked the relationship between discrimination of speech units for SV task and their first- and second- order statistics. , Our observations show that there exists an obvious relationship between discrimination ability and those statistics as shown in Figure 1 and Table 1. The best system performance is obtained from speech units with best statistics, and vice versa. With this motivation, we proposed a new method in speaker verification and evaluated it based using HMMs.

2.1. Speech models and features

The isolated whole word hidden Markov model for English alphabetical characters is used in our system. In this case, the speech unit is a spoken English alphabetic character. In total, there are 26 speech unit (character) models for each user. A database of feature vectors is collected for each frame in each word by computing 12^{th} order mel-cepstral coefficients. One left-to-right Gaussian-mixture HMM with 2 to 3 states per character and two Gaussians per state is used to represent each alphabetical character of each user.

2.2. Speech Unit Confusion Matrix and Ranked Verification Speech Units

By the nature of the SV systems, amplifying the difference between the true claimer and other users has been the aim of the proposed model. In general, there are two possible approaches: embedding considerations in the acoustic model such as accent, emotion or other intra-speaker related attributes (Wang, Chen and Chi, 2002), or using speaker-specific information at the system level (Roland and Parris, 1999, Ben Zeghiba and Bourlard, 2002). In this work, a Speech Unit Confusion Matrix (SUCM) is proposed as a new data set at the system level to achieve the objective. From the observation that for different phrases randomly selected from the alphabet, a large difference in the performance may result, extensive experiments were conducted, and one such example is shown in Figure 1. At first, we compute the following two qualities for each user in case of spoken alpha-character as the speech unit in the SV task: First, compute the log-likelihood $L(\mu_{jk} \mid M_{ik})$ for each utterance of character k from user j $\mu_{jk} : j = 1,2,\cdots N$ against the model of the character k for user i, M_{ik}. Second, compute the log-likelihood $L(\mu_{ik}^{m} \mid M_{ik})$ of each utterance m for character k from this user i, that is μ_{ik}^{m}, in the training set against the model of the character k for this user, that is, M_{ik}. Then, all likelihood values for the same character model of the user are ranked based on the average of two scores: the first one is the mean of the scores across the speaker pool, and the second one is the variance of scores across the training set. We normalize these 26 values and rank them from the most significant confusable character to the least significant one to form the Ranked Verification Speech Unit (RVSU) database for each user. In this case, the most confusable speech unit (character) has the largest chance of being falsely rejected or falsely accepted. The SUCM is built during the enrollment phase and it is used during the verification phase in order to provide a better choice of the prompted passwords, which contain characters that best convey the user's acoustic characteristics.

Therefore, for speaker i, the SUCM is:

$$SUCM(i,k) = \frac{1}{M}\sum_{m} L\{\mu_{ik}^{m} \Big| M_{ik}\} + \tilde{\delta}_{ik} \qquad (1)$$
$$\{i = 1,2,\cdots,N;\ k = 1,2,\cdots,26;\ m = 1,2,\cdots M\}$$

where $\tilde{\delta}_{ik}$ is the variance of score differentia L_{dif}:

$$L_{dif}(i,k) = \frac{1}{N-1}\sum_{j}[L(\mu_{ik} \mid M_{ik}) - L(\tilde{\mu}_{jk} \mid M_{ik})] \quad (2)$$

$$\{i, j = 1,2,\cdots,N;\ k = 1,2,\cdots,26;\ i \neq j\}$$

If the speech unit is the alpha-character, then M_{ik} is the k^{th} character model of speaker i, and μ_{ik} is the utterance of character k from speaker I, while $\tilde{\mu}_{ik}$ is the utterance of same character k but from different speaks. That is, we consider the first component of equation (1) for False Rejection (FR) and the second component, $\tilde{\delta}_{ik}$, for False Acceptance (FA). The size of the SUCM is an N*26 matrix in the case of character. Figure 2 shows the SUCM for a certain user.

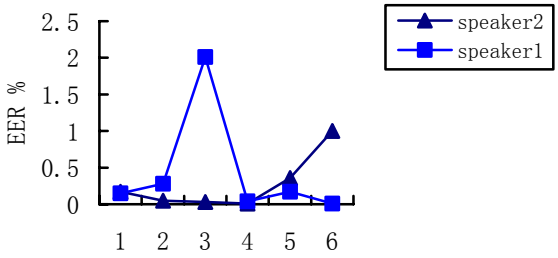

Figure 1. Performance of different test phrases (6 phrases) with alphabetic character as the speech unit in the SV task

Table 1. Normalized Statistics associated with different test phrases for speaker in Figure 1

Phrase #	1	2	3	4	5	6
$mean(x)$	0.805	0.95	0.34	0.99	0.81	0.72
$mean(\hat{x})$	0.01	0.02	0.31	0.01	0.33	-0.3
$\mathrm{var}(\hat{x})$	0.11	0.32	0.04	0.01	0.16	0.16

RVSU for the user i should be based on $SUCM(i,k)$:

$$RVSU(i,q): \{C_q, s_q\} \quad C_q \in \{A,B,C\cdots X,Y,Z\} \quad (3)$$

$$q = 1,2,\cdots 10 \quad s_1 > s_2 > s_3 > \cdots s_{10}$$

$$s_q = SUCM(i,k) \quad for\ character\ k \quad (4)$$

The RVSU is then used for selecting the verification password for that particular user and also as a weighting factor in the scoring phase.

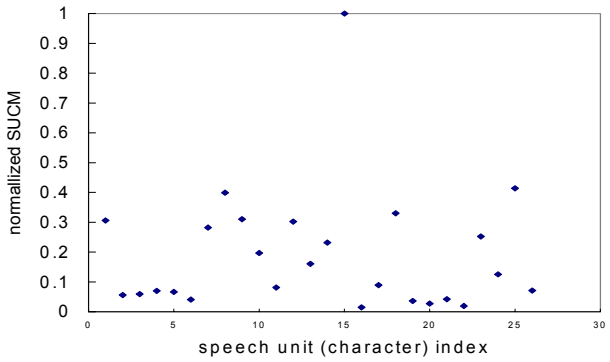

Figure 2. Speech Unit Confusion Matrix for a specific user

For each speaker, the first top ranking speech units are always selected. How many of them are employed in the system depends on the context of the application. In other words, the SV system uses the RVSU to select the prompted password for the user that best convey his/her particular acoustic characteristics.

2.3. Testing claimed utterance with RVSU

The most common performance measure referred to in the SV literature is the Equal Error Rate (EER). This involves an important step in SV, which is the application of a posteriori threshold T_{EER}. The likelihood of utterance μ for a given claimed speaker X and anti-speaker $\overline{X}$ are computed as likelihoods of the speech segment for the sequence of word models that compose the expected linguistic content of the utterance W. The likelihood function $\vartheta(\mu \mid X,W)$ is the estimation of $P(\mu \mid X,W)$ which is the PDF of the speaker-dependent distribution of the acoustic features for the sequence of W. Generally, for setting the decision thresholds the dependence on W has not been considered before, and it was just used as $\vartheta(\mu \mid X)$. With the assumption of equiprobable a priori claimer and imposter distributions and a priori equal cost of FA and FR, the acceptance decision strategy has been:

$$\frac{\vartheta(\mu \mid X)}{\vartheta(\mu \mid \overline{X})} > \Theta \quad (5)$$

Here, the selected a Speaker-Dependent (SD) EER method provides better performance in SV. W is included into the consideration of the decision procedure, not directly in the PDF or likelihood computation but in the decision likelihood function as $\vartheta(\mu \mid X,W)$. The score for each speaker is evaluated by multiplying the likelihood by its corresponding weighing factor from the RVSU database for the claimed user as

$$\vartheta(\mu \mid X, W) = \frac{1}{N}\sum_{i=1}^{N} c_i l_i \quad i = 1,2,\cdots,N \quad , \qquad (6)$$

where N is the number of characters in the verification utterance, l_i is the likelihood of the character i and c_i is the weighting factor from the RVSU database for character i, and the score on anti-speaker model $\vartheta(\mu \mid \overline{X}, W)$ as well.

RVSU is used not only as a weighting factor in the scoring process but also for the selection of the verification password for a particular user. In other words, the SV system uses the RVSU to select the prompted user password that best convey his/her acoustic characteristics. Moreover, the system uses the RVSU database to improve the scoring algorithm.

2.4. Online parameters updating

The training of the statistical speaker model in our proposed method includes old issues and it tries to overcome new challenges. Variance limiting for the HMM has already been proposed elsewhere (Bimbot et al. 2000; De Veth and Gallopyn, 1993; Chen, 2003). However our observation was that when the estimation of mean of mixture components fits one of the feature vectors in the M step (of the EM algorithm), the probability of other vectors in the E step will become too small and they may cause singularities in the covariance matrix. Therefore, selecting the flooring of probability as a means for the survival of the training procedure becomes a critical problem. Here, we apply a "momentum" to push the next estimation of the mean of the mixture off the overfitting point. In our method, we use 0.005 as a universal value for variance limiting and 0.0001 for probability limiting.

Another challenge from the online updating of SUCM was overcome by updating the statistics of system only with the history data and incoming imposter sample n+1 and claimer data sample m+1 as in the work by Chen (2003), that is:

$$mean_{M+1} = \frac{m * mean_M + L_{M+1}}{m+1} \qquad (7)$$

$$\widetilde{mean}_{n+1} = \frac{n * \widetilde{mean}_n + \widetilde{L}_{n+1}}{n+1} \qquad (8)$$

$$\widetilde{\delta}_{n+1}^2 = \frac{n(n+1)\widetilde{\delta}^2 + n(\widetilde{L}_{n+1} - \widetilde{mean}_n)^2}{(n+1)^2} \qquad (9)$$

Details on the proof of this procedure are found in the work Chen (2003). This will hold true both for the HMM case and the GMM case. With these updated statistics, the SUCM for the character case is updated by:

$$\widetilde{SUCM}(i,k) = mean_{M+1} + \widetilde{\delta}_{n+1} \quad \{i = 1,2,\cdots,N;\ k = 1,2,\cdots,26;\ m = 1,2,\cdots M\} \qquad (10)$$

3. Experiments and results

All experiments were performed on the NIST TI-46 speech corpus. Each user has 26 records per character, in which are used 10 for training and 16 for testing. For each user we have 16^N FR testing samples, with each test phrase consisting of N characters. To ensure that the result of experiment are within an 80% confidence level, we referred to the work by Higgins et al. (1991). We selected the three-letter phrase in the experiment, that is, for each test phrase, the size of FR experiment is 4096 and the size of cross gender FA experiment is 61440 (in same gender case, it is 28672). The anti-speaker model (world model) is build from the speaker collection of TI-46 and extra collection of 10 speakers.

When testing a claimed identity, just as in other approaches, the user first claims an identity; the system consults the RVSU database for the claimed identity, and then prompts the user for a verification utterance. The procedure is the same as in the training phase: the likelihood of the input word against the trained word model was computed and this process is repeated for each character in the verification utterance. Finally, the score obtained for each character in the verification utterance is weighted with a factor obtained from the normalized RVSU database of the claimed identity.

Table 2 shows the percentage improvement in the EER by applying the SUCM to three testing phrases for each user. Significant improvement was shown for each user when compared to the general EER without SUCM.

Table 3 shows the comparison of results for four types of EER%, as described in the work of CAVE [2]. For same gender experiments or cross gender experiments SUCM shows its ability to improve performance.

4. Discussion

Performance improvement by this new approach for speaker verification is self-evident. The SUCM provided the best reference to test text selection and scoring in the sense that the characters selected are most unlike to be recognized as another character produced by the user. We also improve the scoring algorithm by increasing the correlation between the overall likelihood and the higher ordered speech unit (character). As a result, the False Rejection and False Acceptance Rates of the system are simultaneously decreased. However, it is not yet clear as to whether the SUCM could make the claimed user more distinct among the user-set when they use the same or almost same test utterance. Another possible issue arises when the size of user set increases. Then, it is possible that there maybe more than one users which have similar or even identical SUCM index. Resolving this source of potential confusion is important.

Nevertheless, the method can be easily extended to other speech units, such as phonemes, syllables, and digits, in conjunction with a speech recognizer.

The proposed approach does increase the processing requirements with the computation of SUCM in the enrollment phase. For N users we need at least N*26 calculation of HMM likelihoods before we get the SUCM and RVSU for users. In addition, the SUCM also needs to be updated after several new verification passes by each user.

Table2. EER% performance comparison by using the SUCM for each user

User	Same gender Test		Cross gender Test	
	Without SUCM	With SUCM	Without SUCM	With SUCM
F1	0.78	0.04	0.98	0.08
F2	0.09	0	0.74	0
F3	0.325	0.01	0.47	0.74
F4	0	0	0.01	0
F5	0.78	0	0.57	0.07
F6	0.0375	0.05	0.27	0.23
F7	0.0625	0.0	0.06	0.01
F8	0.0001	0.0	0.11	0
M1	0.0575	0.0	1.28	0.06
M2	0	0	0.09	0
M3	2.887	3.112	7.21	7.98
M4	0.946	0	0.29	0.10
M5	0	0	0.82	0
M6	3.3	2.95	0	0.20
M7	0.236	0.03	0.31	0.22
M8	2.54	0.75	0.87	0.21

Table 3 EER% performance for the proposed system with HMM as speaker model

	Same Gender				Cross Gender	
	MM EER		FF EER		EER	
	Without SUCM	With SUCM	Without SUCM	With SUCM	Without SUCM	With SUCM
EER %	1.25	1.01	0.259	0.12	0.88	0.62

References

Auckenthaler, R.; Parris, E.S.; 1999: Improving a GMM speaker verification system by phonetic weighting, *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, pp.313-316.

BenZeghiba, M.F.; Bourlard, Herve; 2002: User-customized Password Speaker verification Based on HMM/ANN and GMM verification, *Proceedings of the 2002 International Conference on Spoken Language Processing, ICSLP 2002*, Vol.3, pp. 1325-1329.

Bimbot, F.; Blomberg , M.; Boves , L.; Genoud , D.; Hutter , H.P.; Jaboulet, C.; Koolwaaij, J.; Lindberg, J; Pierrot, J.B.; 2000: An overview of the CAVE project research activities in speaker verification, *Speech Communication,* Vol.31 pp.155-180.

Chen, K.; 2003: Towards better making a decision in speaker verification, *Pattern recognition*, Vol. 36, pp.329-346.

De Veth, J.; Gallopyn, G.; Bourlard, H. *1993:* Limited parameter hidden Markov models for connected digit speaker verification over telephone channels**,** *Proceedings of the 1993 IEEE International Conference on Acoustics, Speech and Signal Processing,* Vol.2 pp. 247 –250.

Matsui, T.; Furui, S; 1994: Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition**,** *Proceedings of the 1994 IEEE International Conference on Acoustics, Speech and Signal Processing,* Vol. I, pp.125 -128.

Higgins, A.; Bahler, L.; and Porter, J.; 1991: Speaker Verification Using Randomized Phrase Prompting. *Digital Signal Processing*, Vol. 1, p. 89-106.

Matsui, T.; Nishitani, T.; Furui, S.; 1999: Study of models and a priori threshold updating in speaker verification, *Systems and Computers in Japan*, Vol. 30, N. 13, Nov. 1999, pp. 96-105.

Douglas A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models", *Speech Communication*, V.17, pp. 91-108, 1995.

Reynolds, D.A.; Rose, R.C.; 1995: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models, *IEEE Transactions on Speech and Audio Processing*, Vol.3, No1. pp. 72-83.

Rosenberg, A.; Soong, F.; 1987: Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. *Computer Speech and Language*, Vol. 22, 1987, pp. 143-157.

Rosenberg, A. E.; Lee, C.-H. and Gokcen, S; 1991: Connected Word Talker Verification Using Whole Word Hidden Markov Models, *Proceedings of the 1991 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-91*, 1991, Vol.2, pp. 247 –250.

Wang, L.; Chen, K.; Chi, H; 2002: Capture Interspeaker Information With a Neural Network for Speaker Identification, *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp. 436-445.