PASSWORD-DEPENDENT SPEAKER VERIFICATION USING QUANTIZED ACOUSTIC TRAJECTORIES

Luc Gagnon, Peter Stubley, and Ghislain Mailhot

Locus Dialogue
460, Sainte-Catherine Ouest
Montréal, Québec, Canada H3B 1A7
{Luc.Gagnon,Peter.Stubley,Ghislain.Mailhot}@locusdialogue.com

ABSTRACT

Speaker verification requires either two steps (identity claim and verification) or the use of speech recognition to determine the password phrase. The single step method using speech recognition is text- and language-dependent. We describe a novel single-step method based on Gaussian mixture models and quantized acoustic trajectories that does not use any linguistic knowledge and is thus text- and language-independent. Although a two-step process can be more accurate, our approach is significantly better than speaker identification and is more convenient than a two-step process.

1. INTRODUCTION

In speaker verification (SV), the authentication process usually requires two steps, an identity claim followed by verification that the speaker is who they claim to be. It is more convenient for users to simply specify a password and have the password phrase both identify and verify the speaker.

In previous work, this has been accomplished with one of two methods. The first uses speaker identification, but this approach can be expensive and less accurate, even when the number of subscribers is only moderately large. The second approach uses automatic speech recognition (ASR) with speaker-independent models to recognize the password phrase (such as the subscriber's name or PIN) and thus the speaker [1, 2]. The speaker verification algorithm is then applied to the phrase. This approach is effective and efficient, but is text- and language-dependent (the words of the password phrase must be including in the ASR's vocabulary).

In this work, we describe a method based on quantized acoustic trajectories (QAT) that does not rely on linguistic knowledge. The QAT approach has the benefits of the ASR approach, but is text- and language-independent. Thus, users are not restricted in their choice of password, and indeed, do not even need to use words. As long as they can consistently reproduce the sounds, any sequence of sounds can be used as the password phrase.

A Gaussian mixture model (GMM) forms the basis of the algorithm. The GMM is used to both determine the QAT and score the utterance to validate the speaker. In the next section, the training and verification processes are described. Section 3 gives some experimental results and discussion. Section 4 gives our conclusions and outlines future work.

2. THE SPEAKER VERIFICATION ALGORITHM

The Quantized Acoustic Trajectory Gaussian Mixture Model (QATGMM) technique builds voiceprints by capitalizing on two components. The first component is a biometric voiceprint made of a parametric Gaussian mixture model (GMM) [3] trained on acoustic features using Bayesian adaptation [4]. The second component is a quantized acoustic trajectory which models the acoustic path of the user-selected password phrase. The most likely mixture sequence which quantizes the acoustic vectors is termed the quantized acoustic trajectory. The QAT actually models the password phrase without using any linguistic knowledge. The QAT is therefore text and language independent. Subscribers could in fact use any sounds to make their password and are not limited to phones associated with languages.

2.1. The GMM model

The seed acoustic model (SAM) is a 512-state GMM trained on approximately 10 hours of speech, using a combination of K-means clustering and the EM algorithm. The feature vectors are cepstral coefficients determined by a 32ms Hamming window, a 256-point FFT resulting in 32 linear frequency cepstrum coefficients. The frame advance rate is 8 ms. As well, cepstral mean subtraction is used.

2.2. Voiceprint model creation (biometric training)

The SAM is used as the prior distribution for MAP training of the subscriber's biometric voiceprints. It also serves as the imposter model for score normalization in the SV scoring stage. As well, the SAM is used to determine the QAT.

The SAM is described as a mixture of M Gaussian pdfs (or states), each with a diagonal covariance matrix:

$$\lambda \;=\; \{p_i, N(\mu_i, \sigma_i^2), i = 1, \ldots, M\},$$

where p_i is the probability of state i, μ_i is the mean, and σ_i^2 is the diagonal covariance matrix.

The biometric voiceprint for a speaker is created by performing MAP adaptation of the initial SAM with a sequence of acoustic observation vectors generated from the training speech samples of the speaker,

$$C \;=\; \{C_1, C_2, \ldots, C_T\},$$

where C_t is the normalized cepstral vector for frame t. A fixed relevance factor r is used to adapt the mixture weights, means, and variances.

The outline of the MAP process follows the development in [3, 4]. For state j, the SAM provides the prior parameters $\{p_p, \mu_p, \sigma_p^2\}$, where p_p is the mixture weight, μ_p is the mean, and σ_p^2 is the vector corresponding to the diagonal covariance matrix. The probabilistic count

$$n \;=\; \sum_{t=1}^{T} \Pr(j|C_t)$$

defines the total likelihood of state j, given the sequence of speaker observations.

The adaptation coefficient α is defined by

$$\alpha \;=\; \frac{n}{n+r},$$

where r is the relevance factor. The relevance factor adds robustness to the training process and is typically set to a value between 1 and 25.

The adapted mixture weight p_a is given by

$$p_a \;=\; \gamma[\alpha p_s \;+\; (1-\alpha)p_p],$$

where $p_s = n/T$, and γ is computed over all states to ensure that the mixture weights sum to 1.

Similarly, the adapted mean μ_a is

$$\mu_a \;=\; [\alpha\mu_s \;+\; (1-\alpha)\mu_p],$$

where

$$\mu_s \;=\; \frac{1}{n}\sum_{t=1}^{T} C_t \Pr(j|C_t).$$

Finally, the adapted variance σ_a^2 is given by

$$\sigma_a^2 \;=\; [\alpha\sigma_s^2 \;+\; (1-\alpha)(\mu_p^2 + \sigma_p^2)] \;-\; \mu_a^2,$$

where

$$\sigma_s^2 \;=\; \frac{1}{n}\sum_{t=1}^{T} C_t^2 \Pr(j|C_t).$$

The adaptation equations are applied to each state of the SAM to produce the adapted model λ_x for speaker x.

The one-pass MAP technique has several advantages:

- MAP is robust to limited training data. When a state has a low probabilistic count n, there is a de-emphasis of the new (potentially under-trained) parameters and an emphasis of the prior parameters. Conversely, when n is large, the new parameters are emphasized, and the prior parameters are de-emphasized.
- it is robust to noisy training data, due in part to the relevance factor. Before a state is updated, it must see a certain number of similar acoustic events. Bad, or out-lying, vectors are softly rejected.
- it can be used to adapt the acoustic models to new environments with limited data. MAP can be re-run on the new data to re-adapt the SAM.
- the SAM provides a *de facto* cohort, or world, model.
- it allows fast computations of the relevant states. This way, a system can afford more states for a given computational limit.
- the SAM provides a technique to quantize acoustic trajectories.

2.3. Voiceprint model creation (acoustic trajectory training)

A standard GMM is password independent because it learns the acoustic properties of the speaker's voice without making any assumptions about the order in which observation vectors will be observed. To create a password-dependent system, the model must encode the sequence of the observations that are typical of the password phrase.

To add the coding of the password phrase to a GMM, we use the quantized acoustic trajectory (QAT). The QAT is the most likely sequence of SAM states, given the observation vectors:

$$QAT(C) \;=\; \{I_1, I_2, \ldots, I_T\},$$

where I_t is the index of the state that minimizes the Mahalanobis distance to the observation C_t. This is equivalent to using vector quantization, with the states as the codebook and the Mahalanobis distance to choose the appropriate entries. Each training sample provided by the speaker is used to create a QAT.

The password-dependent voiceprint consists of the adapted SAM voiceprints and the QATs. For speaker x, we get the composite voiceprint model

$$\Lambda_x \;=\; \{\lambda_x, QAT_x\}.$$

2.4. Pattern recognition and scoring

To perform the verification, an unknown speaker provides a speech sample,

$$C_o \;=\; \{c_1, c_2, \ldots, c_T\},$$

assumed to be a password phrase for the speaker. The password phrase provides the identity of the speaker as well as the properties used to verify the speaker. In order to be accepted, it must be the speaker who says the password. The observation vectors, as with training, consist of 32 normalized linear frequency cepstral coefficients, and T is the number of vectors after noise and silence frames have been removed.

The first step is to quantize C_o with the SAM to obtain the QAT of the incoming utterance. Then, a fast match algorithm is used to select a small set of QATs that best match the test utterance. The scoring uses dynamic programming where the local distance between elements of the QAT indices are provided by table lookup. This table is precomputed using the Mahalanobis distances between the SAM state means.

The scoring algorithm is a DTW-based approach. As with the training process, the speech samples C_o are converted to the QAT, I_o. From the registration process, there is a set of reference QATs, $R_s,\ s = 1, \ldots, S$, where S is the number of known speakers. Each reference QAT is scored against I_o, and the K best-matching QATs, $R_x,\ x \in \{x_1, x_2, \ldots, x_K\}$, are chosen. For the DTW scoring algorithm, the local distance measure, $d(I_t, R_{x,j})$ is derived from the Mahalnobis distance between the means of states I_t and $R_{x,j}$. This scoring is very fast, because the local distances can be precomputed. As well, a hierarchical or clustered approach can be used to organize the reference QATs so that references that are obviously far from the speech sample need not be scored.

The adapted GMMs corresponding to the best-matching QATs are loaded and scored using likelihood ratio testing (LRT). For each subscriber x with a QAT in the short list, we compute the likelihhood ratio

$$\theta_x \;=\; \frac{\Pr(\lambda_x|C_o)}{\Pr(\lambda_{sam}|C_o)},$$

where λ_x is the GMM for speaker x. Using Bayes' rule,

$$\Pr(\lambda_x|C_o) \;=\; \frac{\Pr(C_o|\lambda_x)\Pr(\lambda_x)}{\Pr(C_o)}$$

and

$$\Pr(\lambda_{sam}|C_o) \;=\; \frac{\Pr(C_o|\lambda_{sam})\Pr(\lambda_{sam})}{\Pr(C_o)}.$$

Therefore,

$$\theta_x \;=\; \frac{\Pr(C_o|\lambda_x)}{\Pr(C_o|\lambda_{sam})},$$

because we do not use an a priori probability of x being an imposter. The likelihoods are calculated using the standard scoring of a GMM:

$$\Pr(c_t|\lambda_x) \;=\; \sum_{i=1}^{M} p_{x,i}\Pr(c_t|\mu_{x,i}, \sigma^2_{x,i}),$$

where $p_{x,i}$ is the mixture weight for state i and $\Pr(c_t|\mu_{x,i}, \sigma^2_{x,i})$ is the Gaussian pdf. Finally, the log-likelihood for the complete utterance is

$$\log\Pr(\lambda_x|C_o) \;=\; \sum_{t=1}^{T}\log\Pr(c_t|\lambda_x).$$

$\log\Pr(\lambda_{sam}|C_o)$ is computed in a similar manner.

Finally, from amongst the K candidate speakers, we select speaker x_i as the most likely if

$$\log\Pr(\lambda_{x_i}|C_o) > \log\Pr(\lambda_{x_j}|C_o),\ \forall j \neq i$$

. The speaker is accepted if

$$\log\Pr(\lambda_{x_i}|C_o) \;-\; \log\Pr(\lambda_{sam}|C_o) >= \theta,$$

where θ is a threshold typically $>= 0$, and is chosen to satisfy the required level of security.

3. EXPERIMENTAL RESULTS

To perform our experiments, we used the Polycost 250 database. A set of four baseline experiments (BE) has previously been defined for Polycost to provide a common ground for speaker recognition experiments and to enable-cross-site comparison [5, 6, 7]. We wanted to evaluate the QATGMM on a password-dependent task, so we decide to use the only BE1 experiment that is a text-dependent speaker verification with a fixed sentence. Each speaker is saying the following sentence: "Joe took father's green shoe bench out."

The regular Polycost baseline experiments used 110 of the 134 speakers from the database to create models. Because we have some restriction on noise level and on the similarities of the four different utterances from the same speaker, we were able to use only 91 speakers from the 110 defined speakers.

For all the speakers, we created a speaker verification model and a QATGMM model. We used 4 utterances to

N	EER
1	3.71%
2	4.44%
3	4.75%
5	5.98%
10	6.44%

Table 1. Equal error rates for the QATGMM where N is the number of best matching QATs that are verified.

create each of the models as defined in the BE1 experiment specification. For the speaker verification only one model was created, but for the QATGMM models, each of the four utterances becomes a specific trajectory. All the QATGMM models (the four trajectories) for the same speaker are kept inside the same file.

Table 1 shows the results of combined identification and verification using the QATGMMs. As described above, each QATGMM is scored against the test utterance. The N speakers with the QATGMMs closest to the utterance are then verified. The model with the highest score is chosen as the speaker, and the speaker is accepted if and only if the score (normalized by the SAM score) is above a threshold. The table shows the equal-error rate (EER) for several values of N.

For the sake of comparison, we implemented a single-step verification using speaker identification. The same GMMs were used, but without the filtering of speakers using the QATs. The resulting EER was 8.33%, double the EER with the QATGMM approach.

Finally, to determine the upper bound, we assumed that we could identify the speaker with perfect accuracy (similar to a two-step verification process). In this case, the EER with the same GMMs is 1.47%, significantly better than the single-step approach. However, we believe the performance of the QATGMM is sufficient for many applications, and that the single-step approach is more convenient for users.

Note that this task is somewhat pessimistic, since all speakers are using the same password. We believe, in this case, that the performance of a system using ASR to do the selection would perform at approximately the same level as the speaker identification approach. This is because the QATGMM approach includes some speaker dependence in the QAT that is used to select the most likely candidates. The ASR-based approach, on the other hand, uses speaker *independent* models to perform the candidate selection.

4. CONCLUSIONS AND FUTURE WORK

We have described a novel single-step identification and verification system. Our approach is password-dependent, but unlike ASR-based approaches, the QATGMM method is language independent since it does not rely on linguistic knowledge. We have shown that the QATGMM method has half the EER of simply using speaker identification. Although two-step verification (with an indentity claim step followed by a separate verification) has better performance, the single-step approach is more convenient for users and achieves sufficiently good results for many applications. As well, the speaker-dependent nature of the QAT will be a benefit in applications where a significant number of users share the same password phrase.

Future work includes optimization of the algorithms to improve their efficiency. As well, more comparisons with ASR-based approaches are needed.

5. REFERENCES

[1] A.E. Rosenberg and S. Parthasarathy, "Speaker identification with user-selected passwords," in *EUROSPEECH 97*, 1997.

[2] L. Rodriguez-Linares and C. Garcia-Mateo, "A novel technique for the combination of utterance and speaker verification systems in a text-dependent speaker verification task," in *ICSLP 98*, 1998.

[3] D.A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *EUROSPEECH 97*, 1997.

[4] J.L. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Transactions on Speech and Audio Proc.*, 1994.

[5] D. Petrovska, J. Hennebert, H. Melin, and D. Genoud, "Polycost: a telephone-speech database for speaker recognition," in *Speaker recognition and its commercial and forensic applications*, 1998.

[6] H. Melin and L. Lindberg, "Guidelines for experiments on the polycost database," in *Application of speaker recognition techniques in telephony*, 1996.

[7] T. Nordstrom, H. Melin, and J. Lindberg, "A comparitive study of speaker verification systems using the polycost database," in *ICSLP 98*, 1998.