Unsupervised Learning of HMM Topology for Text-dependent Speaker Verification

Ming Liu, Thomas Huang

Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801
{mingliu1,huang}@ifp.uiuc.edu

Abstract

Usually, text-dependent speaker verification can achieve better performance than text-independent system because of the constraint that the enrollment and testing utterance share the same phonetic content. However, the enrollment data for text-dependent system usually is very limited. Expectation Maximization(EM) training of HMM will suffer from noisy estimation because of limited enrollment. Adaptation is a popular solution in this scenario. The target model is formed by adapting the generic model based on limited speaker specific training data. Although the adaptation scheme can tolerate much less training data than direct EM method, the traditional method does not account the topology of HMM might be different for different speaker. The topology information further distinguish the target speaker from impostors. In this paper, we propose a unsupervised learning method to learn the topology of HMM for each speaker. The experimental results indicate that with learning the topology, the framework is more effective than traditional adaptation methods. In the pure acoustic matching experiments, the proposed method is the best system under extremely small amount enrollment data(1 training utterance) and moderate training data. That mainly due to explicitly including the label information in background modeling and discriminant capability of unsupervised learning of HMM topology.

Index Terms: speaker verification, HMM topology, unsupervised learning.

1. Introduction

Text-dependent(TD) verification systems usually has better performance than Text-independent(TI) verification system. With the constraint that the same phonetic content has been present in enrollment and test utterances, the TD verification can match the acoustic features of the same phonetic units, which leads to a significant boost in performance compared to text-independent system. The most popular methods for text-dependent speaker verification can be classified into two categories: Dynamic time warping(DTW) based methods and Hidden Markov Model(HMM) based methods [1][2][3][4]. In DTW-based systems, the training utterances are stored as templates. The similarity measure of the testing utterances is the global matching distance between testing utterances and templates via dynamic programming. Generally speaking, DTW-based system is efficient only when the training utterances is few. HMM-based system represent the speaker characteristics by a HMM model which usually contains several states(3∼10 states). Since the HMM has fewer states than the frames of each utterances, HMM-based system is faster than DTW-based system in term of evaluation speed. Moreover, it can easily incorporate multiple enrollment data via EM algorithm. In spite of many advantages, HMM-based system require sufficient enrollment data for good estimation of model parameters. Insufficient enrollment data will degrade the estimation of model parameters via EM algorithm. As well known in speech recognition literature, MAP adaptation can improve the model quality when training data is insufficient to perform EM algorithm. In traditional HMM-based system, the configuration of HMM is set to be the same among all speakers. However, the topology of the HMM should be different among different speakers which actually encoding more discriminative information for speaker verification. In this paper, two unsupervised methods are proposed to learn the HMM topology for different speaker. Then, MAP adaptation is applied to refine the model parameters.

The paper is organized as following. Experiments and results are shown in section 4. Conclusions are in section 5.

2. Baseline System

In this paper, we investigate three baseline systems: DTW, HMM with direct EM, and HMM with MAP adaptation. The target model of first two method is trained individually. However, the third method require the background modeling before generate the target model.

2.1. Background Modeling

In MAP adaptation of speech recognition system, the background model is a set of HMMs trained via EM from the speech of large amount of speakers. It is also considered as a speaker-independent model or world model. Each HMM describes an average pattern of each unit(phoneme, word, etc.). MAP adaptation is performed on the background model to generate speaker-dependent patterns which is considered as speaker model. In HMM-based background model, the same HMM topology, usually left-to-right topology, is shared among all speakers which is not optimal to discriminate the different speaker. The HMM topology of different speaker should be different to encode the unique speaker characteristics.

GMM-based background model is a dominant background modeling method in text-independent speaker verification literature. The GMM-based background model is also called Universal Background Model(UBM), which describes the overall acoustic feature space. Generally speaking, UBM is simpler than HMM-based background model, however UBM ignores the temporal information and the label information by pooling the training frames together. In this sense, the HMM-base background model has better description power than UBM model. However, the UBM do provide a good Gaussian pool for our unsupervised learning scheme.

In order to utilize the label information in GMM-based background model, local-UBMs are trained for every digits. The assembly of these local-UBMs is a local-UBM background model. Compared to single UBM modeling, the assembly of local UBMs has better Gaussian candidates because of the dividing the training frames according to the label files. Yet, it is much simpler than traditional HMM background modeling.

2.2. HMM-based Adaptation

In this paper, the experiments are conducted on a corpus contains connected digits, we model each digit with a HMM. The MAP adaptation for HMM is given by [5]. However, by primary experiments, the mean-only adjustment gives the best performance. The following experiments only show the results of mean-only adjustment MAP. The formula for mean-only adjustment is listed as following.

$$\hat{\mu}_i = \frac{\tau}{\tau + \gamma_i}\mu_i + \frac{\gamma_i}{\tau + \gamma_i}\bar{\mu}_i \qquad (1)$$

where γ_i is the occupation soft-count at state i. μ_i is the mean of background model at state i. $\bar{\mu}_i$ is the mean of the training data.

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} \gamma_i(t) x_t}{\sum_{t=1}^{T} \gamma_i(t)} \qquad (2)$$

Thus, if the occupation soft-count is small, the MAP estimation is close to the value of background model. In the following experiments, the factor $\tau = 1$.

3. Unsupervised Learning of HMM Topology

The assumption is that different speaker should have different HMM topology to further encoding the unique speaker characteristics. To implement the unsupervised learning method, the basic idea is to learning the structure over a pool of Guaranies. As mentioned before, the UBM model and local-UBMs provide a pool of Gaussians which cover the human speech acoustic feature space. To learn the structure of a HMM, firstly we use the training frames to prune the unnecessary Gaussians and learn the transition probability between the surviving Gaussians. In this case, the background model is used to specify the structure and the initial parameters of the HMM.

More specifically, an Index Transformation is performed to initialize the HMM. Given the UBM model and training utterance, the index transformation is to find the best Gaussian index sequence for the training utterance. The procedure is illustrated as following equations.

$$\begin{aligned} [i_1, i_2, ..., i_T] &= F([x_1, x_2, ..., x_T]|\lambda) & (3)\\ i_t &= \arg\max_{1 \le m \le M} w_m N(x_t; \mu_m, \Sigma_m) & (4) \end{aligned}$$

Where $F(\cdot)$ is the index transformation, x_t is the acoustic feature vector at frame t, w_m and $N(:, \mu_m, \Sigma_m)$ are the parameters of a Gaussian in the UBM model. i_t is the best index at frame t. After index transformation, the training utterance X_1^T is converted into a integer sequence I_1^T. The topology of HMM can be derived by analyzing this integer sequence via following equations.

$$P(k) = \frac{\sum_{t=1}^{T} f(i_t = k)}{T} \qquad (5)$$

$$P(m|k) = \frac{\sum_{t=1}^{T-1} f(i_{t+1} = m | i_t = k)}{\sum_{t=1}^{T} f(i_t = k)} \qquad (6)$$

where $f(i_t = k)$ is a indicator function. It equal to 1 if i_k equal to k and 0 if it is not equal. By estimating these probability, the structure of HMM model is so defined.

For local UBM background, these index sequence encodes the speaker information. Different speaker may have different index sequence for the same digit. For global UBM background, these index sequence encodes the speaker and digit information simultaneously.

Considering each Gaussian in the UBM model as a state, the UBM model can be treated as a HMM except the transition probability and the initial probability is not defined. The Gaussian index sequence is used to count the initial probability $P(k)$ and transition probability $P(m|k)$. These probabilities define the topology of the HMM. Furthermore, those states which has no observations are eliminated after processing the index sequence. In GMM-based adaptation, the survived states also encode the speaker characteristics. Different speakers has different survive states set. In spite of sharing the same HMM among all speakers in HMM-based MAP adaptation, the initial HMMs are different in GMM-based MAP adaptation which enhance the discriminant capability of the HMMs.

Each state of HMM is a single Gaussian which is similar with the stochastic trajectory model in [6]. After defining the HMM topology, the MAP adaptation is performed to generate the speaker model. In our experiments, we found that the mean-only adaptation give the best performance. We investigate two learning method: global UBM based method and local UBM based method. The only difference between these two methods is that the Gaussian pool is share for all digits in global UBM while the different Gaussian pool is available in local UBM method. Since the local UBM dividing the Gaussian pools according to the label information, so it is expected to outperform the global UBM based method.

4. experiments and results

To evaluate the proposed methods, the comparison experiments are conducted on a Internal corpus which contain 36 speakers(18 males and 18 females). There are around 2 minutes studio recording of isolate digits and continuous digits(phone number) per speaker. The frontend processing is done with HTK toolkits[10]. The final acoustic feature is MFCC and its derivatives ΔMFCC. The overall dimension is 36.

A set of experiments are designed to evaluate the proposed methods. The impostors are assumed to know the exact digits sequence(PIN) of the target speaker. These experiments are designed to measure the pure acoustic matching performance for text-dependent speaker verification. The experiments conditions include different numbers of training utterances and different numbers of digits in the utterances.

Table 1 shows the results of different methods when there is one training utterance. The results indicate that local-UBM method is the best performer overall. The HMM-based adaptation achieves the same level of performance as DTW system. HMM-EM system is the worst performer which is as expected because of limited enrollment data. The global UBM method is inferior than HMM-based system because the HMM-based background model has better description about the prior knowledge. As mentioned in section 2.1, the HMM-based background model incorporates the label knowledge of each training utterances. The global UBM method ignores the label information, hence, the global UBM model is not as precise as the HMM-based model.

Table 2 shows the results of different method when there is two training utterances. The results indicate that DTW is not able to catch up with all HMM-base systems. HMM-EM is still worse than the adapted HMM system. Still local UBM based method is the best performer in this experiment.

Table 3 shows the results of different method when there is five training utterances. Except for utterances contain one digits, the HMM-EM system has the similar performance as HMM-adaptation system which indicate that five training utterances is sufficient to achieve good estimation of model parameters via EM algorithm. The global UBM system outperform than HMM-adaptation system when the utterances contains more than one digit which suggests that the discriminant topology of HMM in global UBM system becomes a dominant factor since the training data is sufficient. By combining the advantage of global UBM system and HMM-adaptation system, the local UBM system achieve the best performance in these experiments.

5. Conclusion and Future direction

Although text-dependent speaker verification can achieve better performance than text-independent system, the enrollment data for text-dependent system usually is very limited. The traditional Hidden Markov Model(HMM) with Expectation Maximization(EM) training suffers from noisy estimation because of limited observations. MAP adaptation can partly solve this problem by introducing a background model. However, the HMM topology of different speaker have to the same in traditional adaptation framework. In this paper, two unsupervised learning methods are proposed to learn the HMM topology for further enhancing the descriminative capacity of HMM modeling. The experimental results indicate that these two methods are effective method to tolerate insufficient observations by incorporating the background modeling and HMM topology learning. The local UBM method overall is the best system under different training conditions. Given moderate training data, the global UBM method is able to outperform the HMM-based adaptation. The local UBM method is better than global UBM method because of explicitly including the label information in background modeling and discriminant capability in HMM initialization. In the future, we are planning to run the experiments on large scale database(200∼300 speakers) to verify the effectiveness of proposed learning method. Incorporating the traditional password security scheme to improve the final performance will be another direction in the near future.

6. Acknowledge

This work was partially supported in part by National Science Foundation Grant IIS 00-859890 and CCF 04-26627.

Table 1: Performance comparison of different methods on the first sets of experiments when there is one training utterance, the performance measure is Equal Error Rate(%). Note: the number of states and number of Gaussians in each state is tuned to the optimal at each training conditions. HMM-EM is the HMM baseline system, DTW is the DTW based system, HMM-Adapt. is the HMM-based background model and MAP adaptation system, UBM-(local) is the local-UBM based unsupervised learning method, UBM-(global) is the global-UBM based unsupervised learning method. "One digit" column represents that each utterance contain one digit. "Three digits" column represents that each utterance contain three digits. "Five digits" column represents that each utterance contain five digits.

$N_{tr} = 1$	One digit	Three digits	Five digits
HMM-EM	22.91	17.35	15.21
DTW	17.08	10.00	6.88
HMM-Adapt.	16.88	10.58	8.32
UBM-(local)	**14.37**	**9.38**	**6.67**
UBM-(global)	18.40	12.64	9.23

Table 2: Performance comparison of different methods on the first sets of experiments when there is two training utterances, the performance measure is Equal Error Rate(%).

$N_{tr} = 2$	One digit	Three digits	Five digits
HMM-EM	10.28	5.28	4.02
DTW	13.51	6.67	5.39
HMM-Adapt.	**9.34**	4.09	3.70
UBM-(local)	9.38	**3.96**	3.13
UBM-(global)	9.79	4.93	**3.05**

Table 3: Performance comparison of different methods on the first sets of experiments when there is five training utterances, the performance measure is Equal Error Rate(%).

$N_{tr} = 5$	One digit	Three digits	Five digits
HMM-EM	6.18	1.81	0.83
DTW	8.75	3.96	1.85
HMM-Adapt.	4.74	1.8	0.83
UBM-(local)	**4.37**	**1.18**	**0.42**
UBM-(global)	4.72	1.32	0.63

7. References

[1] S. Furui,"Cepstral Analysis Technique for Automatic Speaker Verification," IEEE trans. on ASSP vol. 29, no. 2, Apr. 1981

[2] G.R. Doddington,"Speaker Recognition-Identifying People by Their Voices," Proceedings of IEEE, vol. 73, no. 11, pp. 1651-1644, 1986.

[3] K. Yu, J. Mason, J. Oglesby,"Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation," IEE Vision Image and Signal Processing, vol. 142, no. 5, Page(s):313 - 318, Oct. 1995.

[4] J.M. Naik, L.P. Netsch, and G.R. Doddington, "Speaker verification over long distance telephone lines," In Proceeding of ICASSP, pages 524–527, 1989.

[5] J.L. Gauvain, C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," IEEE Transactions on Speech and Audio Processing, vol. 2, no. 2, pp. 291 - 298, Apr. 1994.

[6] I. Illina, Y. Gong, "Elimination of Trajectory Folding Phenomenon: HMM, Trajectory Mixture HMM and Mixture Stochastic Trajectory Model," In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing , vol. 2, pages 1395-1398, 1997.

[7] M. Hebert, D. Doies,"T-NORM FOR TEXT-DEPENDENT COMMERCIAL SPEAKER VERIFICATION APPLICATIONS: EFFECT OF LEXICAL MISMATCH," In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing , 2005.

[8] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn,"Speaker verification using adapted gaussian mixture models," Digital Signal Processing, 2000.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, pp. 1–38, 1977.

[10] "http://htk.eng.cam.ac.uk/," .