A Novel Framework of Text-independent Speaker Verification based on Utterance Transform and Iterative Cohort Modeling

Ming Liu, Huazhong Ning, Thomas S. Huang

IFP, Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL, 61801
[mingliu1,hning2,huang]@ifp.uiuc.edu

Zhengyou Zhang

Communication and Multimedia System group
Microsoft Research
Redmond, WA, 98052
zhang@microsoft.com

Abstract

A novel framework for text-independent speaker verification is proposed. The framework is based on a new interpretation of Universal Background Model. The UBM in our framework actually defines a transform which maps the variable length observation into a fixed dimensional supervector(*supervector space*). Each speech utterance is then mapped into a point in this supervector space. The similarity measure in this vector space is progressively refined via an iterative cohort modeling scheme. The experiments on NIST 2002 corpus show the effectiveness of this new framework. Overall the EER drops from the baseline system(with T-Norm) 9.21% to final improved system(without T-Norm) 8.07%. The new framework can effectively reduce the data dependence in the final output score which is clearly indicated in the second sets of experiments. The EER after T-Norm of final system marginally increases by relatively 1.73% compared to the EER of baseline system drops 16.12% relatively after T-Norm. Also, the relative improvement of DCF after T-Norm is marginal for the final improved system (2.47%) compared to 33.68% in baseline system. It clear shows that the iterative cohort modeling effectively reduce the data dependence of the final scores, so that T-Norm will not further improve the system performance. Also, the performance of novel frame clearly increases as the iteration grows which suggest that the framework progressively refine the similarity measure on the supervector space with the iterative cohort modeling.

Index Terms: speaker verification, utterance transform, iterative cohort modeling.

1. Introduction

Speaker verification is a procedure of verifying the claimed identity of a speaker based on the speech signal from the speaker(voiceprint). The major factors affecting the performance of speaker verification system are discriminative capacity of modeling method and robustness to different channel recording, utterance recording and different target speaker. A ideal system of speaker verification will give 1 if the trial is from the target speaker while 0 if the trial is from imposter speaker. The score of speaker verification system actually depends on target speaker and testing utterance. The speaker model essentially is derived from training utterance, we denote the score using $m(U_t, U_e)$ where $m(\cdot)$ is a similarity measure, U_t is the training utterance and U_e is the testing utterance. From this formulation, it clearly that the score $m(U_t, U_e)$ depends on U_t and U_e. To reduce the dependence of measure from U_t and U_e, [1] use background model (Universal Background Model) to form a ratio operation $LLR(U_e) = \log P(U_e|\lambda_1) - \log P(U_e|\lambda_0)$ where λ_1 is obtained via Maximum a Posterior (MAP) from the background model λ_0. This framework is the well known UBM-MAP in the literature. Although, theoretically the ratio operation can reduce the dependence of the final score measure to U_t and U_e. However, T-Norm proposed in [2] further reduces the dependence of the $LLR(X_2)$ using extended cohort modeling techniques. It suggests that the $LLR(U_e)$ still depends on the data U_t and U_e. In T-Norm, only test utterance are processed via cohort modelling method which imply that the T-Normed log likelihood score $\bar{LLR}(U_e)$ most likely depends on the target speaker which means that the different target speaker has different T-Normed score distribution.

In this paper, the background model is not considered as a Gaussian Mixture Model in traditional viewpoint. The UBM is viewed as a mapping function which transfer the variable length observation, feature sequences, into a fixed dimensional observation. Although we derive this idea from approximation the UBM-MAP framework system, it is very similar to the *Fisher mapping* applying on feature sequence given the background model[3][4]. Instead of kernel methods applied in [4], the similarity measure of the transformed supervector is carefully designed in our proposal to take account for the noise and lack of observation in the supervector. After transferring the variable length feature sequences into this fixed dimensional feature space – *supervector space*. The similarity measure of training and testing utterances become to a similarity measure between two supervectors. In this supervector space, it is intuitive to apply the cohort modelling techniques to reduce to data dependence of $m(U_t, U_e)$. However, in order to choose good cohort points, we need a good similarity measure to search in the cohort pool which make the problem be a chicken-egg problem. To circumvent this difficulty, a iterative cohort modeling framework is proposed to progressively fine tuning the similarity measure in the supervector space. In a result, the EER and DCF of the system is reduced gradually. The experiments shows the equal error rate drops from 9.21% of baseline system to 8.07%.

2. UBM-MAP framework

The UBM-MAP framework is a dominant method in the literature of Text-independent Speaker Verification. The UBM is a Gaussian Mixture Model(GMM) which serves as a background distribution

This work was supported by National Science Foundation Grant CCF 04-26627

of human acoustic feature space. It can be represented as follows:

$$
\begin{aligned}
P(x|\lambda) &= \sum_{i=1}^{M} w_i P_i(x|\lambda) && (1)\\
&= \sum_{i=1}^{M} w_i \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)} && (2)
\end{aligned}
$$

where x is the feature vector with D dimension and λ is the parameter of Gaussian Mixture Model. M is the number of Gaussian components in the model. Parameter λ includes the prior probability of each component w_i, the mean vector of each component μ_i and the covariance matrix of each component Σ_i. $P_i(\cdot|\lambda)$ denotes the likelihood function of the ith component which is a multivariate Gaussian in a GMM. For simplicity, the covariance matrix Σ_i is usually set to be a diagonal matrix to lower the computation load. The maximum likelihood(ML) estimation of the parameters can be obtained via EM algorithm[5].

In the UBM-MAP framework, the target speaker model is generated by the Maximum A Posterior (MAP) adaptation [6] [1]. The mean-only MAP adaptation was the best method compared with other types of MAP adaption such as the fully MAP adaptation [1]. After the target speaker model is generated, a log-likelihood ratio between the target speaker model and the UBM model is then used to evaluate testing utterances. The log-likelihood ratio is computed as follows

$$
LLR(U_e) = LLR(o_1^T) = \frac{1}{T}\sum_{t=1}^{T} \log \frac{P(o_t|\lambda_1)}{P(o_t|\lambda_0)} \quad (3)
$$

where (o_1^T) are the feature vectors of the observed utterance – test utterance U_e, λ_0 is the parameter of UBM and λ_1 is the parameter of the target model. Essentially, the verification task is to construct a generalized likelihood ratio test between hypothesis H_1 (observation drawn from the target) and hypothesis H_0 (observation not drawn the target). The UBM model is usually considered as a background model which provides a description of acoustic feature space. Therefore, the likelihood of testing utterances on this UBM model $P(o_1^T|\lambda_0)$ can serve as an estimation of $P(o_1^T|H_0)$.

3. Utterance Transform

In conventional UBM-MAP framework, UBM is a background model to describe acoustic feature space of human speech. Actually, the procedure of MAP adaptation suggest another view of the purpose of UBM. For mean-only MAP adaptation[1], the MAP procedure is listed as follow

$$
\begin{aligned}
\gamma(i|o_t) &= \frac{w_i P_i(o_t|\lambda)}{\sum_{j=1}^{M} w_j P_j(o_t|\lambda)} && (4)\\
\gamma(i) &= \sum_{t=1}^{T} \gamma(i|o_t) && (5)\\
\bar{\mu}_i &= \frac{1}{\gamma(i)} \sum_{t=1}^{T} \gamma(i|t) o_t && (6)\\
\hat{\mu}_i &= \mu_i + \frac{\gamma(i)}{\gamma(i)+\alpha}(\bar{\mu}_i - \mu_i) && (7)
\end{aligned}
$$

where $\gamma(i|o_t)$ is the posterior probability of ith component given the observation o_t. $\gamma(i)$ is the soft count of observations which belong to ith component. $\bar{\mu}_i$ is the sample mean of ith component given the observations sequence $U = (o_t)_{t=1}^T$ and $\hat{\mu}_i$ is the adapted mean of ith component from the background mean μ_i. The smooth factor $\frac{\gamma(i)}{\gamma(i)+\alpha}$ is design to incorporate the number of observations into the final adapted mean. The basic idea is that the adapted mean should rely on background mean if observation is few. If the observation is sufficient, the adapted mean should prefer the sample mean for better data fidelity.

By examining this procedure, we shall notice that the $(\gamma(i), \bar{\mu}_i)_{i=1}^M$ is the sufficient statistic of utterance $U = (o_1^T)$. In this sense, the UBM serves as a transform which maps the variable length observation $U = (o_1^T)$ to a fixed dimensional supervector $(\gamma(i), \bar{\mu}_i)_{i=1}^M$. And this supervector is the sufficient statistics of the speech utterance. To simply the following derivation, we define $\delta(i) = \bar{\mu}_i - \mu_i$ which is the adjustment between sample mean $\bar{\mu}_i$ and background mean μ_i. Now the sufficient statistic of speech utterance is $(\gamma(i), \delta(i))_{i=1}^M$. The vector space of this sufficient statistics are called *supervector space*. This *utterance transform* is listed as follows.

$$
\psi(U) = X = (\gamma(i), \delta(i))_{i=1}^M \quad (8)
$$

which map the utterance U to a supervector X.

Although the test utterance is not mapped into its sufficient statistic explicitly in UBM-MAP framework, the log-likelihood score of a test utterance can be bounded by a function only depends on the sufficient statistics of the training and testing utterances. The derivation is as follows.

$$
\begin{aligned}
\log \frac{P(o_t|\lambda_1)}{P(o_t|\lambda_0)} &= \log \frac{\sum_{i=1}^{M} w_i P_i(o_t|\lambda_1)}{\sum_{k=1}^{M} w_k P_k(o_t|\lambda_0)} && (9)\\
&\geq \sum_{i=1}^{M} \gamma(i|o_t) \log \frac{P_i(o_t|\lambda_1)}{P_i(o_t|\lambda_0)} && (10)\\
LLR(U_e) &\geq \frac{1}{T}\sum_{i=1}^{M} \gamma(i)(\hat{\mu}_i - \mu_i)^T \Sigma_i^{-1}(\bar{\mu}_i - b_i) && (11)
\end{aligned}
$$

where $\hat{\mu}_i$ is the adapted target mean of ith component and $\bar{\mu}_i$ is the sample mean of testing utterance, $b_i = \frac{\hat{\mu}_i + \mu_i}{2}$. So, this bound is a function only depends on the supervectors of training utterance and testing utterance.

Recall the *Fisher mapping* of a observation sequence is defined as

$$
\Phi(o_1^T) = \nabla_\lambda \log P(o_1^T|\lambda) \quad (12)
$$

In our scenario, the free parameter of UBM are component weights and component mean, so the fisher mapping of utterance defined by

$$
\begin{aligned}
\nabla_{w_i} \log P(o_1^T|\lambda) &= \frac{\gamma(i)}{w_i} && (13)\\
\nabla_{\mu_i} \log P(o_1^T|\lambda) &= 2\gamma(i)\Sigma_i^{-1}\delta(i) && (14)
\end{aligned}
$$

which is very similar to the sufficient statistic of speech utterance. Instead of treat $\nabla_{w_i} \log P(o_1^T|\lambda)$ and $\nabla_{\mu_i} \log P(o_1^T|\lambda)$ equally, we carefully design a similarity measure for the sufficient statistic to take account of noise and lack of observation. The measure is defined on the supervector space as

$$
\begin{aligned}
m_0(X_t, X_e) &= \frac{\sum_{i=1}^{M} w(\gamma_t(i), \gamma_e(i))\delta_t(i)\Sigma_i^{-1}\delta_e(i)}{\sum_{i=1}^{M} w(\gamma_t(i), \gamma_e(i))} && (15)\\
&= \frac{\sum_{i=1}^{M} w(\gamma_t(i), \gamma_e(i)) m_0^i(X_t, X_e)}{w(\gamma_t(i), \gamma_e(i))} && (16)
\end{aligned}
$$

where X_t and X_e are the supervector of training and testing utterances. The main two parts of this similarity measure are local score $m_0^i(X_t, X_e) = \delta_t(i)\Sigma_i^{-1}\delta_e(i)$ and $w(\gamma_t(i), \gamma_e(i)) = \frac{\gamma_t(i)}{\gamma_t(i)+\alpha}\frac{\gamma_e(i)}{\gamma_e(i)+\alpha}$. The first part is a scaled cross correlation which is designed to account for the similarity of adjustments at each Gaussian component. The second part is designed to incorporate the number of observations, the more features are observed at one Gaussian component, the more contribution of the corresponding cross correlation term is added into the overall score. Although $\gamma_t(i)\gamma_e(i)$ are valid candidate as weighting factor, $\frac{\gamma_t(i)}{\gamma_t(i)+\alpha}\frac{\gamma_e(i)}{\gamma_e(i)+\alpha}$ is found a better choice. In our experiments, $\alpha = 16$.

4. Iterative Cohort Modeling

As shown in previous section, the likelihood of UBM $\log P(U_e|\lambda_0)$ is a estimation of the likelihood of H_0 $\log P(U_e|H_0)$. Ideally, after ratio operation, $LLR(U_e) = \log P(U_e|\lambda_1) - \log P(U_e|\lambda_0)$ will be data independent. However, in practice this ratio operation does not reduce all the data dependence. T-Norm[2] extends *cohort modeling* method to further normalize this score which suggest that the cohort modeling is a effective scheme to reduce the data dependence of the output score. The cohort modeling is to estimate the likelihood of H_0 by a small representative set of models. The main issues of cohort modeling are similarity measure between speaker models, size of the cohort set and fusion of individual scores of the cohort set[7][8][9]. In [10], a robust local scoring function is proposed which indeed is a special case of the iterative cohort modeling method.

The basic problem of applying the cohort modeling technique in our framework is to find a appropriate similarity measure in the supervector space. First of all, to define a good similarity in this space, we need to select a cohort set for each speaker model which in supervector space is one sample point. To select a good cohort set, we need a good similarity measure. To solve this chicken-and-egg problem, we adopt a *iterative cohort modeling* scheme. Equation 15 is a similarity between two supervector in the transformed supervector space. This measure is used as an initial similarity measure. The initial cohort set C_0 is selected base on this initial measure. After obtaining the initial cohort set, how do we refine the similarity measure? The scheme we adopted is thresholding for normalizing the score of different components. Basically, the threshold of each component is computed as follows.

$$t_0(i) = \frac{1}{N_{cohort}} \sum_{m=1}^{N_{cohort}} m_0^i(X_m, X_t) \qquad (17)$$

where N_{cohort} is the number of cohort speakers, X_m is the supervector of the training utterance of mth cohort speaker and X_t is the supervector of the training utterance of target speaker and $t_0(i)$ is the threshold of ith component. After getting the threshold for each component, the local score $m_0^i(X_t, X_e)$ is refined to $m_1^i(X_t, X_e) = m_0^i(X_t, X_e) - (t_0^t(i) + t_0^e(i))/2$ where $t_0^t(i)$ is the threshold of training utterance and $t_0^e(i)$ is the threshold of testing utterance. By this formulation, we treat the training and testing utterance equally, and the testing utterance is also processed with the cohort modeling which unify the online cohort modeling and offline cohort modeling method in a principle way.

Now, the refined similarity measure is

$$m_1(X_t, X_e) \quad = \quad \frac{\sum_{i=1}^{M} w(\gamma_t(i), \gamma_e(i)) m_1^i(X_t, X_e)}{\sum_{i=1}^{M} w(\gamma_t(i), \gamma_e(i))} \qquad (18)$$

With this refined similarity measure $m_1(X_t, X_e)$, we can reselect a new cohort set C_1 and compute a new threshold $t_1(i) = \frac{1}{N_{cohort}} \sum_{m=1}^{N_{cohort}} m_1^i(X_m, X_t)$. Now the refined local measure $m_2^i(X_t, X_e) = m_1^i(X_t, X_e) - (t_1^t(i) + t_1^e(i))/2$ and the refined measure is

$$m_2(X_t, X_e) \quad = \quad \frac{\sum_{i=1}^{M} w(\gamma_t(i), \gamma_e(i)) m_2^i(X_t, X_e)}{\sum_{i=1}^{M} w(\gamma_t(i), \gamma_e(i))} \qquad (19)$$

This iteration can be repeated further, the empirical finding is that the improvement become marginal after $m_4(X_t, X_e)$. In our experiments, the $N_{cohort} = 10$ based on empirical searching. The iteration can be repeated as many as possible, however, we found after the fourth iteration, the performance can not be improved.

5. Experiments and Results

In order to show the effectiveness of the novel framework, the experiments are conducted on the NIST 2002 Speaker Recognition corpus [11]. The frontend processing is done with HTK toolkit[12] to extract MFCC+DeltaMFCC feature, the total dimension is 24. Feature warping[13] is applied after MFCC extraction. The UBM is a 1028 component Gaussian Mixture Model trained on NIST01 training set which contains 174 speakers and roughly 2 minute speech per speaker. These 174 speakers also serves as cohort speaker pool. For T-Norm[2], these 174 speakers serves as the T-Norm Speaker pool. The NIST 2002 corpus contains 330 speakers and 39105 trials. The training utterance of each speaker is a telephone conversation that lasts from 60 sec. to 120 sec. The testing utterance lasts from 3 sec. to 120 sec.

The baseline system is a UBM-MAP system with the log-likelihood ratio scoring and T-Norm version. In order to verify that the iterative cohort modeling can effectively reduce the data dependence, we also perform T-Norm for each similarity measure $m_i(\cdot)$. The argument is that if the measure is data independent, then T-Norm will not improve the system performance.

Table 1 shows the experimental results. The performance was measure with two criteria: equal error rate(EER) and minimum Detection Cost Function(DCF)[11]. Table 5 shows the results of different iterations compared to the baseline system. It turns out that the T-Normalization do improve the system performance in first few iterations. Comparing the T-Normed baseline system and $m_4(\cdot)$ without T-Norm system, the improvement over the T-Normed baseline is 12.38% in EER and 0.81% in DCF. Figure 1 shows the DET curves of T-Normed baseline and $m_4(\cdot)$ system(without T-Norm). The improvement of the novel framework is fairly consistent over most range of DET curve.

In order to further investigate the performance of the novel framework, we draw the T-Norm improvement of above experiments. Comparing the T-Norm improvement for $m_i(\cdot)$ system, we will able to find the detail structure of the new framework. If the T-Norm bring more improvement, the $m_i(\cdot)$ score has more data dependence. Ideally, from $m_0(\cdot)$ to $m_4(\cdot)$ the T-Norm improvement will drop to marginal. Table 2 shows the experimental results. Clearly, the T-Norm improvement drop from 23.68% to -1.73% in terms of EER and 33.24% to 2.47% in terms of DCF. This results verify our claim that the iterative cohort modeling method unify the offline and online cohort modeling together and reduce the data dependence in one principal way.

EER/DCF	w/o T-Norm	w T-Norm
Baseline	$10.98\%/52.23(10^{-3})$	$9.21\%/34.64(10^{-3})$
$m_0(\cdot)$	$14.61\%/65.47(10^{-3})$	$11.15\%/43.71(10^{-3})$
$m_1(\cdot)$	$9.06\%/41.22(10^{-3})$	$8.83\%/35.38(10^{-3})$
$m_2(\cdot)$	$8.28\%/36.35(10^{-3})$	$8.38\%/33.58(10^{-3})$
$m_3(\cdot)$	$8.25\%/34.87(10^{-3})$	$8.38\%/33.31(10^{-3})$
$m_4(\cdot)$	$8.07\%/34.36(10^{-3})$	$8.21\%/33.51(10^{-3})$

Table 1: Performance comparison between baseline and novel framework with different iterations

	Rel. Improvement(EER)	Rel. Improvement(DCF)
$m_0(\cdot)$	23.68%	33.24%
$m_1(\cdot)$	2.54%	14.17%
$m_2(\cdot)$	-1.21%	7.62%
$m_3(\cdot)$	-1.58%	3.7%
$m_4(\cdot)$	-1.73%	2.47%

Table 2: Relative performance improvement of novel framework after T-Norm with different iterations

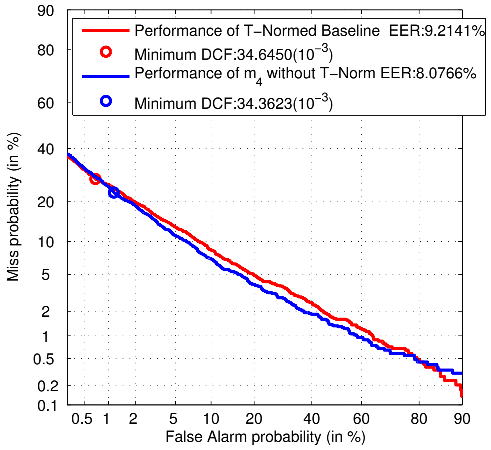

Figure 1: DET curves of T-Normed Baseline and $m_4(\cdot)$ without T-Norm

6. Conclusion and Future direction

In this paper, we propose a novel framework for text-independent speaker verification. The framework is based on the new interpretation of Universal Background Model. The UBM in our framework actually defined a transform which maps the variable length observation into fixed dimensional supervector. With this transform, each utterance is mapped into a point in the transformed supervector space. To define a good similarity measure in this space, an iterative cohort modeling scheme is adopted to progressively refine the similarity measure. The experiments on NIST 2002 corpus clearly show the effectiveness of this new framework. The new framework achieve 12.58% improvement on EER($9.21\% \rightarrow 8.07\%$) and marginal improvement on DCF($34.64(10^{-3}) \rightarrow 34.36(10^{-3})$). To further investigate the effectiveness of this novel framework, we draw a table to show the improvement after T-Norm. With iteration grows, after T-Norm, the improvement of EER drops from 23.68% to -1.73%, and the improvement of DCF drops from 33.24% to 2.47%. This results confirm that the new framework can further reduce the data dependence in the final output score as the iteration grows which suggests the iterative cohort modeling method is able to reduce the data dependence in one principal way by unifying the offline and online cohort modeling.

Since the UBM is used to define a mapping function, the Gaussian Mixture Model may not be the optimal choice. In the near future, we will investigate different clustering techniques to find a optimal choice to define the mapping. Also, the similarity measure $m4(\cdot)$ can be used as a kernel function and apply the well known support vector machine in the transformed supervector space for speaker verification.

7. References

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, pp. 19–41, January 2000.

[2] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, pp. 42–54, January 2000.

[3] Tommi S. Jaakkola and David Haussler, "Exploiting generative models in discriminative classifiers," in *Proceedings of the 1998 conference on Advances in neural information processing systems II*, Cambridge, MA, USA, 1999, pp. 487–493, MIT Press.

[4] Vincent Wan and Steve Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Transactions on Speech and Audio Processing*, pp. 203–210, March 2005.

[5] 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.

[6] J. Gauvain and C. Lee., "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions SAP*, pp. 291–298, 1994.

[7] E. H. C. Choi and J. Song, "Successive cohort selection(scs) for text-independent speaker verification," in *Proceeding, ICSLP 2000*, 2000, pp. 442–445.

[8] Y. Zigel and A. Cohen, "On cohort selection for speakerverification," in *Proceeding, Eurospeech 2003*, 2003, pp. 2977–2980.

[9] Tomi Kinnunen, Evgeny Karpov, and Pasi Franti, "Efficient online cohort selection method for speaker verification," in *Proceeding, INTERSPEECH*, 2004, pp. 2401–2404.

[10] Ming Liu, Zhengyou Zhang, and Thomas S. Huang, "Robust local scoring function for text-independent speaker verification," in *Proc. International Conference of Pattern Recognition*, 2006.

[11] "http://www.nist.gov/speech/tests/spk/," .

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

[13] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceeding, A Speaker Odyssey*, 2001.