EXPLORING DISCRIMINATIVE LEARNING FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

Ming Liu, Zhengyou Zhang, Mark Hasegawa-Johnson, Thomas S. Huang*

University of Illinois at Urbana-Champaign, Urbana, IL, USA
*Microsoft Research, Redmond, WA, USA
[mingliu1,jhasegaw,huang]@ifp.uiuc.edu
zhang@microsoft.com

ABSTRACT

Speaker verification is a technology of verifying the claimed identity of a speaker based on the speech signal from the speaker (voice print). To learn the score of similarity between each pair of target and trial utterances, we investigated two different discriminative learning frameworks: fisher mapping followed by SVM learning and utterance transform followed by Iterative Cohort Modeling (ICM). In both methods, a mapping is applied to map speech utterance from a variable-length acoustic feature sequence into a fixed dimensional vector. SVM learning constructs a classifier in the mapped vector space for speaker verification. ICM learns a metric in this vector space by incorporating discriminative learning methods. The obtained metric is then used by a Nearest Neighbor classifier for speaker verification. The experiments conducted on NIST02 corpus show that both discriminative learning methods outperform the baseline GMM-UBM system. Furthermore, we observe that the ICM-based method is more effective than the SVM-based method, indicating that the metric learning scheme is more powerful in constructing a better metric in the mapped vector space.

1. INTRODUCTION

Speaker verification is a technology of verifying the claimed identity of a speaker based on the speech signal from the speaker (voice print). The fundamental task of speaker verification is to assign a similarity score for each speech utterance pair, namely target utterance and trial utterance. To construct the similarity scoring function, many approaches are proposed in the literature. Among them, a generative probabilistic model – Gaussian Mixture Model (GMM) is a widely adopted method [1]. With this method, the likelihood of trial utterance on target GMM serves as a similarity between the pair. The likelihood ratio of trial between target model and Universal Background Model (UBM) is the similarity score in the GMM-UBM framework [2]. This ratio operation normalizes the range of the similarity score and results in a substantial performance improvement over the GMM method. Besides the normalization due to ratio operation, feature level normalization [3, 4] turns out to be also very effective for the system performance, as well as the score level normalization [5].

Although the generative method is rather successful in speaker verification task, the combination of generative and discriminative methods gains more and more research effort in recent years. This is one of the new trends in speaker verification literature. Among all different discriminative machine learning methods, Support Vector Machine (SVM) is a well established learning method to achieve optimal classification according to a predefined criterion. In addition to solid theoretical foundation, there are many successful practical applications of SVM on regular classification problems such as text classification, speech recognition, face detection and pedestrian detection. At the early stage, researchers tried to classify frame by frame and average the results for the final decision [6, 7], this scheme turns out not efficient for speaker recognition task. S. Fine and J. Navrátil [8] adopt the *fisher mapping* to map a whole utterance into a fixed dimensional vector and perform classification in this new vector space. The mapping step is a very crucial point for the success of the later SVM-based method. The underlying motivation of this mapping is to treat the utterance as a whole object, which leads to a similarity score on utterance pairs instead of frame pairs. V. Wan [9] extends the fisher mapping to score space mapping and achieves better improvement. Based on the same idea of obtaining similarity between an utterance pair, W. Campbell [10] uses generalized linear discriminant sequence kernels. J. Louradour and K. Daoudi [11] use a VQ-based kernel function for SVM learning. D. E. Sturim [12] maps the speaker utterance into a fixed dimensional score vector. Each element of this score vector is the likelihood on one anchor model. In this paper, we compare a metric learning-based framework – Iterative Cohort Modeling (ICM) [13] to the well known fisher mapping followed by SVM learning framework. In stead of fisher mapping, [13] adopted the sufficient statistics of speech utterance as a mapping function, which has similar formulation as fisher mapping. However, the major difference is that the soft count and sample mean are separated in sufficient statistics (see Sect. 4), which gives more flexibility in designing an optimal similarity function. In the ICM framework, the global similarity of an utterance pair is a weighted combination of local similarities which compare the statistics belong to the same component. Also, each local similarity is thresholded to normalize its contribution to the global similarity. The

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

experimental results on NIST 2002 corpus show the effectiveness of our framework: the EER drops from 10.98% to 8.07% and the DCF drops from $52.23(10^{-3})$ to $34.36(10^{-3})$.

2. BASELINE SYSTEM

The GMM-UBM framework is a very successful method in the literature of text-independent speaker verification. The UBM is a Gaussian Mixture Model (GMM) which serves as a background distribution of human acoustic feature space. It can be represented as follows:

$$P(x|\lambda) = \sum_{i=1}^{M} w_i P_i(x|\lambda) \quad (1)$$

$$= \sum_{i=1}^{M} \frac{w_i}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \quad (2)$$

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 In the UBM-MAP framework, the target speaker model is generated by the Maximum A Posterior (MAP) adaptation. The mean-only MAP adaptation was the best method compared with other types of MAP adaption such as the fully MAP adaptation. 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

$$LLR(U) = LLR(x_1^T) = \frac{1}{T}\sum_{t=1}^{T} \log \frac{P(x_t|\lambda_1)}{P(x_t|\lambda_0)} \quad (3)$$

where (x_1^T) are the feature vectors of the observed utterance – trial utterance U, λ_0 is the parameter of UBM and λ_1 is the parameter of the target model. In GMM-UBM framework, the verification task is essentially 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). As the background model provides a description of acoustic feature space, therefore the likelihood of a trial utterance on this background model $P(x_1^T|\lambda_0)$ can be used as an estimation of $P(x_1^T|H_0)$.

Although, the ratio operation reduces the variance of the likelihood, feature level normalization and score level normalization are still effective. Standard speech feature, such as MFCC, may be distorted in channel mismatch conditions. Feature warping is to map these MFCC features to an new feature space according to a non-linear function. In this new feature space, each dimension of the feature vector will have an identical distribution, such as the standard normal distribution. T-Norm is a score normalization to reduce the dependence of the final score on different sessions. It is closely related to the cohort-base modeling. The T-Norm estimates the mean and variance of the impostor scores for each trial utterance based on a large pool of impostor speakers. Ideally, with this estimation, the T-Norm is able to normalize all the impostor scores into a standard normal distribution.

3. FISHER MAPPING AND SVM LEARNING

In a conventional GMM-UBM framework, the UBM is a background model to describe acoustic feature space of human speech. Instead of this traditional probabilistic model interpretation, several researchers [8, 9, 13] suggest a different viewpoint about the function of UBM model: it defines a mapping function from variable-length speech utterance into a fixed dimensional vector. Fisher mapping is applied in [8] to map speech utterance, while score space mapping is used in [9]. The *Fisher mapping* of a observation sequence is defined as

$$\phi(U) = \nabla_\lambda \log P(x_1^T|\lambda) \quad (4)$$

$$\nabla_{w_i} \log P(x_1^T|\lambda) = \frac{\gamma(i)}{w_i} \quad (5)$$

$$\nabla_{\mu_i} \log P(x_1^T|\lambda) = 2\gamma(i)\Sigma_i^{-1}\delta(i) \quad (6)$$

which is very similar to the utterance mapping. However, in utterance mapping, the soft count $\gamma(i)$ and adjustment $\delta(i)$ are separated, with carefully designed similarity measure, it has more flexibility than fisher mapping.

3.1. Learning with Support Vector Machine

A Support Vector Machine is a learning algorithm for pattern recognition and regression problems. SVM aims to maximize the margin between positive and negative training samples with penalty term of complex hypothesis (high VC dimension). Based on this principle, the SVM adopts a systematic approach to find a linear function that belongs to a set of functions with lowest VC dimension. With the kernel method, the SVM can also approximate a non-linear function.

Given a set of samples $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ where x_i $(x_i \in R^d)$ is the input vector of a d-dimensional space and y_i is its label $(y_i \in \{-1, 1\})$, for classification, SVM is to find the optimal hyperplane that leaves the largest possible fraction of data points of the same class on the same side while maximizing the distance of either class from the hyperplane (margin). Vapnik[14] shows that finding the optimal hyperplane is equivalent to a constrained optimization problem and can be solved using quadratic programming

techniques. The optimal hyperplane is in the form

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i k(x, x_i) + b \tag{7}$$

where $k(,)$ is a kernel function and the sign of $f(x)$ determines the label of x. Any vector x_i which corresponds to a nonzero α_i is a support vector (SV) of the optimal hyperplane. One desirable feature of SVM is that the number of support vectors is usually small, thereby producing a compact classifier.

For a linear SVM, the kernel function is just the simple dot product of vectors in the input space while the kernel function in a nonlinear SVM projects the samples to a feature space with higher (possible infinite) dimensions via a nonlinear mapping function:

$$\Phi : R^d \rightarrow R^p, p >> d \tag{8}$$

and construct a hyperplane in R^p. The motivation is that it is more likely to find a linear function, as done in linear SVM, in the high dimensional feature space. Using Mercer's theorem, the expensive calculations in projecting samples into high dimensional space can be reduced significantly by using a suitable kernel function

$$k(x, x_i) =< \Phi(x), \Phi(x_i) > \tag{9}$$

where $\Phi()$ is a nonlinear projection function. Several kernel functions, such as polynomial functions and radial basis functions, have been shown to satisfy Mercer's theorem and been used in nonlinear SVMs. In this paper, libsvm[15] is used to train a linear SVM classifier in the transformed vector space.

4. UTTERANCE MAPPING AND ITERATIVE COHORT MODELING

Instead of treating all the dimensions of the vector in a single shot, the utterance mapping in [13] firstly defines a local scoring function over each component and a combination of all local scores is used as the final similarity score. This similarity function is similar to the kernel function defined in [11] which applies a VQ model to define a kernel in the concatenated mean space. However, the similarity function in [13] also takes account of the number of observations to handle the lack of observations for some components.

Based on sufficient statistics of speech utterance, the mapping from variable-length speech utterance into a fixed dimensional vector space is defined as

$$\Phi(U) = (\gamma(i), \delta_i)_{i=1}^{M} \tag{10}$$

$$\gamma(i|x_t) = \frac{w_i P_i(x_t|\lambda)}{\sum_{j=1}^{M} w_j P_j(x_t|\lambda)} \tag{11}$$

$$\gamma(i) = \sum_{t=1}^{T} \gamma(i|x_t) \tag{12}$$

$$\bar{\mu}_i = \frac{1}{\gamma(i)} \sum_{t=1}^{T} \gamma(i|t) x_t \tag{13}$$

where $\gamma(i|x_t)$ is the posterior probability of ith component given the observation x_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 $U = (x_t)_{t=1}^{T}$. $\delta_i = \bar{\mu}_i - \mu_i$ is the adjustment of the ith component.

4.1. Iterative Cohort Modeling

The similarity in the mapped vector space is a weighted combination of thresholded local similarity of each component.

$$S(U, V) = \sum_{i=1}^{M} \alpha_i(U, V)(s_i(U, V) - \theta_i(U, V)) \tag{14}$$

$$\alpha_i(U, V) = \frac{\gamma_U(i)\gamma_V(i)}{(\gamma_U(i) + \rho)(\gamma_V(i) + \rho)} \tag{15}$$

$$s_i(U, V) = \delta_U(i)\Sigma_i^{-1}\delta_V(i) \tag{16}$$

$$\theta_i(U, V) = (\theta_i(U) + \theta_i(V))/2 \tag{17}$$

where U and V are the speech utterances. The three main parts of this similarity measure are local score $s_i(U, V)$, threshold $\theta_i(U, V)$ and weighting factor $\alpha_i(U, V)$. And ρ is a smoothing factor to reduce the dynamic range of the soft count γ_i. In this paper we set $\rho = 16$.

To estimate the threshold $\theta_i(U)$, a cohort set is selected for the speech utterance U and the average local similarity over this cohort set is used as a threshold. After estimation of threshold $\theta_i(U)$, the similarity is refined. Another cohort set is selected based on this refined similarity which leads to a new estimation of threshold $\theta_i(U)$. This iteration can go further, but the empirical finding is that the improvement becomes marginal after four iterations.

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

EER/DCF	w/o T-Norm	w T-Norm
GMM-UBM	$10.98\%/52.23(10^{-3})$	$9.21\%/34.64(10^{-3})$
Fisher-SVM	$9.14\%/38.27(10^{-3})$	$8.62\%/35.30(10^{-3})$
ICM_0	$14.61\%/65.47(10^{-3})$	$11.15\%/43.71(10^{-3})$
ICM_2	$8.28\%/36.35(10^{-3})$	$8.38\%/33.58(10^{-3})$
ICM_4	$8.07\%/34.36(10^{-3})$	$8.21\%/33.51(10^{-3})$

Table 1. System Performance Comparison on NIST02 corpus

[16]. The frontend processing is done with HTK toolkit[17] to extract MFCC+DeltaMFCC feature. The total dimension is 24. Feature warping[3] 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 of speech per speaker. These 174 speakers also serves as cohort speaker pool. For T-Norm [5], 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 GMM-UBM system with log-likelihood ratio scoring. The performance was measured with two criteria: equal error rate (EER) and minimum Detection Cost Function (DCF) [16]. The compared systems are fisher mapping followed by SVM modeling and utterance transform followed by ICM modeling.

Table 5 shows the experimental results. Compared with the GMM-UBM baseline, fisher mapping-SVM system achieves better EER on both T-Norm and without T-Norm scenarios. The improvement is 16.75% relative reduction in EER. However, the baseline does achieve better DCF at T-Norm scenario. With different iterations, the ICM-based method has different performance. The ICM at 4th iteration is the best performer in term of EER and DCF. Compared with the baseline system, the ICM at 4th iteration achieves the relative improvement of 26.5% in term of EER. Furthermore, the T-Norm procedure is least effective in ICM_4 system, indicating that the ICM-based method is a very effective score normalization scheme that can be used without T-Norm procedure. Thus, the computation of T-Norm can be saved in the ICM-based method.

6. CONCLUSION AND FUTURE DIRECTION

In this paper, we compare two discriminative learning frameworks for text-independent speaker verification. The framework based on fisher mapping and SVM learning achieves better performance in term of EER than the GMM-UBM baseline. While the framework based on utterance transform and Iterative Cohort Modeling is able to outperform the GMM-UBM system and fisher-mapping system on NIST02 task. The ICM based method achieves 26.5% relatively improvement on EER ($10.98\% \rightarrow 8.07\%$). In both fisher mapping and ICM based methods, the universal background model defines a mapping function from variable-length speech utterance to a fixed dimensional vector space. The Gaussian Mixture Model trained with conventional EM algorithm may not be the optimal background model for this purpose. In the near future, we will investigate different training methods of the background model and different structures of the background for searching a better mapping function.

7. REFERENCES

[1] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[2] 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.

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

[4] Bing Xiang, Upendra V. Chaudhari, Jiri Navratil, Ganesh N. Ramaswamy, and Ramesh A. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proceedings, ICASSP*, 2002.

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

[6] M. Schmidt and H. Gish, "Speaker identification via support vector classifiers," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1996, pp. 105–108.

[7] V. Wan and W. Campbell, "Support vector machines for speaker verification and identification," in *Proceeding of Neural Networks for Signal Processing X*, 2000.

[8] S. Fine, J. Navrátil, and R. A. Gopinath, "A hybrid gmm/svm approach to speaker identification," in *Proc. ICASSP*, 2001, pp. 417–420.

[9] 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.

[10] W. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. International Conference on Acoustics Speech and Signal Processing*, 2002, pp. 161–164.

[11] J. Louradour and K. Daoudi, "Svm speaker verification using a new sequence kernel," in *Proc. European Signal Processing Conference*, 2005.

[12] D. E. Sturim, D. A. Reynolds, E. Singer, and J. P. Campbell, "Speaker indexing in large audio databases using anchor models," in *Proceedings of ICASSP*, 2001, pp. 429–432.

[13] 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.

[14] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998, forthcoming.

[15] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001.

[16] “http://www.nist.gov/speech/tests/spk/,” .

[17] “http://htk.eng.cam.ac.uk/,” .