RC25139 (W1104-040) April 11, 2011
Computer Science

IBM Research Report

Conversational-Side-Specific Inter-Session Variability Compensation

Mohamed Kamal Omar, Jason Pelecanos
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

LIMITED DISTRIBUTION NOTICE: This report has been submitted for publication outside of IBM and will probably be copyrighted if accepted for publication. It has been issued as a Research Report for early dissemination of its contents. In view of the transfer of copyright to the outside publisher, its distribution outside of IBM prior to publication should be limited to peer communications and specific requests. After outside publication, requests should be filled only by reprints or legally obtained copies of the article (e.g., payment of royalties). Copies may be requested from IBM T. J. Watson Research Center , P. O. Box 218, Yorktown Heights, NY 10598 USA (email: reports@us.ibm.com). Some reports are available on the internet at http://domino.watson.ibm.com/library/CyberDig.nsf/home .

CONVERSATIONAL-SIDE-SPECIFIC INTER-SESSION VARIABILITY COMPENSATION

Mohamed Kamal Omar and Jason Pelecanos

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
{mkomar,jwpeleca}@us.ibm.com

ABSTRACT

Inter-session variability compensation techniques in speaker recognition systems are typically crucial for achieving a satisfactory performance. General techniques for inter-session variability compensation may not capture session and channel information specific to a given conversational side. This paper investigates three methods for estimating a conversational-side-specific projection or affine transform to compensate for session and channel effects. In the first, we estimate the projection based on an estimate of the within-class covariance matrix from the statistics of a conversational-side-specific subset of the development data. In the second, we use a discriminative objective function to estimate the projection parameters. We present an iterative algorithm similar to the expectation maximization (EM) algorithm to estimate the projection parameters which maximize this objective function. An affine transform of the observation vectors of each conversational side is estimated using maximum likelihood estimation in the third method. The maximum likelihood objective function is estimated on a selected subset of the training data. We present several experiments that show how these three techniques perform compared to our baseline system on the interview tasks of the NIST 2008 and the NIST 2010 speaker recognition evaluations. The best method of these techniques gives a performance improvement of up to 20% relative compared to the baseline system.

1. INTRODUCTION

Improved user security in speech-driven telephony applications can be achieved with automatic speaker verification systems. Degradation in the performance of these systems due to inter-session variability has been one of the main challenges to the deployment of speaker verification technologies. We investigate how integrating more information about the development and the test sets into the speaker recognition system may improve its performance and robustness.

In this work, we propose three approaches for inter-session variability compensation. In the first approach, the conversational-side-specific projection of the supervector representation of each enrollment and verification conversational side is constructed by using a within-class covariance matrix estimated using a conversational-side-specific subset of the development data. This subset is selected based on a measure of similarity between the development set and the enrollment or the verification conversational side. Subsequently, this method is called the side-specific within-class covariance projection (SWCCP) approach. The approach is motivated by the assumption that many of the sources of inter-session and intra-speaker variability have similar values across conversational sides with high dot-product scores.

In the second approach, we examine a discriminative objective function for estimating the projection parameters. Many speaker verification systems use some variations of principal component analysis, probabilistic principal component analysis, nuisance attribute projection, and linear discriminant analysis to achieve a subspace representation of the conversational sides and compensate for inter-session variability. The popularity of these techniques is attributed to the existence of efficient algorithms to implement them, such as eigen-value decomposition. Discriminative training offers an alternative which optimizes an estimate of the training data recognition error. Discriminative training has been used in SVM-based speaker recognition systems [1] and in training the UBM parameters [2] and provided noteworthy improvements compared to the MLE GMMs. In this work, we integrate information about the speakers of the development set into the objective functions used for training the projection parameters discriminatively and describe an efficient iterative algorithm to estimate the projection parameters.

In the third approach, an affine transform of the observation vectors of each enrollment and verification conversational side is estimated. The parameters of the affine transform are estimated using maximum likelihood (ML) estimation. The ML objective function is estimated for each test side on a subset of the training data selected based on a similarity measure with respect to that side.

In the next section, we describe the main architecture of the speaker verification system used in this work. In Section 3, we formulate the problem and describe our objective criteria for the three approaches. The experiments performed to evaluate the performance of the systems are described in Section 4. Finally, Section 5 contains a discussion of the results and future research.

2. THE SPEAKER VERIFICATION SYSTEM

In this work, the speaker recognition systems are based on the use of GMM supervectors. These GMM supervectors are formed from the concatenation of the MAP [3] adapted means that are normalized according to a mapping proposed in [1]. Nuisance Attribute Projection (NAP) [1] is applied to remove supervector directions that correspond to large intra-speaker variability. In all the systems reported in this work, 128 nuisance directions were removed. These nuisance directions, as per our submission in the NIST 2008 speaker recognition evaluation [4], correspond to the eigenvectors with the largest eigenvalues of the average within-class covariance matrix [5]. These supervectors are constructed as follows

$$\Phi_k = \sqrt{w_k}\mathbf{\Sigma}_k^{-\frac{1}{2}}\left(\mu_k^{adapt} - \mu_k^{ubm}\right), \quad (1)$$

$$\Phi = \left[\Phi_1^T \Phi_2^T \ldots \Phi_G^T\right]^T, \quad (2)$$

where w_k is the weight of the kth Gaussian component in the GMM, μ_k^{adapt} is the MAP adapted mean for this component, μ_k^{ubm} is the universal background model (UBM) mean for this component, and $\mathbf{\Sigma}_k$ is the diagonal covariance matrix of the kth Gaussian component in the GMM. We use the single iteration MAP adaptation presented by Reynolds [3] to generate the conversational-side-specific adapted means, $\{\mu_k^{adapt}\}$, from the UBM means, $\{\mu_k^{ubm}\}$.

Before score normalization, the output scores of the speaker verification systems can be represented by some kind of generalized inner product of two vectors representing the verification and the enrollment sides [1]. This can be described by the relation

$$s = \Phi_e^T \mathbf{K} \Phi_v, \tag{3}$$

where Φ_e is the supervector representing the enrollment side, Φ_v is the supervector representing the verification side, $\mathbf{K}$ is the NAP projection matrix, and s is the score corresponding to this pair of recordings.

For all the systems reported in this work, the UBM consists of 1024 mixture components. The UBM parameters are trained using Maximum Likelihood (ML) training [6]. Both Z-Norm and T-Norm [7] score normalization approaches were applied separately for each gender. Further details about the various systems are described in the experiments section.

3. THE THREE APPROACHES

In this section, we describe the three approaches for conversational-side-specific inter-session variability compensation presented in this paper. We start by describing the method we used for selecting a subset of the training data to estimate the objective function in each of the three techniques. This is followed by a detailed description of the three techniques.

3.1. Training data selection

In each of the three methods of the conversational-side-specific inter-session variability compensation presented here, a subset of the training data is selected to be used for estimating the objective function and the parameters of the variability compensation technique. The selection of this subset is based on how close each of its members is to the enrollment or the verification conversational side of interest. This is determined based on a similarity measure equal to the dot-product score of the supervector representation of the training and the test conversational sides. We experimented with both setting a threshold on the dot-product score to select the subset of the training data and selecting the K-nearest-neighbors as members of the subset training data. Both approaches give similar results. We report here the results based on the K-nearest-neighbors algorithm.

3.2. Side-Specific within-class covariance projection (SWCCP)

In this approach, an estimate of the within-class covariance matrix is calculated using the statistics collected from the subset of the training data selected as described before. The statistics are weighted for each training conversation side by a monotonically increasing function of the dot-product score of the training conversation side and the test conversation side. This approach is a conversational-side-specific variation of our implementation of the nuisance attribute projection (NAP) approach.The details of the steps followed to estimate the SWCCP projection for each enrollment and verification conversational side are:

- Estimate a similarity measure with each training conversational side based on their dot-product score.
- Calculate a conversational-side-specific projection using an eigen decomposition of the weighted within-class covariance matrix estimated using the data of similar training speakers.
- Calculate the final dot-product scores using the projected representation of the conversational sides.

3.3. Side-Specific discriminatively trained projection (SDTP)

Estimating the projection parameters using the average within-class covariance matrix does not directly target reducing the speaker verification errors on the training data. In this approach, we use a discriminative criterion regularized with the log-likelihood objective function. The discriminative criterion reduces the value of the imposter scores corresponding to the selected subset of the training data. The parameters of the projection are updated using an EM-like algorithm to maximize the regularized objective function

$$O = -\sum_{j=1}^{J} e^{bs_j} + \lambda L, \tag{4}$$

where L is the log likelihood of the test conversational side data estimated using the UBM, $\lambda > 0$, λ is the regularization parameter, s_j is the jth imposter score, b is the scaling parameter of the imposter scores, and J is the number of imposter scores. The scaling parameter is estimated on a held-out set to provide proper conditioning of the imposter scores. In the experiments reported here, we used a value for λ which is double the value that ensures all the imposter scores are positive in the first iteration and is kept the same for the remaining iterations. Also the imposter scores are the speaker recognition scores without NAP compensation and without ZT normalization. We investigated using the NAP-compensated and ZT-normalized scores in the objective function but we keep the discussion and the results in this work to the simpler case of scores without NAP compensation and without ZT normalization.

We use an iterative algorithm similar to the EM algorithm to estimate the projection parameters that maximize the objective function in Equation 4. It can be shown that the update equations for the projection parameters are

$$\hat{v_{ij}} = \frac{-\sum_{u=1}^{U}\left[\frac{\partial O}{\partial x_{iu}}\sum_{d\neq i} v_{dj}\phi_{du} + \phi_{iu}\sum_{l\neq i}\frac{\partial O}{\partial x_{lu}} v_{lj}\right]}{\sum_{u=1}^{U}\left[\frac{\partial O}{\partial x_{iu}}\phi_{ju} + \phi_{iu}\frac{\partial O}{\partial x_{iu}}\right]}, \tag{5}$$

where $\hat{v_{ij}}$ is the new estimate of the element of the rejected subspace matrix in the ith row and the jth column, $\frac{\partial O}{\partial x_{iu}}$ is the partial derivative of the objective function with respect to the ith dimension of the projected supervector representation of the conversational side u, v_{dj} is the element of the rejected subspace matrix in the dth row and the jth column, ϕ_{du} is the dth element of the supervector representation of the conversational side u before projection, U is the number of conversational sides in the selected subset of the training data.

Since the dot-product score is a linear function of the projected supervector representation of the conversational sides, estimating the partial derivative of the discriminative portion of the objective function is straightforward. However, estimating the partial derivative of the log likelihood regularization term with respect to the elements of the projected supervector representation involves inverting the relation in Equation 1 between the supervector representation and the UBM means and between the supervector representation and the posterior probabilities of the UBM Gaussian components given a specific observation vector. After doing that, the partial derivative of the log likelihood term with respect to the ith dimension of the projected supervector representation of the conversational side u can be written as

$$\frac{\partial L}{\partial x_{iu}} = \frac{\partial L}{\partial \mu_i}\frac{\partial \mu_i}{\partial x_{iu}} + \sum_{t=1}^{t=T} \frac{\partial L}{\partial \gamma_t^{k_i}}\frac{\partial \gamma_t^{k_i}}{\partial x_{iu}}, \qquad (6)$$

where μ_i is the ith element of the UBM mean, $\gamma_t^{k_i}$ is the posterior probability of the corresponding UBM Gaussian component given the tth observation of the conversational side u.

3.4. Side-Specific maximum likelihood affine transform

In this approach, an affine transform of the observation vectors of each test conversational side is estimated using maximum likelihood estimation. The objective function and therefore the parameters of the affine transform are estimated using a selected subset of the training data as described before. The affine transform parameters are estimated using the EM algorithm by optimizing the following auxiliary function.

$$Q(\lambda|\lambda^r) = \sum_{g=1}^{G}\sum_{t=1}^{N} \gamma_t^g \left(\log|\mathbf{\Sigma}_g| + (\mathbf{z}_t - \mu_g)^T \mathbf{\Sigma}_g (\mathbf{z}_t - \mu_g)\right), \qquad (7)$$

where $Q(\lambda|\lambda^r)$ is the auxiliary function after the rth iteration,

$$\lambda = \left[\begin{array}{ccccc} \mu_1 & \mathbf{\Sigma}_1 & \dots & \mu_G & \mathbf{\Sigma}_G \end{array}\right]$$

is the set of UBM parameters after transformation, γ_t^g is the posterior probability of the Gaussian component g at time t, μ_g is the mean vector of the UBM Gaussian component g, $\mathbf{\Sigma}_g$ is the covariance matrix of the UBM Gaussian component g, G is the total number of UBM Gaussian components, and N is the total number of observations.

The affine transform is applied to the feature vectors of the conversational side as

$$\mathbf{z} = \mathbf{D}\mathbf{y} + \mathbf{e}, \qquad (8)$$

where **y** is the observed feature vector, **z** is the transformed feature vector, **D** is an $n \times n$ matrix, **e** is an $n \times 1$ vector, n is the dimension of the feature vector.

4. EXPERIMENTS

The three previously discussed methods to compensate for the inter-session variability effects were evaluated on the interview tasks of the core condition of the NIST 2010 and the NIST 2008 Speaker Recognition Evaluations (SRE) [4] and compared to the baseline system.

The development data set for the experiments performed on the NIST 2008 evaluation consists of a combination of audio from the NIST 2004 speaker recognition database, the Switchboard II Phase III corpora, the NIST 2006 speaker recognition database, and the NIST 2008 interview development set. The collection contains 13770 conversational sides: 6038 sides of male speakers and 7732 sides of female speakers. The total number of speakers in the development data is 1769 speakers: 988 female speakers and 781 male speakers. For experiments performed on the NIST 2010 evaluation data, we added the NIST 2008 evaluation data to the development set. In both cases, the development set was used to estimate the UBM parameters, to estimate the expected within-class covariance matrix over all speakers for NAP compensation, as well as for gender-dependent ZT-norm score normalization.

4.1. Baseline system

The baseline speaker recognition system, following [8], has the frontend of the IBM large vocabulary English telephone conversations automatic speech recognition (ASR) system. The 40-dimension features for the IBM ASR system are estimated from sequences of 13-dimensional perceptual linear prediction (PLP) features by using a linear discriminant analysis (LDA) projection, and then applying a maximum likelihood linear transformation (MLLT). The acoustic model consists of 250K diagonal-covariance Gaussian components. In the context of speaker-adaptive training, vocal tract length normalization (VTLN) and feature-space maximum likelihood linear regression (FMLLR) are used. A feature-based minimum phone error (FMPE) transform is applied on top of the utterance-specific FMLLR transforms. A single pass of MLLR adaptation is also performed. The language model is a 72K-vocabulary interpolated back-off 4-gram language model. Each conversational side in both the training and the testing data is represented by a GMM mean based supervector of dimension 40960 as described in Section 2. The UBM is trained using the development set by maximum likelihood estimation. The system performance was measured at two operating points, namely in terms of the Equal-Error Rate (EER) and the minimum Detection Cost Function (DCF) as defined in the evaluation plan for both the NIST 2008 and NIST 2010 evaluations [4].

4.2. Experimental Setup

Four systems are compared in the experiments: the baseline system, the side-specific within-class covariance projection (SWCCP) system with the same structure as the baseline system, the side-specific discriminatively trained projection (SDTP) with the conversational-side-specific projection to compensate for inter-session variability is estimated using the discriminative criterion in Equation 4, and the side-specific maximum likelihood affine transformation (SMLAT) system which uses a conversational-side-specific affine transform for inter-session variability compensation as described in Section 3.

For the SWCCP algorithm, we used the K-nearest-neighbor method with K equal to 120 for subset selection. The conversational sides in the selected training data subset are then used to estimate the average within-class covariance matrix. The six eigenvectors corresponding to the six largest eigenvalues of the estimated within-class covariance matrix form the basis for the rejected subspace. The components of the testing conversational side of interest corresponding to this rejected subspace are then removed. This process is repeated for all the enrollment and the verification conversational sides in the evaluation data.

For the SDTP approach, we used the K-nearest-neighbor method with K equal to 200 for subset selection. The conversational sides in the selected training data subset are then used to estimate the objective function in Equation 4 and the update equation of the projection parameters in Equation 5. The rank of the rejected subspace matrix is set to five. The components of the testing conversational side of interest corresponding to this rejected subspace are then removed. This process is repeated for all the enrollment and the verification conversational sides in the evaluation data.

For the SMLAT approach, the conversational sides in the training data which are selected with the K-nearest-neighbor method are then used to estimate the maximum likelihood objective function and

Task	Description
Int-S	Interview speech from the same microphone in training and test.
Int-D	Interview speech from different microphones in training and test.
Int-NTel	Interview speech in training and normal vocal effort telephone speech in test.
Int-NMic	Interview speech in training and normal vocal effort telephone microphone speech in test.

Table 1. Description of the interview NIST 2010 core condition evaluation tasks reported in our experiments.

	Performance minDCF (x10^3) and EER (%) (in parentheses)			
System	Int-S	Int-D	Int-NTel	Tel-NMic
Baseline	0.39 (3.4)	0.52 (5.1)	0.38 (4.1)	0.45 (3.4)
SWCCP	0.35 (3.1)	0.47 (4.9)	0.36 (3.9)	0.40 (3.2)
SDTP	0.32 (2.9)	0.45 (4.7)	0.36 (3.8)	0.41 (3.3)
SMLAT	0.39 (3.3)	0.49 (5.0)	0.35 (3.8)	0.42 (3.3)

Table 2. The results on the NIST 2010 interview core condition tasks comparing the baseline system with systems using conversational-side-specific intersession variability compensation.

Task	Description
Int-Int-All	Interview speech in training and in test.
Int-Int-S	same-mic. interview speech in training and test.
Int-Int-D	different-mic. interview speech in training and test.

Table 3. Description of the interview NIST 2008 core condition evaluation tasks reported in our experiments.

	Performance minDCF (x10^3) and EER (%) (in parentheses)		
System	Int-Int-D	Int-Int-S	Int-Int-All
Baseline	0.194 (4.2)	0.029 (0.91)	0.193 (4.1)
SWCCP	0.162 (3.2)	0.019 (0.64)	0.164 (3.2)
SDTP	0.159 (3.1)	0.020 (0.73)	0.163 (3.2)
SMLAT	0.176 (3.7)	0.021 (0.77)	0.174 (3.7)

Table 4. The results on the NIST 2008 interview core condition tasks comparing the baseline system with systems using conversational-side-specific intersession variability compensation.

the update equations of the parameters of the transform. The affine transform which is estimated by maximizing the ML objective function is then applied to each feature vector of the test side before estimating the UBM-based supervector representation of the test conversational side. This process is then repeated for all the enrollment and the verification conversational sides in the evaluation data.

The results are first reported on the interview tasks of the core condition of the NIST 2010 speaker recognition evaluation. The description of these tasks is provided in Table 1. In this first set of experiments, the NIST 2008 evaluation data is added to the original set of development data. This combined development set is used to build the UBM and estimate the NAP projection of the baseline system. As shown in Table 2, the performance of the two systems which uses the subspace removal approach outperforms both the baseline system and the system which uses an affine transform on all the tasks. The results in Table 2 show also that the SDTP system outperforms the SWCCP system on the Int-S and the Int-D tasks but not on the other tasks. The difference between the results of the two approaches is mostly insignificant.

The results are then reported on the interview tasks of the core condition of the NIST 2008 speaker recognition evaluation. The description of these tasks is provided in Table 3. As shown in Table 4, the performance of the two systems which use the subspace removal approach outperform both the baseline system and the SMLAT system. However, the results in Table 4 show also that the SMLAT system performs slightly better compared to the results in Table 2.

5. CONCLUSIONS

The SWCCP and the SDTP systems consistently outperforms the baseline system on the NIST 2010 and NIST 2008 interview core condition tasks. Both systems also consistently outperforms the SMLAT system on these evaluation tasks. In both cases, the improvement is achieved by removing a conversational-side-specific subspace of the supervector representation of the test conversational sides. Integrating training speaker information into the estimation of the parameters of the rejected subspace in both cases is thought to account for this improvement. In the SMLAT approach, estimating the parameters of the affine transform using maximum likelihood estimation did not provide significant improvement compared to the baseline on most of the NIST 2010 and the NIST 2008 evaluation tasks.

6. REFERENCES

[1] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[2] M. Omar and J. Pelecanos, “Training universal background models for speaker recognition,” *A Speaker Odyssey, The Speaker Recognition Workshop*, pp. 52–57, 2010.

[3] D. Reynolds, T. Quatieri, and R. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 19–41, 2000.

[4] National Institute of Standards and Technology, “NIST speech group website,” *http://www.nist.gov/speech*.

[5] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-class covariance normalization for SVM-based speaker recognition,” *International Conference on Spoken Language Processing*, 2006.

[6] D. Reynolds and R. Rose, “Robust text-independent speaker identification using Gaussian mixture models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[7] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Processing*, vol. 10, no. 1/2/3, pp. 42–54, 2000.

[8] M. Omar and J. Pelecanos, “A novel approach to detecting non-native speakers and their native language,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.