INTERSESSION VARIABILITY COMPENSATION FOR LANGUAGE DETECTION

Xi Zhou[1]*, Jiří Navrátil*[2]*, Jason W. Pelecanos*[2]*, Ganesh N. Ramaswamy*[2]*, and Thomas S. Huang*[1]

[1]Dept. of ECE, University of Illinois at Urbana-Champaign (UIUC), Urbana, IL 61801, USA
[2]Statistical Content Analytics Group, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

ABSTRACT

Gaussian mixture models (GMM) have become one of the standard acoustic approaches for Language Detection. These models are typically incorporated to produce a log-likelihood ratio (LLR) verification statistic. In this framework, the intersession variability within each language becomes an adverse factor degrading the accuracy. To address this problem, we formulate the LLR as a function of the GMM parameters concatenated into normalized mean supervectors, and estimate the distribution of each language in this (high dimensional) supervector space. The goal is to de-emphasize the directions with the largest intersession variability. We compare this method with two other popular intersession variability compensation methods known as Nuisance Attribute Projection (NAP) and Within-Class Covariance Normalization (WCCN). Experiments on the NIST LRE 2003 and NIST LRE 2005 speech corpora show that the presented technique reduces the error by 50% relative to the baseline, and performs competitively with the NAP and WCCN approaches. Fusion results with a phonotactic component are also presented.

Index Terms— WCCN-LLR, NAP, ISV

1. INTRODUCTION

Spectral (a.k.a. acoustic) modeling belongs to one of the successful approaches applied in automatic language recognition [1]. Recently, Shifted-Delta Cepstra Features in conjunction with Gaussian Mixture Models (GMM) [2] were demonstrated to be highly effective both as individual components as well as in fusion with other modeling approaches [3].

A major error source in GMM-based acoustic modeling for language detection is Inter-Session Variability (ISV) which is also a significant challenge in many other pattern recognition tasks (the term "session" refers to a particular speech recording). A number of techniques have been proposed to solve the problem, including feature warping [4], and feature mapping [5]; as well as score compensation techniques such as HNorm [6] and TNorm [7]. In [8], Kenny proposed using factor analysis to compensate for speaker and channel variability in GMM-based speaker verification. In [9], Hatch introduced Within-Class Covariance Normalization (WCCN) to modify a generalized linear kernel for Support Vector Machine (SVM) based speaker verification. Nuisance Attribute Projection (NAP) [10] is another successful approach to mitigate the cause of variability in the SVM feature space by removing certain subspace components. Noor and Aronowitz [11] defined a session-space and modeled the intra-speaker subspace explicitly for language identification. In [12], Castaldo proposed to alleviate ISV by compensating the observation features for GMM based language identification.

The GMM formulation which simply sums up the likelihood of Gaussian components has a number of underlying assumptions. One assumption is that the feature vectors from a single session from the language being tested are independent and identically distributed to their corresponding language GMM. In one formulation, each language GMM may be trained on a pool of utterances. Here session variability and the pooling of utterances result in significant mismatch and the loss of distribution sharpness which violates the identically distributed assumption. Session variability compensation may help to relax this assumption.

In this paper, we adopt the concept of WCCN/NAP and introduce a derived algorithm to calculate scores, termed WCCN-LLR, in order to compensate for ISV in GMM-based language detection. We formulate the LLR as a function of the GMM concatenated mean supervectors, estimate the distribution of each language in the high dimensional supervector space and subsequently de-emphasize the directions with the largest intersession variability. Experiments obtained on the NIST LRE 2003 and 2005 databases demonstrate the effectiveness of the new technique.

2. NAP AND WCCN

Nuisance Attribute Projection (NAP), as introduced by Solomonoff et al [10], is one of the successful ISV compensation approaches for SVM-based speaker recognition. Firstly, the speaker models are constructed via *maximum-a-posteriori* (MAP) adaptation of the means of the UBM. Using an adapted model, a GMM supervector is constructed by concatenating the means of the adapted mixture components. The supervector can be thought of as mapping a variable length utterance to a fixed-dimensional point in a high-dimensional (supervector) space. A linear kernel based on the supervectors is then used in an SVM classifier [13].

$$K(m^a, m^b) = \sum_{j=1}^{M} \left(\sqrt{\lambda_j}\Sigma_j^{-\frac{1}{2}} m_j^a\right)^T \left(\sqrt{\lambda_j}\Sigma_j^{-\frac{1}{2}} m_j^b\right) \quad (1)$$

where λ_j and Σ_j are the weight and covariance for the j_{th} Gaussian component of an M-component GMM; while m_j^a and m_j^b denote the adapted means for utterance a and b respectively.

The NAP approach constructs a modified kernel matrix that removes subspace components that cause ISV:

$$\begin{aligned} K(m^a, m^b) &= [Pe(m^a)]^t [Pe(m^b)] \\ &= e(m^a)^t Pe(m^b) \end{aligned} \quad (2)$$

where $P = I - vv^t$ is a projection matrix, v is a unit length vector indicating the direction being removed from the SVM expansion space, and $e()$ is the utterance to SVM feature space expansion.

Several criteria have been proposed for estimating P (and correspondingly v) and one possibility is:

$$v^* = \arg\min \sum_{i,j} W_{i,j} \|Pe(m^i) - Pe(m^j)\|_2^2 \quad (3)$$

where $W_{i,j}$ represents the weight for utterance comparison i and j, and can be selected in several different ways [13]. Another version of the NAP approach (as used in these experiments) is to select the subspace to be removed based on the within-class covariance information.

Hatch et al [9] proposed the WCCN approach which achieved significant improvements on SVM based speaker recognition. It considered P in Equation 2 as the inverse of a modified estimate of the within-class covariance matrix of the supervectors. The difference between WCCN and NAP is mainly how to weigh different directions. WCCN introduced a more general framework to split the feature space into two subspaces, and weighted them respectively. NAP can be interpreted as a simplified WCCN where the weights are zero for the directions in the nuisance subspace and one for the directions in the remaining subspace. A detailed explanation of WCCN can be seen in [14].

3. WCCN-LLR

3.1. Supervector LLR formulation

To present the supervector LLR formulation, the probability density of a feature vector x given a GMM representation, Θ, is shown:

$$g(x;\Theta) = \sum_{i=1}^{M} \omega_i \mathcal{N}(x;\mu_i,\Sigma_i) \tag{4}$$

where ω_i, μ_i and Σ_i are the weight, mean and covariance for the i_{th} Gaussian component respectively.

The Log-Likelihood Ratio (LLR), calculated over an entire utterance, $X = \{x_1, x_2, \ldots, x_T\}$, is one of the de-facto standard verification measures in state-of-the-art systems.

$$\begin{aligned} LLR(X) &= \sum_{t=1}^{T} \log \frac{g(x;\Theta^s)}{g(x;\Theta^u)} \\ &= \sum_{t=1}^{T} \log \frac{\sum_{j=1}^{M} \omega_j \mathcal{N}(x_t;\mu_j^s,\Sigma_j)}{\sum_{j=1}^{M} \omega_j \mathcal{N}(x_t;\mu_j^u,\Sigma_j)} \end{aligned} \tag{5}$$

where superscript s and u denote a specific language model and UBM respectively. Moreover, only the mixture component means are adapted from the UBM, while the weights and diagonal covariances of the language model are kept the same as in the corresponding UBM.

Let n_{kt} denote the posterior probability of Gaussian k given an observed x_t for the UBM (i.e., $n_{kt} = \frac{\omega_k \mathcal{N}(x_t;\mu_k^u,\Sigma_k)}{\sum_{j=1}^{M} \omega_j \mathcal{N}(x_t;\mu_j^u,\Sigma_j)}$). We consider a lower bound on the LLR, $L(X)$, as follows:

$$\begin{aligned} LLR(X) \geq L(X) &= \sum_{t=1}^{T}\sum_{j=1}^{M} n_{jt} \log \frac{\omega_j \mathcal{N}(x_t;\mu_j^s,\Sigma_j)}{\omega_j \mathcal{N}(x_t;\mu_j^u,\Sigma_j)} \\ &= -\frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{M} n_{jt} \Big\{ (x_t-\mu_j^s)^T \Sigma_j^{-1} (x_t-\mu_j^s) \\ &\qquad -(x_t-\mu_j^u)^T \Sigma_j^{-1} (x_t-\mu_j^u) \Big\} \\ &= \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{M} n_{jt} \left(2x_t - \mu_j^s - \mu_j^u\right)^T \Sigma_j^{-1} \left(\mu_j^s - \mu_j^u\right) \\ &= \frac{1}{2}\sum_{k=1}^{M} n_k \left(2\bar{x}_k - \mu_k^s - \mu_k^u\right)^T \Sigma_k^{-1} \left(\mu_k^s - \mu_k^u\right) \end{aligned} \tag{6}$$

with

$$n_k = \sum_t n_{kt} \tag{7}$$

$$\bar{x}_k = \frac{1}{n_k}\sum_t n_{kt} x_t \tag{8}$$

Let us make the following assignments. The $vec()$ operator is used here to specify the column-wise concatenation of a matrix into a single column vector form.

$$m_x = vec\left(\bar{x}_1 \quad \bar{x}_2 \quad \ldots \quad \bar{x}_M\right) \tag{9}$$

$$\mu^s = vec\left(\mu_1^s \quad \mu_2^s \quad \ldots \quad \mu_M^s\right) \tag{10}$$

$$\mu^u = vec\left(\mu_1^u \quad \mu_2^u \quad \ldots \quad \mu_M^u\right) \tag{11}$$

$$\Sigma = \begin{pmatrix} \frac{1}{n_1}\Sigma_1 & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2}\Sigma_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{n_M}\Sigma_M \end{pmatrix} \tag{12}$$

This gives the following representation for the LLR bound:

$$\begin{aligned} L(X) &= (2m_x - \mu^s - \mu^u)^T \Sigma^{-1} (\mu^s - \mu^u) \\ &= \left(\Sigma^{-\frac{1}{2}} \left(2(m_x - \mu^u) - (\mu^s - \mu^u)\right)\right)^T \\ &\quad \left(\Sigma^{-\frac{1}{2}} (\mu^s - \mu^u)\right) \end{aligned} \tag{13}$$

For simplicity, let us denote $m = \Sigma^{-\frac{1}{2}}(m^x - \mu^u)$ and $\mu = \Sigma^{-\frac{1}{2}}(\mu^s - \mu^u)$, then

$$L(X) = \frac{1}{2}(2m-\mu)^T \mu \tag{14}$$

where m is the vector containing the normalized sufficient statistics of an utterance; which means each utterance becomes a vector in this high-dimensional space. Hence, the LLR approximation is a simple dot product in the high-dimensional space.

3.2. WCCN and LLR Combination

For estimating the correlation between the parameters across different components, let us consider the detection task as a two class (target and non-target) classification problem in the high-dimensional super-vector space. Suppose the distribution for each language (including UBM) in the high dimensional space is Gaussian and with shared covariance matrix S. Then the LLR for the vector m is

$$\begin{aligned} L(m) &= \log \frac{g(m;\Theta^s)}{g(m;\Theta^u)} = \log \frac{\mathcal{N}(m;\nu^s,S)}{\mathcal{N}(m;\nu^u,S)} \\ &= \frac{1}{2}(2m - \nu^s - \nu^u)^T S^{-1} (\nu^s - \nu^u) \\ &= \frac{1}{2}(2m-\mu)^T S^{-1} \mu \end{aligned} \tag{15}$$

Notice that $\nu^s - \nu^u = \mu$ according to the normalization of μ. By comparing Equations 14 with 15, the only difference between the two measures is the precision matrix S^{-1}. Via the covariance matrix S in Equation 15 correlations between the parameters across different components are incorporated into the calculation. However, the

size of S in the formulation is too large to be robustly estimated. Taking the settings used in these experiments in which the feature dimension is 56 and the number of Gaussian components is 1024, the resulting supervector dimension would be 57,344, with the number of free parameters in S being of the order of 10^9.This is clearly an impractical size to provide a reliable estimate given the relatively small number of utterances to train with. Instead of using Equation 15 directly, we exploit part of the information in this matrix in two steps as follows.

Firstly, we assume that a diagonal covariance matrix can be estimated relatively robustly (compared to estimating the parameters of a full covariance matrix). The S matrix may be written in the form of a diagonal covariance matrix D and a correlation matrix R.

$$S = D^{\frac{1}{2}} R D^{\frac{1}{2}} \tag{16}$$

Secondly, we de-emphasize the directions with largest variance by applying a mixed eigenvector and unit-residual representation for R^{-1}. This follows from the WCCN work proposed by Hatch [9]. The representations for $R^{-\frac{1}{2}}$ and R^{-1} are given accordingly:

$$\text{Let} \quad R^{-\frac{1}{2}} = K\Lambda^{-\frac{1}{2}}K^T + (I - KK^T) \tag{17}$$

$$\text{then} \quad R^{-1} = K\Lambda^{-1}K^T + (I - KK^T) \tag{18}$$

where Λ is a diagonal matrix containing the most significant eigenvalues and K represents the corresponding top few eigenvectors of the (within-class variance-normalized) development data. Equations 17 and 18 each have two terms; the first relates to the space described by the eigenvectors that are to be de-emphasized; the second term relates to the residual subspace that remains untransformed. The modified score function is now:

$$\begin{aligned} L(X) &= (2\mathrm{m} - \mu)^T S^{-1}(\mu) \\ &= \left(D^{-\frac{1}{2}}(2m - \mu)\right)^T R^{-1} \left(D^{-\frac{1}{2}}\mu\right) \\ &= \left(R^{-\frac{1}{2}} D^{-\frac{1}{2}}(2m - \mu)\right)^T \left(R^{-\frac{1}{2}} D^{-\frac{1}{2}}\mu\right) \end{aligned} \tag{19}$$

This normalization is the same as WCCN with this variant being constructed for a log-likelihood ratio function instead. Later on, we refer to this technique as WCCN-LLR. Note that with this form of subspace re-weighting, there is no guarantee that the function can be represented as an equivalent log-likelihood ratio.

4. EXPERIMENTS

4.1. Speech corpus and baseline system

The development data set used in these evaluations is the CallFriend corpus, which consists of 1800 utterances from 12 languages. In these experiments, the system is trained on the first 5 minutes of each utterance.

The described methods are evaluated on the 2003 and 2005 NIST Language Recognition Evaluation (LRE) test sets. The baseline used in this study is a GMM/UBM system with mean-only MAP adaptation. Each language model has 1024 Gaussian components. The Shifted Delta Cepstral (SDC) features configured as 7-1-3-7 plus 7 static cepstral features are used, as described in [12].

4.2. Results

Table 1 shows Equal Error Rates (EER) corresponding to different ISV compensation methods across evaluation sets and duration conditions. The baseline represents the performance of a traditional GMM/UBM-MAP system; the NAP-Kernel uses the NAP and a kernel-based score; the WCCN-Kernel uses WCCN for ISV and a kernel-based score; the WCCN-LLR (red curve) uses WCCN with LLR-based scoring. Notice that WCCN-LLR achieves the best performance among ISV compensation methods on each data set and is the only technique outperforming the baseline for the very short (3 sec) tests.

Figures 1 and 2 show the DET curves obtained from the NIST LRE 2003 and 2005 30 second test sets. The proposed WCCN-LLR ISV compensation method achieved more than 50% relative error reduction compared to the baseline system. Furthermore, results on both data sets provide evidence of competitive performance of the WCCN-LLR configuration with the WCCN and NAP kernels. Based on these results, the effectiveness of the WCCN-LLR method may be attributed to at least two potential sources: 1) the use of a soft-weighted (WCCN-like) subspace removal and 2) the use of a LLR-

Table 1. Equal error rates across conditions

EER(%)	lid05			lid03		
	30s	10s	3s	30s	10s	3s
Baseline	17.6	21.0	26.9	12.3	16.5	23.6
NAP-Kernel	13.6	22.3	33.6	7.4	16.4	28.4
WCCN-Kernel	11.5	19.5	30.6	5.7	14.4	26.1
WCCN-LLR	8.7	15.6	25.3	5.4	12.0	22.4

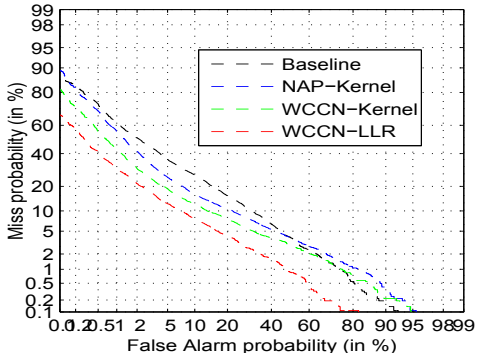

Fig. 1. DET for NIST LRE-05 30 sec

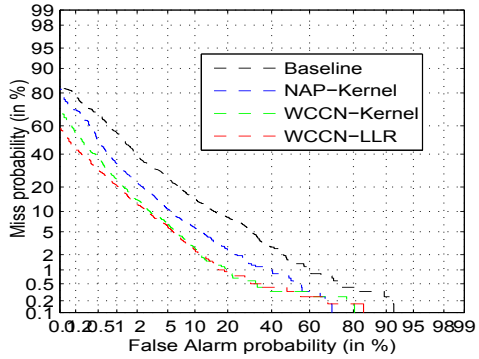

Fig. 2. DET for NIST LRE-03 30 sec

Table 2. Fusion results

EER(%)	lid03-30	lid05-30
Phonotactic	6.2	9.8
Acoustic	5.4	8.7
Fusion	3.4	7.0

related scoring metric, as opposed to a kernel-based metric.

Figure 3 shows the EER corresponding to compensating for a different number of subspace dimensions for the NIST LRE 2005 - 30 second test set. The results indicate that WCCN-LLR performs better than NAP-Kernel over the range of dimensions evaluated. The EER curve for WCCN-LLR also seems smoother in contrast to NAP. Another point is the considerable performance gain as a result of using just the first few dimensions across these methods. The EER for the baseline on the 30-sec test is 19.9% and 17.6% without and with T-Norm, respectively. For the WCCN-LLR case, when the subspace dimension is 12, the EER is 12.5% and 9.2%, without and with T-Norm, respectively. The latter result is already close to the best performance indicating that most of the impact for this type of intersession variability model resides in the first few dimensions.

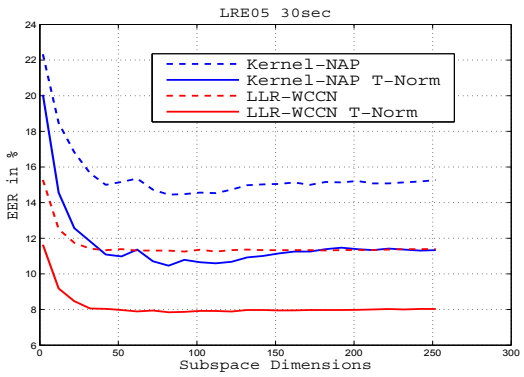

Fig. 3. EER v.s. the number of subspace dimensions considered for NIST LRE-05 30 second task.

4.3. Fusion Results

To demonstrate the effectiveness of the described acoustic technique in an overall system, we also present fusion results with phonotactics. For this purpose, a phonotactic language detector was used as described in [15] and the fusion was performed as an equal-weight linear combination of the acoustic and phonotactic scores.

The EERs for the acoustic and the phonotactic systems, and their fusion are shown in Table 2. Notably, the performance of the acoustic system with ISV compensation, and the phonotactic system in isolation are comparable. Their combination leads to a further significant improvement, thus confirming previous observations [3].

5. CONCLUSION

In this work, we applied an intersession variability compensation technique for GMM based language detection. We adopted the ideas of NAP and WCCN, which was introduced for SVM-based speaker verification, and extended this to be tested using a modified LLR-based scoring function. Experiments show that the new ISV compensation methods achieved improvements over the baseline and performed competitively with the WCCN and NAP approaches.

6. REFERENCES

[1] J. Navratil, *Multilingual Speech Processing*, chapter Automatic Language Identification, Academic Press, 2006, Eds. T. Schultz & K. Kirchhoff, ISBN-978-0120885015.

[2] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller Jr., "Language identification using Gaussian mixture model tokenization," in *ICASSP02*, 2002, pp. I: 753–756.

[3] W. Campbell, T. Gleason, J. Navrátil, D. Reynolds, W. Shen, E. Singer, and P. Torres-Carrasquillo, "Advanced language recognition using cepstra and phonotactics: MITLL system performance on the NIST 2005 language recognition evaluation," in *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006.

[4] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *ODYSSEY01*, 2001, pp. 213–218.

[5] D.A. Reynolds, "Channel robust speaker verification via feature mapping," in *ICASSP-2003*, 2003, pp. II: 53–56.

[6] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker 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, pp. 42–54, 2000.

[8] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Odyssey-04*, 2004, pp. I: 219–226.

[9] A. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: Application to speaker recognition," in *ICASSP06*, 2006, pp. V: 585–588.

[10] A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for SVM speaker recognition," in *Odyssey-04*, 2004, pp. I: 57–623.

[11] E. Noor and H. Aronowitz, "Efficient language identification using anchor models and support vector machines," in *Odyssey-2006*, 2006.

[12] F. Castaldo, E. Dalmasso, P. Laface, D. Colibro, C. Vair, and Politecnico di Torino, "Language identification using acoustic models and speaker compensated cepstral-time matrices," in *ICASSP07*, 2007, pp. IV: 1013–1016.

[13] W. M. Campbell, D. E. Sturim, D.R Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *ICASSP06*, 2006, pp. I: 97–100.

[14] A. O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *ICSLP06*, 2006, pp. 1471–1474.

[15] J. Navratil, "Recent advances in phonotactic language recognition using binary-decision tree," in *INTERSPEECH06*, 2006, pp. 1338–Mon2CaP.6.