

MAXIMUM CONDITIONAL MUTUAL INFORMATION MODELING FOR SPEAKER VERIFICATION

Mohamed Kamal Omar, Jiri Navrátil, Ganesh Ramsawamy

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA

mkomar, jiri, ganeshr@us.ibm.com

ABSTRACT

This paper describes a novel approach for class-dependent modeling and its application to automatic text-independent speaker verification. This approach maximizes the conditional mutual information between the model scores and the class identity given some constraints on the scores. It is shown in the paper that maximizing the differential entropy of the scores generated by the classifier or the detector is an equivalent criterion. This approach allows emphasizing different features, in the feature vector used in the detection, for different target speakers. In this paper, we apply this approach to the NIST 2003 1-speaker verification task. Compared to the baseline system, around 10% relative improvement in the minimum detection cost function (DCF) is obtained.

1. INTRODUCTION

Maintaining data security and authenticity in speech-driven telephony applications can be done effectively through speaker verification. Current automatic speaker verification systems face significant challenges caused by adverse acoustic conditions. Telephone band limitation, channel/transducer variability, as well as the natural speech variability all have a negative impact on the performance. Degradation in the performance of speaker verification and recognition systems due to channel mismatch has been one of the main challenges to actual deployment of speaker verification and recognition technologies.

Conventional systems for speaker verification use features extracted from very short segments of speech, and model these features using Gaussian mixture models (GMM). In [1], a particularly effective way of speaker modeling via Bayesian adaptation from speaker-independent models combined with a likelihood-ratio detector was introduced, improving robustness of the verification system trained with limited training data. However, this approach still suffers significant performance degradation in the presence of handset variability. Score normalization techniques like H-norm and T-norm [2] and feature mapping approaches try to compensate for these channel mismatch effects. These approaches try to compensate for channel effects and speech variability in the training and testing utterances. However, these approaches do not take into consideration that different speakers may need different models or different emphasize on specific features in the feature vector.

In this paper, we describe a discriminative approach to improve the performance of speaker verification systems. The previous discriminative approaches for improving automatic speaker verification can be classified into two main categories: model-

based and feature-based. Model-Based approaches include both using a discriminative classifier like support vector machines and training the parameters of a generative model like GMM to optimize a discriminative criterion. Examples of using discriminative classifiers like SVM in the automatic speaker verification task are [3], [4]. In [5], minimum classification error (MCE) criterion was used to estimate the parameters of a hidden Markov model (HMM) system for automatic speaker identification. In [6], the error probability was directly approximated using sets of target and imposter sets and the generative model parameters were trained to minimize an estimate of the verification error. In feature-based approaches, a transform of the features is applied to improve the ability of the features to discriminate among various speakers. Examples of these transforms include linear discriminant analysis (LDA) [7] and heteroscedastic linear discriminant analysis (HLDA) [8]. In [9], a discriminative feature design using neural networks was also proposed.

In this paper, we present a novel score-based approach for improving the performance of automatic speaker verification systems. It maximizes the conditional mutual information of the log likelihood ratio scores and the speaker identity given some constraints on the scores. We show that maximizing this objective function is equivalent to maximizing the differential entropy of the log likelihood ratio scores of the utterance given the speaker. We estimate a weighting to the log likelihood ratio scores in each dimension of the feature vector to optimize our objective criterion. It is interesting to note that, as these weights are speaker-dependent, our approach improves the performance of the system without changing either the features or the classifier, but only with creating better scores for each speaker.

In the next section, we will formulate the problem and describe our objective criterion. In section 3, the algorithm used in estimating the weighting coefficients to optimize our objective criterion is described. The experiments performed to evaluate the performance of our approach are described in section 4. Finally, Section 5 contains a discussion of the results and future research.

2. PROBLEM FORMULATION

In this section, we will discuss the relation between the performance of our automatic speaker verification system and the mutual information of the log likelihood ratio scores and the speaker identity and then show how the problem can be reduced to a maximum differential entropy problem.

The mutual information of the log likelihood ratio scores and the speaker identity is

$$I(S, P) = H(P) - H(P|S), \quad (1)$$

where S is the log likelihood ratio score of the utterance, P is the identity of the speaker, $H(P)$ is the entropy of the speakers, and $H(P|S)$ is the conditional entropy of the speakers given the log likelihood ratio score. From this relation, and knowing that $H(P|S) \geq 0$, we get

$$I(S, P) \leq H(P), \quad (2)$$

with equality if and only if $H(P|S) = 0$. This can take place only if we have a zero verification error and there exist a one-to-one mapping from the log likelihood ratio scores to the speaker identity. So maximizing the mutual information is equivalent to minimizing the verification error on the set of given speakers.

Rewriting the mutual information of the log likelihood ratio of a given utterance given the speaker and the speaker identity as

$$I(S, P) = H(S) - H(S|P), \quad (3)$$

where $H(S)$ is the differential entropy of the log likelihood ratio scores and $H(S|P)$ is the conditional differential entropy of the log likelihood ratio scores given the speaker identity.

Due to having a small amount of training data per speaker, we can not have a reliable estimate of the conditional differential entropy of the log likelihood ratio scores given the speaker identity. Therefore, we choose to maximize the conditional mutual information of the log likelihood ratio of a given utterance given the speaker and the speaker identity given that $H(S|P)$ is constant. This is equivalent to maximizing the differential entropy of the scores $H(S)$. It can be shown that the assumption that $H(S|P)$ is constant is valid in the case of using T-norm with sufficiently large number of speakers.

Since the performance of the automatic speaker verification system is insensitive to any scaling of the values of the log likelihood ratio scores, we choose to put a constraint of constant variance on these scores. Given this constraint, the differential entropy of these scores is maximized if and only if the probability density function (PDF) of the scores are Gaussian. Hence, maximizing the differential entropy of the scores becomes a maximum likelihood problem. In which, we maximize the likelihood that the log likelihood ratio scores are Gaussian random variables.

Therefore our objective function to be maximized is

$$L = -\frac{NM}{2} \log \sigma^2 - \sum_{i=1}^M \sum_{k=1}^N \frac{1}{2} \frac{(s_k^i - \mu)^2}{\sigma^2}, \quad (4)$$

where N is the number of training utterances, M is the number of training speakers including the current target speaker, μ is the mean of the scores, and σ^2 is the variance of the scores, s_k^i is the log likelihood ratio score using the i th speaker model on the k th utterance.

The log likelihood scores are the summation of the log likelihood scores due to different elements in the feature vector. So we introduce a weighting vector for each speaker to weight the scores corresponding to different elements of the feature vector. This can be written as

$$s^i = \sum_{j=1}^J w_j^i s_j^i, \quad (5)$$

where s^i is the log likelihood score for speaker i , w_j^i is the weight for speaker i assigned to the score from the j th feature, s_j^i is the log likelihood score for speaker i corresponding to the j th element in the feature vector, and J is the length of the feature vector.

3. IMPLEMENTATION

In this section, we present our implementation of the approach described in the previous section. Our goal is to calculate the speaker-dependent weights of the log likelihood scores, $\{w_j^t\}_{t=1, j=1}^{i=T, j=J}$, where T is the number of target speakers, which maximize our objective function in Equation 4, i.e.

$$\hat{W}^t = \arg \max_{W^t} - \sum_{i=1}^M \sum_{k=1}^N \frac{(s_k^i - \mu)^2}{\sigma^2}, \quad (6)$$

for $t = 1, 2, \dots, T$, $W^t = \{w_1^t, \dots, w_J^t\}$, J is the dimension of the feature vector, and M is the number of the training speakers including the current target speaker.

To calculate these weights, we use the gradient decent algorithm to estimate the weights that will maximize our objective function. The gradient of the objective function with respect to the speaker-dependent weight vectors is

$$\frac{\partial L}{\partial W^t} = - \sum_{k=1}^N \frac{(s_k^t - \mu)}{\sigma^2} S_k^t, \quad (7)$$

for $t = 1, 2, \dots, T$, where $S_k^t = \{s_{k1}^t, \dots, s_{kr}^t, \dots, s_{kT}^t\}$, s_{kT}^t is the log likelihood score for the k th utterance corresponding to the r th feature and the t th speaker, W^t is the weighting vector for target speaker t .

Using the gradient decent approach, the value of the weights are updated in each iteration using the relation

$$W_{n+1}^t = W_n^t + \alpha \frac{\partial L}{\partial W^t} |_{W^t=W_n^t}, \quad (8)$$

where W_{n+1}^t is the weight vector at iteration $n + 1$, W_n^t is the weight vector at iteration n , and α is a step size that should be chosen small enough to guarantee convergence and large enough to reduce the number of iterations required to achieve convergence.

Our speaker verification system is based on the Gaussian mixture model structure and the widely successful approach of adapted target speaker models [1]. In order to alleviate mismatch problems due to channel variability, we apply several steps of compensation and normalization during feature extraction and score calculation based on the Gaussianization technique [10] [11] and the T-Norm [2]. A coupled-Gaussian scoring technique [12] combined with a discriminative feature-space transform [13] help optimize our system for the NIST-defined operating region of low false alarms. Furthermore a cellular codec simulation software was utilized to adapt landline-telephone data to the cellular task thus helping the overall data augmentation and a better testing-condition match. Extracting features from the speech signal consists of generating the classic MFCC frame sequence comprising 18 static plus 18 derivative dimensions followed by either a marginal or a full Gaussianization transform as described in [10]. The Gaussianization step seeks to achieve normal distribution of the features within

any window of a predetermined length - in our case three seconds. This leads to a redundancy reduction due to channel effect suppression alongside a partial information loss due to removal of long-term speaker properties, further extending the process of feature standardization. Individual speaker models are estimated as Maximum-A-Posteriori (MAP) adapted models using a speaker-independent Gaussian Mixture Model (GMM) trained in a relatively extensive fashion covering a variety of acoustic conditions, such as landline and cellular channels, male and female speakers, etc. In our system, the training of such a speaker-independent model, also known as Universal Background Model (UBM) [14], consists of the following steps:

1. Pre-Cluster the training data set via a fast deterministic top-down algorithm using binary data splits with an eigenvector-based criterion, as described in [12]
2. Iterate via K-Means using the Euclidean distortion measure until convergence
3. Carry out one iteration of the Expectation-Maximization algorithm on the mean and covariance parameters calculated from final K-Means clusters. Only the diagonals of the covariance matrices are computed
4. Estimate a feature-space Maximum-Likelihood Linear Transformation (MLLT) using the EM parameters from the previous step. The MLLT achieves optimum feature space for diagonal covariance modeling [15]
5. Repeat Step 3 and 4 iteratively until convergence, or until a maximum number of iterations is reached.

Using the speaker training data, the MAP adaptation is applied on the mean parameters of the UBM in the MLLT-transformed space to create each individual speaker model characterized by a set of adapted mean vectors with the diagonal covariances and the Gaussian weights being shared with the UBM.

To process a particular trial, i.e. a test utterance against a claimed target identity, the system calculates the (component-wise) Gaussian likelihoods of the vectors in the sequence using the gender-dependent UBMs. As described in [13], for each feature vector, only the maximum-likelihood Gaussian component is taken into account in each system, followed by a likelihood ratio of that component and its corresponding adapted counterpart in the target speaker GMM. Such tying of single components was shown experimentally to cause a counterclockwise rotation of the DET curve, which is favorable to the operating region of interest. The frame-wise likelihood ratios are then averaged over the complete utterance and output as real-numbered scores.

Beyond speaker properties, the resulting likelihood ratio values are typically influenced by irrelevant information such as channel properties and the generally non-linearly shaped acoustic space. The T-Norm [2] is applied to the ratio scores in order to compensate for certain shifts and scales due to the acoustic space. The T-Norm involves a set of additional speaker models created consistently with the regular target models, which serve as a basis for calculating the first and second order statistics of the score distribution at the given test point in the acoustic space. The trial score is normalized by subtracting the mean and dividing by the standard deviation of the T-Norm model scores. Due to the fact that the T-Norm takes the acoustic conditions of the given test into account, it is effective in suppressing score distribution shifts and scaling due to channel variations. In the final evaluation, we used a total number of 234 T-Norm speakers, whereby a gender-matched

System	min. DCF	EER
Baseline	0.0350062	8.5194%
MCMI	0.032044	8.25227%

Table 1. Comparison of the Baseline and MCMI systems

subset was used for each given trial. Furthermore a weighting and pruning scheme of the individual T-Norm scores is applied as described in [16].

4. EXPERIMENTS

The performance of the described method was evaluated using data from the cellular part of the Switchboard (SWB) telephone corpus, as defined by NIST for the 1-speaker cellular detection task in the 2003 Speaker Recognition Evaluations (SRE) [17]. The 2003 set consists of 356 speakers, and a total of 37664 verification trials.

The 2001 cellular SRE, the 1996 SRE landline-quality dataset and an internal cellular-quality data collection served as the data for the estimation of the substrate model (UBM) and score normalization via T-Norms. The 2001 set consists of 60 development, 174 test speakers, and a total of 20380 verification trials with the GSM codec being the prevalent type. The 2002 set contains 330 targets and 39105 trials with a majority of CDMA-quality utterances. The SWB-I (1996 SRE) part containing landline-telephone recordings (4 hrs), and an internal cellular database (2 hrs) were used to create the UBM of the baseline systems. Further details about the NIST SRE CT can be found in [17].

In the training phase, the weighting vector for each target speaker is estimated by maximizing the objective function which is calculated using a set of training utterances and speakers.

In the detection phase, log likelihood ratio scores are calculated given each test utterance, target model and the corresponding substrate model. Furthermore, the T-Norm score normalization technique is applied. A total of 234 speakers from the 2001 cellular SRE served as T-Norm speakers in both systems and are used in a gender-matched fashion in the test.

The system performance was measured at two operating points, namely in terms of the Equal-Error Rate (EER) and the minimum Detection Costs Function (DCF) as defined in the evaluation plan [17]. As shown in Table 1 and Figure 1, our maximum conditional mutual information (MCMI) approach decreases the minimum DCF by 10% and the EER by 3.2%.

5. RESULTS AND DISCUSSION

In this paper, we examined an approach for class-dependent modeling which maximizes the conditional mutual information of the classifier's scores and the class identity given that the conditional differential entropy of the scores given the class identity is constant. We applied this approach to the NIST 2003 automatic 1-speaker verification task. This approach decreased the minimum DCF by 10% compared to the baseline system and the EER by 3.2%. This improvement can be attributed to decreasing the dependency of the speaker-dependent models on the channel, hand set, and speech variabilities of the training utterances used to generate these models by emphasizing scores corresponding to robust features in the feature vector.

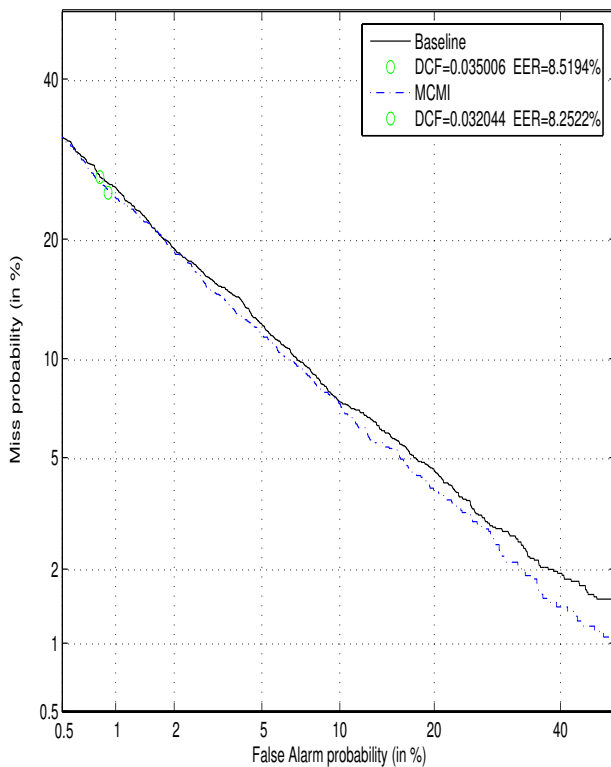


Fig. 1. Comparison of the MCMI and Baseline Systems

Further investigation of the performance of our approach on other evaluation tasks will be our main goal. We will consider also many other classification and recognition applications to our approach like in speech recognition, and text classification.

6. REFERENCES

- [1] D.A Reynolds, T.F. Quatieri, and R.B. Dunn, Speaker verification using adapted gaussian mixture models, *Digital Signal Processing*, 10(1-3):19–41, January/April/July 2000.
- [2] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, January/April/July 2000.
- [3] W. M. Campbell, Generalized Linear Discriminant Sequence Kernels for Speaker Recognition, In *Proc. of International Conference on Audio, Speech, and Signal Processing (ICASSP)*, Florida, May 2002.
- [4] D. A. Reynolds, W. Campbell, T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, A. Adami, The 2004 MIT Lincoln Laboratory Speaker Recognition System, In *Proc. of International Conference on Audio, Speech, and Signal Processing (ICASSP)*, Philadelphia, March 2005.
- [5] Olivier Siohan, Aaron E. Rosenberg, and S. Parthasarathy, The 2004 MIT Lincoln Laboratory Speaker Recognition System, In *Proc. of International Conference on Audio, Speech, and Signal Processing (ICASSP)*, Philadelphia, March 2005.
- [6] L. Heck, and Y. Konig, Discriminative Training and Minimum Cost Speaker Verification Systems, In *Proc. of RLA2-ESCA*, pp. 93–96, Avignon, France, 1998.
- [7] Q. Jin, and A. Waibel, Application of LDA to speaker recognition, In *Proc. of International Conference on Speech, and Language Processing (ICSLP)*, Beijing, China, October 2000.
- [8] Sachin S. Kajarekar, Luciana Ferrer, Elizabeth Shriberg, Kemal Sonmez, Andreas Stolcke, Anand Venkataraman, and Jing Zheng, SRI's 2004 NIST Speaker Recognition Evaluation System, In *Proc. of International Conference on Audio, Speech, and Signal Processing (ICASSP)*, Philadelphia, March 2005.
- [9] L. P. Heck, Y. Konig, M. K. Sonmez, and M. Weintraub, Robustness to Telephone Handset Distortion in Speaker Recognition by Discriminative Feature Design. *Speech Communication*, vol. 31, pp. 181–192, 2000.
- [10] B. Xiang, U.V. Chaudhari, J. Navrátil, G.N. Ramaswamy, and R.A. Gopinath. Short-time gaussianization for robust speaker verification. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, May 2002. IEEE.
- [11] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proc. Speaker Odyssey 2001*, Crete, Greece, June 2001.
- [12] J. Navrátil, U.V. Chaudhari, and G. Ramaswamy. Speaker verification using target and background dependent linear transforms and multi-system fusion. In *Proc. of the European Conference on Speech Communication and Technology (EUROSPEECH)*, Aalborg, Denmark, September 2001.
- [13] J. Navrátil and G.N. Ramaswamy. Detac - a discriminative criterion for speaker verification. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, September 2002.
- [14] D.A. Reynolds, R.B. Dunn, and J.J. McLaughlin. The lincoln speaker recognition system: Nist eval2000. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, October 2000.
- [15] U.V. Chaudhari, J. Navrátil, and S.H. Maes. Multi-grained modeling with pattern-specific maximum likelihood transformations for text-independent speaker recognition. *IEEE Trans. Speech and Audio Processing*, 2002.
- [16] J. Navrátil, U.V. Chaudhari, and S.H. Maes. A speech biometric system with multi-grained speaker modeling. In *Proc. KONVENS-2000*, Ilmenau, Germany, September 2000.
- [17] (URL). <http://www.nist.gov/speech/tests/spk/index.htm>.