DISCRIMINATIVE TRAINING OF MINIMUM COST SPEAKER VERIFICATION SYSTEMS

Larry Heck and Yochai Konig

Speech Technology and Research Laboratory
SRI International
Menlo Park, CA 94025

RÉSUMÉ

Ce papier présente une nouvelle méthode d'apprentissage pour les systèmes de vérification du locuteur. Cette méthode améliore les travaux précédents dans le domaine de vérification du locuteur en (1) développant un nouvel algorithme d'apprentissage discriminant *a posteriori*, et en (2) étendant l'algorithme pour optimiser directement les performances de la vérification du locuteur. L'élément clé de ce nouvel algorithme d'apprentissage améliorant l'état de l'art de la technologie initialise le système avec un modèle mélangé de Gauss modifié par des Bayesiens. L'algorithme d'apprentissage discriminant ajuste alors les paramètres de ces modèles pour directement minimiser une fonction du coût de la vérification (VCF) représentant le coût attendu des fausses acceptations des imposteurs et des faux rejets des locuteurs acceptables. Les résultats présentés proviennent du corpus de l'évaluation de la reconnaissance du locuteur du NIST en 1997 indiquant que la performance de la VCF peut être améliorée mais au depend d'une ré2duction de performance d'autres parties du système (différents coûts des fausses alarmes et des faux rejets).

ABSTRACT

This paper presents a new training procedure for speaker verification systems. The procedure extends previous speaker verification work by (1) developing a new discriminative *a posteriori*-based training algorithm, and (2) extending the algorithm to directly optimize speaker verification performance. The key features of the new training algorithm include leveraging current state of the art technology by initializing the system with Bayesian-adapted Gaussian mixture models. The discriminative training algorithm then adjusts parameters of these models to directly minimize a verification cost function (VCF) representing the expected costs of falsely accepting impostors and falsely rejecting true claimants. Results are presented from the 1997 NIST Speaker Recognition Evaluation corpus indicating that the VCF performance can be improved with this procedure, but at the expense of reduced system performance at other operating points (different false alarm and false rejection costs).

1. INTRODUCTION

In many applications, the goal of a speaker verification system is *not* to minimize classification error, but rather to minimize the expected cost of making an error. This expected cost can be expressed as a function of the two error types: the cost of false alarms (falsely accepting an impostor speaker) and the cost of false rejections (falsely rejecting a true speaker). If the costs associated with each error type differ, then minimizing classification error will not minimize the expected cost of making an error.

For example, minimizing the expected cost of making an error (rather than classification error) is important in detecting credit card fraud. In this application, the problem is to determine whether the voice of a person who is attempting to purchase an item with a credit card is in fact the voice of the person authorized to use that credit card. There is a real cost (in dollars) to the credit card company for missing a fraudulent user (the system accepted a transaction from a person not authorized to use the card). On the other hand, falsely accusing a valid customer of fraud (the system rejected a transaction from a person authorized to use the card) could result in lost business. For this application, it would be desirable to design a speaker verification system that considers both the cost of false acceptances and false rejections.

The expected cost of verification errors can be expressed as a verification cost function (VCF)

$$\begin{aligned}\text{VCF} = C(\text{miss})\ P(C)\ P(\text{miss}) + \\ C(\text{fa})\ P(I)\ P(\text{fa})\end{aligned} \quad (1)$$

where $C(\text{miss})$ and $C(\text{fa})$ are the costs of missing a claimant speaker and falsely accepting an impostor, respectively, $P(C)$ and $P(I)$ are the *a priori* probabilities of a claimant speaker and an impostor speaker, and $P(\text{miss})$

and $P(\text{fa})$ are the probabilities of missing a claimant and falsely accepting an impostor.

Given claimant and impostor models, much of the current work attempts to minimize the VCF by setting appropriate thresholds on the output scores, i.e., speakers who score above the threshold are accepted, and those who score below the threshold are rejected. However, this approach only attempts to affect the VCF during testing. A better approach would be to change the speaker and impostor models to minimize the VCF during training as well. Developing a discriminative framework to accomplish this is the focus of this paper.

Several approaches to discriminative training of speaker verification systems have been investigated, including the study by Liu, et. al [1] and a study by Korkmazskiy, et. al [2] on the use minimum classification error (MCE) techniques. Liu's paper extended the previous work of Juang et. al [3] by applying MCE to the speaker verification problem (Juang's original application was to speech recognition). Korkmazskiy's work focused on applications where adaptation data was available to the system for model refinement, while our focus is on the initial training step of the system. Our paper extends Liu's and Korkmazskiy's work in two ways. First, we formulate the discriminative training in an *a posteriori* framework to preserving the connection of the verification scores to probabilities. Probabilities, as opposed to a distance measure, are meaningful in an absolute sense, providing a measure of confidence about an accept/reject decision. A second extension of previous work is in our development of a training procedure that optimizes verification performance directly rather than minimizing classification error. More specifically, we minimize both false acceptances and false rejections during training according to their costs (specified by the application). The approach leverages current state of the art Bayesian-adapted Gaussian mixture model (GMM)-based speaker verification systems [6] by utilizing GMMs to represent the underlying probability density functions.

2. DISCRIMINATIVE TRAINING PROCEDURE FOR MINIMIZING COST

We can express the probability of miss and the probability of false alarms in Equation (1) as

$$\begin{aligned} P(\text{miss}) &= \int_{\mathcal{C}} P(\text{miss}, \vec{x}) d\vec{x} \\ &= \int_{\mathcal{C}} P(\text{miss} \mid \vec{x}) p(\vec{x}) d\vec{x} \end{aligned} \tag{2}$$

and

$$\begin{aligned} P(\text{fa}) &= \int_{\mathcal{I}} P(\text{fa}, \vec{x}) d\vec{x} \\ &= \int_{\mathcal{I}} P(\text{fa} \mid \vec{x}) p(\vec{x}) d\vec{x} \end{aligned} \tag{3}$$

where $\mathcal{C}$ and $\mathcal{I}$ are the sets of observations over $\vec{x}$ where the classifier decides that the observation is a claimant and impostor, respectively. Given a particular $\vec{x}$ in the claimant decision set $\mathcal{C}$, we can express the probability of miss as

$$\begin{aligned} P(\text{miss} \mid \vec{x}) &= 1 - P(C \mid \vec{x}) \\ &= P(I \mid \vec{x}). \end{aligned} \tag{4}$$

where $P(C \mid \vec{x})$ is the probability of the claimant speaker given the observation $\vec{x}$, and $P(I \mid \vec{x})$ is the probability of an impostor speaker given the observation. Likewise, for a particular $\vec{x}$ in the impostor decision set $\mathcal{I}$, we can express the probability of false alarm as

$$\begin{aligned} P(\text{fa} \mid \vec{x}) &= 1 - P(I \mid \vec{x}) \\ &= P(C \mid \vec{x}) \end{aligned} \tag{5}$$

Substituting Equations (4-5) into Equation (1), we obtain

$$\begin{aligned} \text{VCF} = {} & C(\text{miss}) P(C) \int_{\mathcal{C}} P(I \mid \vec{x}) p(\vec{x}) d\vec{x} \quad + \\ & C(\text{fa}) P(I) \int_{\mathcal{I}} P(C \mid \vec{x}) p(\vec{x}) d\vec{x}. \end{aligned} \tag{6}$$

We can approximate Equation (6) with a large training sample by

$$\begin{aligned} \text{VCF}^* = {} & C(\text{miss}) P(C) \frac{1}{N_T} \sum_{k=1}^{N_T} P(I \mid \vec{x}_{T_k}) \quad + \\ & C(\text{fa}) P(I) \frac{1}{N_I} \sum_{l=1}^{N_I} P(C \mid \vec{x}_{I_l}) \end{aligned} \tag{7}$$

where N_T and N_I are the number of observations in the training sample where the classifier decides claimant and impostor, respectively, $\vec{x}_{T_k}$ is the kth observation from the claimant decision set, and $\vec{x}_{I_l}$ is the lth observation from the impostor decision set.

To minimize the detection cost function directly, we can use a stochastic steepest descent algorithm, where model parameters at the k+1 iteration can be written as

$$\theta_{k+1} = \theta_k - \gamma \frac{\partial VCF^*}{\partial \theta} \tag{8}$$

where θ_{k+1} is the set of model parameters at the k+1st iteration and γ is (an experimentally determined) learning rate. The gradient term is computed *at a sentence level* and is given as

$$\begin{aligned} \frac{\partial VCF^*}{\partial \theta} = {} & C(\text{miss}) P(C) \frac{1}{N_T} \sum_{k=1}^{N_T} \frac{\partial P(I \mid \vec{x}_{T_k})}{\partial \theta} \quad + \\ & C(\text{fa}) P(I) \frac{1}{N_I} \sum_{l=1}^{N_I} \frac{\partial P(C \mid \vec{x}_{I_l})}{\partial \theta} \end{aligned} \tag{9}$$

We can also implement the steepest descent algorithm at the frame level, that is,

$$\frac{\partial VCF^*}{\partial \theta} = C(\text{miss})P(C)\ \frac{1}{N_T}\frac{\partial P(I \mid \vec{x}_{T_k})}{\partial \theta} + C(\text{fa})P(I)\ \frac{1}{N_I}\frac{\partial P(C \mid \vec{x}_{I_l})}{\partial \theta} \quad (10)$$

The task that remains is to derive an expression for the partials of the posteriori probabilities with respect to model parameters θ. First, we need to specify a suitable (parametric) model to represent the posteriori probability density functions. Using Bayes' rule, the posteriori probabilities in Equation (10) can be expressed as

$$P(C|\vec{x}_{T_k}) = \frac{p(\vec{x}_{T_k}|c)P(C)}{P(\vec{x}_{T_k}|c)P(C) \ + \ P(\vec{x}_{T_k}|I)P(I)} \quad (11)$$

In this work, we model the likelihood density functions $p(\vec{x}_{T_k}|c)$ and $p(\vec{x}_{T_k}|I)$ with a GMM. We use a GMM because GMMs give state of the art performance in speaker recognition applications [6], and GMMs can be efficiently initialized with the EM maximum likelihood training algorithm.

The likelihood of an observation given the model for a claimant speaker can be expressed by a GMM as

$$p(\cdot \mid c) = \sum_{m=1}^{M} \pi_m b_m(\cdot \mid \mu_m, \sigma_m) \quad (12)$$

where $b_m(\cdot \mid \vec{\mu}_m, \vec{\sigma}_m)$ is the mth Gaussian of the mixture parameterized by a mean vector $\vec{\mu}_m$ and standard deviation $\vec{\sigma}_m$. Substituting Equation (12) into Equation (11) and taking the partial with respect to each model parameter gives the expressions for the partials of the posterior probabilities.

3. EXPERIMENTAL DATABASE

The results in this paper are from text-independent experiments on the database used for the June 1997 NIST Speaker Recognition Evaluation [5]. The database is from Switchboard-II, a conversational-style corpus of telephone calls. The database consists of 401 claimant speakers (167 male and 234 female). During the testing of one of the claimant speakers, the impostors are simulated by using speech from the other 166 claimants. Only training data from the claimant and an impostor development database (excluding all claimant speakers) is used during the training of the claimant and impostor models, and the setting of decision thresholds. For all the tests in the evaluation, there are approximately 25,000 target speaker trials and 250,000 impostor trials. For the claimant model, 2 minutes of training data is available.

4. PRELIMINARY RESULTS

The parameters of the claimant and impostor models are initialized with a nondiscriminative training procedure. The initialization procedure is as follows: multiple speaker-independent GMMs are used to represent the impostor speakers. These speaker-independent GMMs are trained with an EM algorithm to maximize the likelihood of the data observations given the model. Separate impostor models are trained to represent speakers talking in various acoustic environments and gender. For example, in telephone speech, separate impostor models are trained for each handset transducer type (carbon button, electret) and each gender. Using multiple models to represent impostor speakers in various acoustic environments (in combination with a detector to automatically determine the correct acoustic environment) greatly enhances the robustness of speaker verification systems [4]. Each model used 1024 Gaussians, and was trained with approximately 5 hours of speech from the 1997 NIST Speaker Recognition development set.

Claimant speaker models are then initialized by adapting the impostor GMM that has the same handset type and gender type as the claimant. The adaptation is accomplished with an unsupervised Bayesian adaptation with the training data of the claimant [6]. The result is a 1024-Gaussian claimant model.

After initialization, the parameters of the GMMs representing the claimant and the impostor speakers are updated with the discriminative training procedure described in Section 2. Table 1 shows the costs for false alarms and false rejections, as well as the prior probabilities for claimant and impostor that we used in training (these values were specified in the 1997 NIST evaluation). To provide a balanced training sample, we used a 2:5 ratio of data for claimants vs. impostors (the prescribed 2 minutes for the claimants, and 5 minutes of impostor data from the 1996 NIST Speaker Recognition Evaluation, 3-second 1-session test).

Cost of False Reject	10
Cost of False Alarm	1
Prior Prob. of Claimant	0.01
Prior Prob. of Impostor	0.99

Table 1: Cost of false alarms and false rejections, and prior probabilities for claimants and impostors in the 1997 NIST Speaker Recognition Evaluation.

To complete the discriminative training, we divided the training data into two sets: 90% used to optimize the parameters, and 10% used as a cross validation set. As described earlier, the training procedure was iterative, with a stopping criterion based on the relative improvement (reduction) of the VCF on the cross validation set. The

initial learning rate, γ in Equation (8), was experimentally determined, and updated automatically depending on the relative improvement of the VCF in cross validation.

To illustrate the properties of the discriminative minimum cost training procedure developed in this paper, Table 2 shows the performance on training and cross validation for one of the claimant speakers for each epoch of training (complete passes through the frames of data in the training set). The training procedure resulted in an 8.6% improvement (decrease) in the VCF score for the training set, and a 2.3% improvement in the VCF score for the cross validation set. It is interesting to note that the VCF scores and the probability of the correct class are not necessarily monotonic with the training epochs. For example in the cross validation results, the probability of the correct class worsens (gets smaller) from the first to the second epoch while the VCF improves. This suggests that training the classifier to maximize the probability of correct (minimize probability of error) does not necessarily result in the best verification performance (VCF).

Training Set		
Epoch #	Avg. Posterior Prob. of Correct Class	VCF (x10^3)
0	0.743	107.7
1	0.734	107.5
2	0.773	98.4
Cross Validation Set		
0	0.726	102.6
1	0.762	103.0
2	0.750	100.2

Table 2: Performance of discriminative training procedure on 1997 NIST Evaluation (male, 10 second test, 1 session training) for one speaker on training and cross validation sets.

Table 3 shows results for the 1-session (2 minutes of training from one phone call), 10-second male test in the 1997 NIST Speaker Recognition evaluation where the same telephone was used in by the claimant in training and testing. Table 4 shows results for the same training case, but with tests from a different telephone than used in training. The first column of each table shows the equal error rate (EER), the second column shows the probability of false alarm at a 10% miss (false reject) rate, and the third column shows the VCF score. (The VCF scores are in a different range from the training set because the training VCF was computed at the frame level). The discriminative procedure reduces the VCF by 5.3% for the matched case and 2.8% for the mismatched case as compared to our baseline system (EM-trained impostor model, with Bayesian adaptation to the claimant model). This gain is at the cost of increased errors for other operating conditions (different costs), as illustrated at the probability of false alarm at a 10% miss rate.

System	EER	P_{fa}(@10% miss)	VCF (x10^3)
Baseline	7.0	4.1	34.0
Min. Cost (Discrim.)	7.1	4.6	32.2

Table 3: 1 session, 10 second male test, 1997 NIST Speaker Recognition Evaluation corpus (matched telephone numbers)

System	EER	P_{fa}(@10% miss)	VCF (x10^3)
Baseline	16.6	29.4	60.9
Min. Cost (Discrim.)	16.5	31.4	59.1

Table 4: 1 session, 10 second male test, 1997 NIST Speaker Recognition Evaluation corpus (mismatched telephone numbers)

5. DISCUSSION AND FUTURE WORK

We presented a new *a posteriori*-based training procedure for speaker verification systems that optimizes verification performance directly. The framework is flexible and facilitates a principled design of the classifier to improve performance at a desired operating point (specified by the costs of false alarms and rejections) by giving up performance at other operating points. In addition to training the classifier to directly optimize verification performance, future work will extend the training procedure to include the design of the feature extractor.

6. REFERENCES

[1] W. Chou B.-H.Juang C.-S. Liu, C.-H. Lee and A.E. Rosenberg, "A study on minimum error discriminative training for speaker recognition," *JASA*, 97(1):637–648, 1995.

[2] F.Korkmazskiy and B.-H.Juang, "Discriminative adaptation for speaker verification," *ICSLP*, 1996.

[3] B.-H.Juang and S.Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Proc.*, 40(12):3043–3054, 1992.

[4] L.P. Heck and M. Weintraub, "Handset dependent background models for robust text-independent speaker recognition," *ICASSP*, 1997.

[5] *NIST Speaker Recognition Workshop*, Linthicum Heights, Maryland, 1997.

[6] D.A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," *EUROSPEECH*, 1997.