SPEAKER IDENTIFICATION USING MINIMUM CLASSIFICATION ERROR TRAINING

Olivier Siohan *Aaron E. Rosenberg* *S. Parthasarathy*

AT&T Labs – Research
180 Park Avenue, Florham Park
NJ 07932-0971, USA

ABSTRACT

In this paper we use a Minimum Classification Error (MCE) training paradigm to build a speaker identification system. The training is optimized at the string level for a text-dependent speaker identification task. Experiments performed on a small set speaker identification task show that MCE training can reduce closed-set identification errors by up to 20-25% over a baseline system trained using Maximum Likelihood Estimation. Further experiments suggest that additional improvement can be obtained by using some additional training data from speakers outside the set of registered speakers, leading to an overall reduction of the closed-set identification errors by about 35%.

1. INTRODUCTION

In a hidden Markov model (HMM)-based speaker identification system, each speaker to be identified is modeled by a set of HMMs denoted $\Lambda_k = \{\lambda_k^w, 1 \leq w \leq W\}$, in which λ_k^w is an HMM characterizing the speech unit w of speaker k. According to the application, the speech unit can be a subword unit, a word unit, or an utterance unit. In the following, we denote W the number of speech units and K the number of speakers.

Speaker identification is usually carried out via likelihood (or equivalently log-likelihood) computation. Given a sequence O of acoustic observations, the log-likelihood of O, $L(O;\Lambda_k)$, is evaluated using each set Λ_k. The identified speaker is the one whose set of HMMs leads to the highest likelihood. More formally, for closed-set speaker identification this scenario corresponds to the following identification rule where $\hat{k}$ is the identified speaker:

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, L(O;\Lambda_k). \quad (1)$$

In conventional speaker identification systems, each set of hidden Markov models Λ_k is derived through Maximum Likelihood Estimation (MLE). Given a collection of training acoustic observations $\mathcal{O}_k$ from speaker k, the parameters of Λ_k are estimated so that the likelihood $L(\mathcal{O}_k;\Lambda_k)$ of the training data is maximized. The MLE training aims at approximating the underlying distribution of acoustic units of each speaker. This is a sub-optimal procedure for classification [1] since the estimated distribution deviates from the true one due to incorrect modeling assumptions and insufficient training data. Hence, the "optimal" MLE criterion for density estimation does not imply an "optimal" classifier design.

To overcome some limitations of the MLE training criterion, a classifier design procedure called discriminative learning has been introduced [2, 1]. The goal of discriminative training is to estimate model parameters which minimize the classification errors of the training data. This so-called Minimum Classification Error (MCE) criterion maximizes the separation between speaker models and is therefore consistent with the goal of speaker identification. MCE training has been successfully applied for speech recognition [3], string verification [4, 5], speaker verification [6, 7] and closed-set speaker identification [8, 9]. In this paper, we propose a string-based minimum classification error training for a text-dependent speaker identification task, and discuss related issues such as the choice of the misclassification measure and the effect of variance reestimation. We also show how the use of additional training data from outside the set of registered speakers can improve the overall discrimination among speakers. Experiments carried out on a large telephone speech database show that the proposed approach leads to improved performance over a standard MLE-based system.

The organization of the paper is as follows. Section 2 introduces the general framework of minimum classification error training. Experiments and results are presented in section 3 based on a fixed-password speaker identification task. Section 4 concludes the paper.

2. FRAMEWORK OF MINIMUM CLASSIFICATION ERROR TRAINING

The goal of minimum classification error training is to derive a set of speaker models $\{\Lambda_k, 1 \leq k \leq K\}$ which minimizes the classification errors of the training data. This is achieved by deriving an approximation of the total number of misclassification errors of the training corpus as a functional form of all model parameters. Then, a generalized probabilistic descent (GPD) algorithm is applied to minimize this function with respect to these parameters. To derive this functional form, GPD is based on 3 functions, defined as follows.

First, a set of *discriminant functions* $g_k(O;\Lambda_k), 1 \leq k \leq K$, is defined, where $g_k(O;\Lambda_k)$ usually represents the log-likelihood of the observation sequence O given the models Λ_k for speaker k. An observation sequence O is classified using the following decision rule:

$$\hat{k} = \underset{k}{\operatorname{argmax}}\, g_k(O;\Lambda_k). \quad (2)$$

Then, a set of *misclassification measures* is defined for each speaker, which attempt to evaluate how likely an observation spoken by speaker k is misclassified:

$$d_k(O;\Lambda) = -g_k(O;\Lambda_k) + G_k(O;\Lambda), \quad (3)$$

where Λ denotes the set of all speaker HMMs, and $G_k(O;\Lambda)$ is the anti-discriminant function for speaker k. $G_k(O;\Lambda)$ is defined

so that $d_k(O;\Lambda)$ is non-positive if O is correctly classified and $d_k(O;\Lambda)$ is positive if O is mis-identified. For speech recognition problems, $G_k(O;\Lambda)$ is usually defined as a collective representation of all competing classes, as follows:

$$G_k(O;\Lambda) = \log\left[\frac{1}{K-1}\sum_{j\neq k}\exp[g_j(O;\Lambda)\eta]\right]^{1/\eta}, \quad (4)$$

where η is a positive coefficient used to control the weight of competing classes.

It has been argued that for a speaker identification task, the use of misclassification measures based on individual representation of competing speakers (instead of a collective representation of competing speakers) might be more appropriate [9]. Instead of using a single misclassification measure $d_k(O;\Lambda)$ per speaker k, several pairwise misclassification measures can be defined, with respect to competing speakers:

$$d_{kj}(O;\Lambda) = -g_k(O;\Lambda_k) + g_j(O;\Lambda_j), \qquad j \neq k. \quad (5)$$

For a given speaker k, $K-1$ misclassification measures $d_{kj}()$ can therefore be defined. However, in this paper we show how to limit the number of misclassification measures associated to a given speaker k to a smaller subset by using only the misclassification measures involving the N-best ($N < K-1$) competing speakers. Given an observation sequence O and a speaker k, the set of N-best competing speakers is the set of speakers whose corresponding discriminant functions are the closest to $g_k(O;\Lambda_k)$. We denote $\mathcal{K}(O,k,N)$ this reduced set of competing speakers.

Each misclassification measure $d_{kj}(O;\Lambda)$ is embedded into a smooth empirical *loss function* which approximates a loss (between 0 and 1) directly related to the number of classification errors:

$$l_{kj}(O;\Lambda) = \frac{1}{1+\exp(-a d_{kj}(O;\Lambda))}. \quad (6)$$

The positive constant a is used to control the slope of the decision threshold. If d_{kj} is much smaller than zero, which corresponds to a correct classification, the loss is negligible. When d_{kj} is significantly positive, the loss becomes close to 1.

The overall empirical loss associated with a given observation O is given by:

$$l(O;\Lambda) = \sum_k \sum_{j\in\mathcal{K}(O,k,N)} l_{kj}(O;\Lambda)1_k(O), \quad (7)$$

where $1_k(\cdot)$ is the indicator function defined as:

$$1_k(O) = \begin{cases} 1 & \text{if } O \text{ is uttered by speaker } k, \\ 0 & \text{otherwise.} \end{cases} \quad (8)$$

Given a set of observations $\mathcal{O}$, the total loss is:

$$l(\mathcal{O};\Lambda) = \sum_{O\in\mathcal{O}} l(O;\Lambda), \quad (9)$$

a function of all model parameters.

Using a gradient descent algorithm, it becomes possible to derive all model parameters so that the total loss is minimized:

$$\Lambda^{n+1} = \Lambda^n - \epsilon_n \nabla_{\Lambda^n} l(\mathcal{O};\Lambda^n), \quad (10)$$

where n denotes the iteration number of the GPD algorithm, $\nabla_{\Lambda^n} l(\cdot)$ is the gradient of the loss function, and ϵ_n is the step size.

3. EXPERIMENTAL EVALUATION

3.1. Database description

The speech database is part of a large database of spoken phrases recorded digitally over the telephone network by AT&T. Volunteers and paid subjects recorded utterances from their home, office or other phones by dialing a toll-free number, and were encouraged to use a variety of phones, excluding speakerphones. The data used for the purpose of this evaluation consists of a phrase common to all speakers ("I pledge allegiance to the flag").

The database contains utterances spoken by 50 male speakers. Each speaker provided 6 tokens of the common phrase in a single training session. Two tokens of the common phrase were also recorded in each of 25 testing sessions. Consequently, a total of 50 test utterance tokens are available from each speaker.

3.2. Front-end processing

The signal is first passed through a 3200Hz lowpass anti-aliasing filter. A 300Hz highpass filter is then applied to minimize the effect of processing in the telephone network. The resulting signal is pre-emphasized using a first order difference and 10th order linear predictive coding (LPC) coefficients are derived every 10ms over 30ms Hamming windowed segments. The 10 LPC coefficients are converted to 12th order cepstral coefficients (LPCC) and a feature vector of 24 components, consisting of 12 LPCC and their first derivatives is produced at each frame.

3.3. System description

For each speaker, only 3 utterance tokens from the training session are used to train a whole-phrase unit continuous density left-to-right HMM and one silence unit using MLE. The number of model states is 30 for the whole-phrase unit and 3 for the silence unit. The nominal number of mixtures per state is 4. A baseline system is built by estimating the HMM parameters using a standard segmental $\mathcal{K}$-means algorithm [10]. A fixed global diagonal covariance matrix is used in this baseline system.

Speaker identification experiments are performed in closed set mode, meaning that the goal is to identify which of the registered speakers spoke a given utterance. In these experiments, we are interested in identifying the speaker from small groups of speakers. Two different group sizes (5 and 10 speakers) are considered.

In a first set of experiments, 140 5-speaker groups are created by uniformly selecting the speakers from the entire customer population. A given speaker can therefore appear in 14 different groups. The goal is to identify the speaker of 250 (50 test utterances per customer, 5 customers per group) test utterances per group. The identification tasks are carried out for each of 140 groups.

In a second set of experiments, 70 groups of 10 speakers each are created. These groups are obtained by combining the 5-speaker groups. A total of 500 (50 test utterances per speaker, 10 speakers per group) test utterances per group are classified as one of 10 speakers in each of the 70 groups. Experimental results are given in terms of identification errors averaged over all test utterances. The total number of test tokens is about 35000.

# of MCE iterations	20	40	60	80	100
MCE (N=4)	2.40	2.40	2.48	2.55	2.57
MCE (N=1)	2.23	2.14	2.08	2.10	2.06
MLE (baseline)			2.59		

Table 1: Speaker identification error rates (%), group size 5, MCE training using $N = 4$ and $N = 1$ (N denotes the number of competing speakers).

# of MCE iterations	20	40	60	80	100
MCE (N=4)	4.09	3.74	3.60	3.53	3.52
MCE (N=1)	3.49	3.30	3.20	3.19	3.17
MLE (baseline)			4.32		

Table 2: Speaker identification error rates (%), group size 10, MCE training using $N = 4$ and $N = 1$ (N denotes the number of competing speakers).

3.4. Experimental results

3.4.1. MCE setting

For each speaker group, the MCE training is performed by updating all parameters (means, global covariance and mixture weights) of the initial MLE models using (10), where $\mathcal{O}$ is the set of training utterances from the speakers in the group (*e.g.* when the group contains 5 speakers, $\mathcal{O}$ contains $3 \times 5 = 15$ training tokens). However, instead of updating the model parameters once after computing the gradient of the loss over the whole training data set, the loss function is evaluated and the models are updated training token by training token. Once the whole training corpus has been processed, the training is re-iterated a specified number of times. In most experiments, the training is iterated 100 times and the models obtained after every 20 iterations are saved to monitor the convergence of the identification rate. A small slope ($a = 0.1$) is chosen for the sigmoidal loss functions, which forces the models to be updated even if the training data are correctly classified.

3.4.2. Influence of the number of misclassification functions

Two different values for the size N of the competing speaker set $\mathcal{K}(O, k, N)$ are used. Experiments were carried out with $N = 1$ where only one misclassification function per speaker is used and also for $N = 4$, where 4 misclassification functions per speaker are used. Results are given in Tables 1 and 2 for group sizes 5 and 10.

We first observe that for group size 5, when $N = 4$, MCE leads to a slight improvement over MLE when the number of iterations is small (20 or 40). However, it appears that iterating the training does not lead to stable set of models since the improvement is not consistent as the number of iterations increases. We assume that this is related to the small amount of training data and to the use of too many misclassification functions which makes it difficult to simultaneously minimize all loss functions. By reducing the number of misclassification functions to 1 ($N = 1$), the training process appears to stabilize and a monotonic reduction of the identification error is obtained as the number of iterations increases. Compared to the baseline MLE system, MCE training leads to a 20% reduction of the identification error rate.

Better results are also obtained for group size 10 when $N = 1$. Approximatively 25% reduction of the identification error rate is obtained over the standard MLE system. We also point out that the identification error rate monotonically decreases as the number of iterations increases, even when more than one misclassification measure ($N = 4$) is used. This suggests that for group size 10, the amount of training data becomes large enough to get a stable estimate of MCE models. A possible explanation is that when the number of speakers per group increases, the amount of training data used to build a given model also increases. However, using a single misclassification measure leads to better results so N is set to 1 in all subsequent experiments.

# of MCE iterations	20	40	60	80	100
MCE updated variance	2.23	2.14	2.08	2.10	2.06
MCE, fixed variance	2.38	2.28	2.22	2.20	2.14

Table 3: Speaker identification error rates (%), group size 5, $N = 1$, with and without updating the variance during MCE training (first line taken from table 1).

# of MCE iterations	20	40	60	80	100
MCE, updated variance	3.49	3.30	3.20	3.19	3.17
MCE, fixed variance	3.92	3.65	3.47	3.37	3.28

Table 4: Speaker identification error rates (%), group size 10, $N = 1$, with and without updating the variance during MCE training (first line taken from table 2).

3.4.3. Influence of the variance reestimation

The baseline MLE system uses HMMs with a fixed global variance. This is used as the initial estimate in MCE models. In the previous set of experiments, model variances are updated during the MCE training and therefore become specific to each speaker and group. Tables 3 and 4 present identification errors on group size 5 and 10 with and without updating the variances. It appears that adapting the variances leads to better results and to a faster convergence of the identification error rate.

3.4.4. Extended training data set

The amount of training data is increased by adding in each group 20 additional training utterances from 10 speakers outside the group. We assume that these extra-training utterances come from a "dummy" background speaker whose initial model was created by concatenating speaker independent phone-HMMs derived from an independent database. The MCE training is performed as previously,

# of MCE iterations	20	40	60	80	100
Original training corpus	2.21	2.08	2.06	2.01	2.02
Extended training corp.	1.85	1.74	1.68	1.65	1.64

Table 5: Speaker identification error rates (%), group size 5, with and without extended training data set.

# of MCE iterations	20	40	60	80	100
Original training corpus	3.44	3.20	3.16	3.13	3.10
Extended training corp.	3.18	2.95	2.87	2.85	2.78

Table 6: Speaker identification error rates (%), group size 10, with and without extended training data set.

except that there are now 5+1 (or 10+1) speakers in each group[1]. To summarize, the training corpus per group consists of 3 training utterances per speaker and 20 extra background utterances. During the training, all speaker models as well as the background model are updated.

Closed-set identification results are given in Tables 5 and 6 for group size 5 and 10. The first line of the tables correspond to the situation where only the 3 training utterances per speaker for each group are used[2]. It appears that the use of the extended corpus significantly decreases the identification error rate, leading to an overall reduction of 35% for both group size 5, and 10 with respect to the original MLE-based system. It is possible that this extended training set effectively provides more training data to learn the decision rules, and therefore leads to a more robust estimate of the classification boundaries.

4. CONCLUSION

A Minimum Classification Error training paradigm has been applied on a small set text-dependent speaker identification task. Experiments performed on a telephone speech database have shown that the MCE-based system outperforms the MLE-based system by about 20 to 25% on a closed-set evaluation. It appears that additional improvement can be obtained by using some additional training data from speakers outside each group. This suggests that the MCE training scenario can be useful when the number of training data is very small and the resulting MLE-models are poorly estimated.

[1]The background model was originally introduced for open-set experiments. In the MCE training process, an additional misclassification measure is actually added to discriminate the background speaker from speakers in the group.

[2]The first lines of Table 5 and 3 (Table 6 and 4 for group size 10) are slightly different since in the extended training set experiments a background model is also trained, while there was no such background model in previous experiments.

5. REFERENCES

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

[2] A. Ljolje, Y. Ephraim, and L. R. Rabiner. Estimation of hidden Markov model parameters by minimizing empirical error rate. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 709–712, Albuquerque, New Mexico, April 1990. ICASSP'90.

[3] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Transactions on Speech and Audio Processing*, 5(3):257–265, May 1997.

[4] M. G. Rahim, C.-H. Lee, and B.-H. Juang. Discriminative utterance verification for connected digits recognition. *IEEE Transactions on Speech and Audio Processing*, May 1997.

[5] R. A. Sukkar, A. R. Setlur, M. G. Rahim, and C.-H. Lee. Utterance verification of keyword strings using word-based minimum verification error (WB-MVE) training. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 518–521, 1996.

[6] F. Korkmazskiy and B.-H. Juang. Discriminative adaptation for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 28–31, Philadelphia, USA, 1996. ICSLP'96.

[7] A. E. Rosenberg, O. Siohan, and S. Parthasarathy. Speaker verification using minimum verification error training. Submitted to ICASSP98.

[8] C. Martin del Álamo, F. J. Caminero Gil, C. de la Torre Munilla, and L. Hernandez Gomez. Discriminative training of GMM for speaker identification. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 89–92, 1996.

[9] C. Martin del Álamo, J. Alvarez, C. de la Torre, F. J. Poyatos, and L. Hernández. Incremental speaker adaptation with minimum error discriminative training for speaker identification. In *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 312–315, 1996.

[10] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang. A segmental K-means training procedure for connected word recognition. *AT&T Bell Labs Tech. J.*, 65(3):21–31, 1986.