DISCRIMINATIVE ADAPTATION FOR SPEAKER VERIFICATION

F. Korkmazskiy *B.-H. Juang*

AT&T Bell Labs, Murray Hill, NJ 07974, USA
yelena@research.att.com

ABSTRACT

This paper describes a speaker verification system in which the talker and imposter models are adapted to achieve maximum discrimination, or equivalently minimum verification error. This goal is accomplished by extending the minimum error classification criterion (MCE) and generalized probabilistic descent (GPD) algorithm to the task of adapting talker model parameters and the corresponding anti-talker model parameters to the test environments so as to minimize an empirical estimate of the verification error rate. We address in the current study adaptation of two types of parameters: the model parameters and the decision threshold. We have obtained substantial improvements in equal error rate by applying combined techniques involving a simplified MAP (maximum a posteriori) method and the GPD algorithm. The equal error rate for a database of 43 talkers with 5 adaptation utterances each was reduced from the previously reported best result of 5.41% [1] to 2.17%. We will discuss several alternative methods that have been investigated in this work to provide comparative insights for the use of discriminative methods in speaker verification tasks.

1. INTRODUCTION

Adaptation in speaker verification system was proved to be one of the most effective tool to comply with changing speaker and channel conditions. But methods, employed so far for the task of speaker verification adaptation, first, used only knowledge about true speaker adaptation samples and, second, they affected only a true talker model and did not adjust an imposter model. So, it is reasonable to include in the adaptation procedure some additional information taken from imposter talkers and to use some kind of discriminative methods to attain better separation between true and imposter models. One possible way to apply discriminative approach for speaker verification system adaptation is to use GPD method [2], which traditionally has been used for the task of training. GPD may use some information taken from the samples, which have already been classified by the speaker classification system (particularly, by the speaker verification system). Based on assumption that the class of the sample is already known, one can adjust model parameters, using stochastic discriminative technique supplied by the GPD method. Similar ideas were used for speech recognition and it was shown to be effective in conjuction with traditional adaptation techique [3].

2. GPD ADAPTATION FOR SPEAKER VERIFICATION

2.1. GPD Adaptation of HMMs

The GPD method [2] is based on gradual adjustment of HMM parameters to minimize the expectation $E[\ell(O|\Lambda)]$ of the classification error $\ell(O|\Lambda)$. For verification we can use the same correction formulas as for recognition, but the formula for classification error $\ell(O|\Lambda)$ estimation has two different forms for true and imposter training data. If O_t and O_i are training samples for the true speaker and the imposters, the corresponding formulas for $\ell(O_t|\Lambda)$ and $\ell(O_i|\Lambda)$ take on such a form:

$$\ell(O_t|\Lambda) = \frac{1}{1+\ell^{-\gamma(d(O_t|\Lambda)+b)}} \tag{2.1.1}$$

$$\ell(O_i|\Lambda) = \frac{1}{1+\ell^{-\gamma(d(O_i|\Lambda)-b)}}. \tag{2.1.2}$$

Here, $\Lambda = \{\Lambda_t; \Lambda_i\}$ is the parameter set of true talker model Λ_t and the corresponding imposter model Λ_i; b and γ are GPD parameters ($\gamma > 0$).

The adjustment of model Λ usually is implemented as follows:

$$\Lambda^{(k+1)} = \Lambda^{(k)} - \xi \bigtriangledown \ell(O|\Lambda)\Big|_{\Lambda=\Lambda^{(k)}}. \tag{2.1.3}$$

Here $\Lambda^{(k)} = \{\Lambda_i^{(k)}; \Lambda_t^{(k)}\}$ is the parameter set of true talker $\Lambda_t^{(k)}$ and imposter $\Lambda_i^{(k)}$ models at the k-th step of the parameter adjustment iteration. The gradient $\bigtriangledown\ell(O|\Lambda)$ may be expressed as follows:

$$\begin{cases} \bigtriangledown\ell(O|\Lambda) = \dfrac{\partial\ell(O|\Lambda)}{\partial d(O|\Lambda)} \cdot \left(\dfrac{\partial d(O|\Lambda)}{\partial g(O|\Lambda_t)} \bigtriangledown g(O|\Lambda_t) + \dfrac{\partial d(O|\Lambda)}{\partial g(O|\Lambda_i)} \bigtriangledown g(O|\Lambda_i) \right) \\ d(O|\Lambda) = \pm g(O|\Lambda_i) \mp g(O|\Lambda_t) \\ \ell(O|\Lambda) = \dfrac{1}{1+\ell^{-\gamma(d(O|\Lambda)\pm b)}}, \quad \text{the upper sign corresponds to } O = O_t \text{ (true speaker sample)} \end{cases} \tag{2.1.4}$$

Continious digit strings were used for the training and testing in the speaker verification task. Two rosters of data (A and B) were used in the experiments. A small set of adaptation data, collected under test conditions, were used for GPD adjustment of verification system parameters. The adaptation samples for imposter were taken from an expanded set of imposter's training data from roster B. The adaptation samples for a true speaker were taken from the 5 first test sentences in roster A. The process of GPD adaptation is a sequence of adjustment steps. At the odd steps of GPD adaptation, the true speaker samples were used to correct both the true talker and the imposter HMM parameters. At the even steps, the imposter speakers were used to adapt the true talker and the imposter HMM parameters. The average equal-error rate after GPD adaptation using test sentences for 43 talkers in the roster A is 2.58%, comparing to the 4.40% without GPD adaptation.

The above result relates to nonsequntial GPD adaptation; that is, for adaptation we use a batch of the adaptation samples, extracted from the several true talkers adaptation sentences. To avoid storage of the adaptation samples we can use GPD adaptation sequentially, sentence by sentence. So, for each adaptation sentence, we adapt *only* those HMMs which corespond to the words in the given adaptation sentence. Let's describe some specific features of this sequential GPD adaptation.

1) The process of sequential GPD adaptation is a sequence of adjustment steps. In our implementation at the odd steps of GPD adaptation the same *true talker samples* for a current words in the adaptation sentence were used. At the even adaptation steps different *imposter samples* for the same current word, randomly selected from a set of imposter samples, were used.
2) HMM parameter adjustment for the current word was acomplished using the same parameter ξ for 2 subsequent odd and even steps of GPD. Then for each next pair of steps the parameter ξ was decreased as follows:

$$\xi = \xi - \frac{2\xi_0}{N}\,, \tag{2.1.5}$$

where ξ_0 is an initial value of the the parameter ξ and N is the total number of GPD adaptation steps for the current word. Each time, starting HMMs adjustment for a new word in the sentence, the parameter ξ was given its initial value ξ_0.

The result of the described sequential GPD adaptation for 43 talkers in the roster A is 2.50% equal-error rate. In all subsequent experiments, described below, only the sequential GPD is used.

The traditional adaptation methodology, that has been used so far for HMMs adjustment in speaker verification tasks, was based on evaluation of updated estimates for mixture components means $\hat{\mu}_{jm}$, weight coefficients $\hat{c}_{jm}$, and frame counts $\hat{N}_{jm}$ ([1]), and may be considered as a simplified version of MAP adaptation [4]. The result of such MAP adaptation for the given digit string database is 2.56% equal-error rate. To improve performance of MAP adaptation it was proposed to use a 2-step adaptation:

- MAP adaptation for all words constituting the current adaptation sentence;
- GPD adaptation for all words constituting the current adaptation sentence.

The result of the experiments with proposed combined MAP $\Rightarrow$ GPD adaptation is 2.24% equal-error rate. It shows that discriminative adaptation outperforms statistical (MAP) adaptation. Note that the quality of GPD adaptation as representative of discriminative methods dependends on the quality of word samples presented for discriminative adjustment of the corresponding HMMs. For the test sentences word level segmentation is usually obtained with the help of speaker independent HMMs. This segmentation may be improved using speaker dependent HMMs updated after MAP adaptation. The results of this combined MAP$\Rightarrow$ segmentation$\Rightarrow$ GPD adaptation is 2.17%.

We can conclude that combined adaptation leads to improved results. In this study, using only 5 adaptation sentences, the reduction in equal-error rate constituted more than 15% (from equal-error rate of 2.56% for MAP adaptation to 2.17% for combined MAP$\Rightarrow$ segmentation$\Rightarrow$ GPD adaptation). GPD based adaptation results may, for example, be compared with results of the MAP adaptation obtained on the same database (i.e. roster A and roster B) as in the current study [1]. Using a cohort representation of the imposter model and 6 true talker adaptation sentences for adaptation of the true speaker HMM, an equal-error rate of $\underline{5.41}\%$ was achieved in [1], compared to the best results of $\underline{2.17}\%$ in our study.

2.2. Verification Threshold Adaptation by GPD

GPD adaptation implies that we optimize HMMs parameters in terms of minimizing the expected verification error $E[\ell(O|\Lambda)]$. Assuming equal probability for true speaker samples O_t and imposter samples O_i we can express $E[\ell(O|\Lambda)]$ as follows:

$$E[\ell(O|\Lambda)] = E[\ell(O_t|\Lambda)] + E[\ell(O_i|\Lambda)], \tag{2.2.1}$$

where $E[\ell(O_t|\Lambda)]$ and $E[\ell(O_i|\Lambda)]$ are the corresponding expectations of the false rejection and the false acceptance errors $\ell(O_t|\Lambda)$ and $\ell(O_i|\Lambda)$ respectively evaluated according to (2.1.1) and (2.1.2).

Usually, the bias parameter b is assumed to be fixed in the procedure, and it is considered only as a parameter of the GPD method. But we can consider b as a variable and try to find its optimal value to obtain minimum of $E[\ell(O|\Lambda_v, b_v)]$ (Λ_v and b_v are HMM parameters and the bias for unit v). To optimize b_v by means of GPD method, we have to derive the corresponding partial derivatives for b_v:

$$\frac{\partial \ell(O_t|\Lambda_v, b_v)}{\partial b_v} = \gamma \ell(O_t|\Lambda_v, b_v)(1 - \ell(O_t|\Lambda_v, b_v)) \tag{2.2.2}$$

$$\frac{\partial \ell(O_i|\Lambda_v, b_v)}{\partial b_v} = -\gamma \ell(O_i|\Lambda_v, b_v)(1 - \ell(O_i|\Lambda_v, b_v)). \tag{2.2.3}$$

For each new sample O, that represents unit v, adjustment of b_v may be done according to formula:

$$b_v^{(k+1)} = b_v^{(k)} - \xi_b \left.\frac{\partial \ell(O|\Lambda_v, b_v)}{\partial b_v}\right|_{b_v = b_v^{(k)}}, \tag{2.2.4}$$

where ξ_b is the correction step size for parameter b. We can, for example, apply GPD adaptation to adjust both HMMs Λ_v and parameter b_v for all units v. Values b_v may be used to obtain adapted verification threshold. This threshold may be considered as an *estimate* of the optimal separation boundary between the true talker and the imposter models.

Assume, the total number of frames for unit v over all training samples is equal to F_v (units may denote words or some phonemes). Then, the common verification threshold T_V for V units may be expressed as follows:

$$T_V = \frac{\sum_{v=1}^{V} F_v \cdot \overline{b}_v}{\sum_{v=1}^{V} F_v}. \tag{2.2.5}$$

Here $\overline{b}_v$ is the optimal value of b_v, obtained for unit v after GPD adaptation. Hereafter we'll refer to threshold T_V, evaluated according (2.2.5), as `common verification threshold`.

Another approach to verification threshold adaptation assumes that verification threshold T_V has to be unique for each verification sequence $W = \{W_1, \ldots, W_n, \ldots, W_N\}$ of speech units (e.g. it should be dependent on the sequence W of words used for speaker verification). Then,

$$T_V = T(W) \tag{2.2.6}$$

Using information about optimal values of parameters $\{\overline{b}_{W_n}\}_{n=\overline{1,N}}$, the function $T(W)$ may be evaluated as follows:

$$T(W) = \frac{\sum_{n=1}^{N} F_{W_n} \cdot \overline{b}_{W_n}}{\sum_{n=1}^{N} F_{W_n}}. \tag{2.2.7}$$

Here, F_{W_n} is the total number of frames for unit W_n in the sequence W. Hereafter we'll refer to threshold $T(W)$, evaluated according (2.2.7), as `variable verification threshold`.

Several experiments were conducted in order to study the effectiveness of the proposed verification thresholds measures (2.2.5) and (2.2.7). In the threshold adaptation experiments a verification sentence was used to adapt verification threshold, *only* if this sentence was classified as being uttered by the true talker. To apply GPD for evaluation of an optimal separation boundary $\overline{b}_v$ for each unit v we need to set some initial value $b_v^{(0)}$ for each b_v. We used for $b_v^{(0)}$ a value, such that in histogram for *imposter* ***adaptation*** samples for the unit v, 5% of the normalized score for all these samples are larger than value $b_v^{(0)}$. Using values $b_v^{(0)}$ for all units v, we can get an estimate for the initial threshold $T_V^{(0)}$ as follows:

$$T_V^{(0)} = \frac{\sum_{v=1}^{V} F_v \cdot b_v^{(0)}}{\sum_{v=1}^{V} F_v} \tag{2.2.8}$$

Here, F_v is the total number of frames for unit v over all *training* samples. Hereafter we'll refer to threshold $T_V^{(0)}$, evaluated according (2.2.8), as `initial verification threshold`.

In our experiments we introduced a parameter λ_i, characterizing the probability of a request from an imposter talker. Obviously, that $\lambda_i + \lambda_t = 1$ (λ_t is the probability of a request from the true talker). Assigning to λ_i some large value we can simulate situation, when speech verification system is attacked by some group of imposters, trying to penetrate into the verification system. In the experiments the simulation of the true talker and imposter requests was conducted by `Monte Carlo` method. Speaker verification was done using 3 groups of trials:

1) *The same* `initial verification threshold` $T_V^{(0)}$ was used for each test sentence. The threshold was not subjected to any adaptation.
2) The `common verification threshold` T_V was used.
3) The `variable verification threshold` $T(W)$ was used.

In the first set of experiments the value of probability λ_i for imposters was selected to be equal to the probability λ_t for the true talker($\lambda_i = \lambda_t = 0.5$). The value of $\lambda_i = 0.5$ allows us to simulate situation, corresponding to the medium rate of imposters attack. The results of the verification experiments, simulating the medium rate of imposters attack, are: false rejection error rate $P_r = 8.10\%$ and false acceptance error rate $P_a = 3.24\%$ for the initial verification threshold, false rejection error rate $P_r = 8.94\%$ and false acceptance error rate $P_a = 2.76\%$ for the common verification threshold, and false rejection error rate $P_r = 9.14\%$ and false acceptance error rate $P_a = 2.08\%$ for the variable verification threshold.

The results of the verification experiments, simulating low rate of imposters attack($\lambda_i = 0.1$), are: false rejection error rate $P_r = 8.10\%$ and false acceptance error rate $P_a = 4.42\%$ for the initial verification threshold, false rejection error rate $P_r = 9.54\%$ and false acceptance error rate $P_a = 1.86\%$ for the common verification threshold, and false rejection error rate $P_r = 9.48\%$ and false acceptance error rate $P_a = 1.63\%$ for the variable verification threshold.

The results of the verification experiments, simulating high rate of imposters attack($\lambda_i = 0.9$), are: false rejection error rate $P_r = 8.06\%$ and false acceptance error rate $P_a = 1.29\%$ for the initial verification threshold, false rejection error rate $P_r = 8.04\%$ and false acceptance error rate $P_a = 1.25\%$ for the common verification threshold, and false rejection error rate $P_r = 8.12\%$ and false acceptance error rate $P_a = 1.25\%$ for the variable verification threshold.

From the derived experimental results we can see that quality of threshold adaptation strongly depends on the value of paramater λ_i. We can conclude that both common threshold adaptation and variable threshold adaptation schemes outperform verification schemes without threshold adaptation. However, the last statement is correct only for the cases of low or medium rate of imposters attack. For a high rate of imposters attack the use of threshold adaptation does not seem to be reasonable. The most significant result, obtained in the experiments for low and medium rate of imposters attack, is that performance for *variable verification threshold* adaptation is better than performance for *common verification threshold* adaptation. That means, additional *a posteriori* information, derived from words duration in the verified sentence, is useful for verification performance improvement. It is important to note, that the values of GPD step size ξ_b were different in the conducted experiments. It was found that the optimal value of the parameter ξ_b for different values of λ_i may be different: the larger value of λ_i, the smaller value of parameter ξ_b has to be chosen in order to obtain positive results of adaptation.

3. PARAMETER OPTIMIZATION FOR GPD ADAPTATION

As was shown in the described experiments with verification threshold adaptation, the quality of adaptation as well as parameters of adaptation strongly depend on the probability λ_i of the appearence of imposter requests. The main point here is how to evaluate the value of *a posteriori* parameter λ_i in order to use this information for the right evaluation of adaptation parameters. If GPD method is used for adaptation, we can adjust, for example, the step size ξ_b to trace properly the change of parameter λ_i. So, the larger is an estimated value of λ_i the smaller value has to be assigned to the parameter ξ_b. We can construct some monotonically decreasing function $\Psi(\lambda_i)$, such that $\xi_b = \Psi(\lambda_i)$.

Assume, p_i is the probability of false acceptance of an imposter request, and p_t is the probability of false rejection of a true talker re-

quest. From the latest history of the speaker verification system, consisiting of N requests, we assume there are N_a accepted requests and N_r rejected requests($N_a + N_r = N$). Among N_a accepted requests, some N_{a_i} requests may belong to accepted imposters and some N_{a_t} to the accepted true talker requests($N_{a_i} + N_{a_t} = N_a$). Similarly, we consider N_{r_i} and N_{r_t} for N_r rejected requests($N_{r_i} + N_{r_t} = N_r$). If the total number of *rejected imposters* requests is N_{r_i}, then the probability $P(N_{r_i}|N_r, \lambda_i)$ to obtain exactly N_{r_i} rejected imposter requests from the total number of N_r rejected requests is subjected to binomial distribution:

$$P(N_{r_i}|N_r, \lambda_i) = \binom{N_r}{N_{r_i}} \cdot \lambda_i^{N_{r_i}} \cdot (1-\lambda_i)^{N_r - N_{r_i}} \qquad (3.1)$$

But expression (3.1) treats the probability of appearence of N_{r_i} rejected imposters regardless of the discriminative properties of the speaker verification system. Such discriminative properties may be characterized by 2 parameters:

1) probability of false rejection of a true talker request p_t;
2) probability of false acceptance of an imposter talker request p_i.

So, rather than evaluate $P(N_{r_i}|N_r, \lambda_i)$ we have to evaluate $P(N_{r_i}|N_r, \lambda_i, p_t, p_i)$:

$$\begin{aligned} P(N_{r_i}|N_r, \lambda_i, p_t, p_i) &= \binom{N_r}{N_{r_i}} \cdot \lambda_i^{N_{r_i}} \cdot (1-\lambda_i)^{N_r - N_{r_i}} \\ &\cdot (1-p_i)^{N_{r_i}} \cdot p_t^{N_r - N_{r_i}} \end{aligned} \qquad (3.2)$$

Here, the term $(1-p_i)^{N_{r_i}}$ corresponds to the probability for N_{r_i} imposter requests to be rejected by the verification system, and $p_t^{N_r - N_{r_i}}$ corresponds to the probability for $N_r - N_{r_i}$ true talker requests to be rejected by the verification system.

Similar to (3.2) an expression may be derived for the probability $P(N_{a_i}|N_a, \lambda_i, p_t, p_i)$ to get exactly N_{a_i} *accepted imposter* requests among the total $N_a(N_a = N - N_r)$ number of accepted requests:

$$\begin{aligned} P(N_{a_i}|N_a, \lambda_i, p_t, p_i) &= \binom{N - N_r}{N_{a_i}} \cdot (1-\lambda_i)^{N - N_r - N_{a_i}} \\ &\cdot \lambda_i^{N_{a_i}} \cdot p_i^{N_{a_i}} \cdot (1-p_t)^{N - N_r - N_{a_i}} \end{aligned} \qquad (3.3)$$

Here, the term $p_i^{N_{a_i}}$ corresponds to the probability for N_{a_i} imposters to be accepted by verification system, and $(1-p_t)^{N - N_r - N_{a_i}}$ corresponds to the probability for $(N - N_r - N_{a_i})$ true talker requests to be accepted by the system. Eventually, the probability $P(N_r, N_a|N, \lambda_i, p_i, p_t)$ to get N_r rejected requests and N_a accepted ones from the total number of N requests in verification system may be evaluated as follows:

$$\begin{aligned} P(N_r, N_a|N, \lambda_i, p_i, p_t) &= \sum_{N_{r_i}=0}^{N_r} \sum_{N_{a_i}=0}^{N_a} (P(N_{r_i}|N_r, \lambda_i, p_i, p_t) \\ &\cdot P(N_{a_i}|N_a, \lambda_i, p_i, p_t)) \end{aligned} \qquad (3.4)$$

We use a maximum likelihood method to estimate the value of λ_i; i.e. our task is to find such an optimal parameter $\tilde{\lambda}_i$, that would maximize the probability $P(N_r, N_a|N, \lambda_i, p_i, p_t)$:

$$\tilde{\lambda}_i = \arg\max_{\lambda_i} P(N_r, N_a|N, \lambda_i, p_i, p_t) \qquad (3.5)$$

Due to high complexity of the derived expression for $P(N_r, N_a|N, \lambda_i, p_i, p_t)$ the most reasonable way to find a solutution of the equation (3.5) is just to use direct evaluation of $P(N_r, N_a|N, \lambda_i, p_i, p_t)$ for some discrete set $\Lambda_i = \{\lambda_i^{(m)}\}_{m=\overline{1,M}}$ of M possible discrete values of λ_i. Then we have to choose from the obtained M values such a value $\check{\lambda}_i$ that brings a maximum value for $P(N_r, N_a|N, \lambda_i, p_i, p_t)$. Obtained decision $\check{\lambda}_i$ may be substituted into function $\Psi(\lambda_i)$ to adjust parameters of the adaptation procedure.

4. SUMMARY

The objective of this study was to find effective ways to optimize adaptation for a speaker verification task by exploiting GPD as a representative of discriminative methods. Significant improvement was obtained by using GPD for HMMs adaptation. Equal-error rate was reduced from 4.40% to 2.17% after applying combined adaptation technique by using traditional simplified MAP adaptation, subsequent resegmentation of an adaptation sentence and final GPD adjustment of HMMs. technique for verification threshold adjustment. Derived formulas allow application of GPD method for estimation of an optimal separation boundary on the word level between the true and imposter talkers samples. The experiments, simulated different levels of imposters attack on the verification system, were conducted. The experiments proved, that for the adjustment of adaptation procedure parameters we need knowledge about the current rate of imposters attack. In the current study it was derived the corresponding formula that allows us to get an approximate estimate for the probability of an imposter requests appearance, thereby offering a numerical value, that may characterize the rate of imposters attack. Derived formula may be used to find optimal adaptation parameters under changing channel conditions.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. A. E. Rosenberg, Dr. S. Parthasarathy for their suggestions and support during the process of this research and Dr. C.H. Lee for his proposal to use combined adaptation.

5. REFERENCES

1. A. E. Rosenberg and J. DeLong, "HMM-Based verification using a telephone network database of connected digit utterances," *AT&T Bell Laboratories Technical Memorandum*, BL011226-931206-23TM, December 6, 1993.
2. B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Processing*, **40**(12), pp. 3043–3054, December 1992.
3. T. Matsui and S. Furui, "A study of speaker adaptation based on minimum classification error training," 4^{th} *European Conference on Speech Communication and Technology, Madrid,* pp. I-81-84, September, 1995
4. C.-H. Lee and J.L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters," *Proc. ICASSP, Minneapolis*, pp. II-558-561, 1993.