INCREMENTAL SPEAKER ADAPTATION WITH MINIMUM ERROR DISCRIMINATIVE TRAINING FOR SPEAKER IDENTIFICATION

C. Martín del Álamo(1), J. Álvarez(1), C. de la Torre(1), F.J. Poyatos(2), L. Hernández(3)

(1) Speech Technology Group. Telefónica I+D. c/ Emilio Vargas, 6. 28043 Madrid. Spain.
(2) Universidad Alfonso X El Sabio. Madrid. Spain.
(3) E.T.S.I. Telecomunicación. Univ. Politécnica de Madrid. Spain.
e-mail: cma@craso.tid.es

ABSTRACT

Minimum Classification Error (MCE) has shown to be effective in improving the performance of a speaker identification system [1]. However, there are still problems to solve, such as the variability of the voice characteristics of a particular speaker through time. In this work, we analyze the degradation of a GMM-based text-independent speaker identification system when using test data recorded over 6 months after the training session. And trying to avoid this degradation we study the use of supervised adaptation based on Maximum a Posteriori (MAP), and MCE. These techniques have been shown to provide good results for speaker adaptation in speech recognition.

The major result we have obtained is that by starting with GMM models trained with only speech from session 1, similar identification results can be obtained for all the other sessions using an incremental adaptation using only 2.5 seconds of speech per speaker and session as data for the MCE training adaptation procedure. We have also found that, in our extreme experimental setup, MAP becomes unhelpful when combined with MCE adaptation.

1. INTRODUCTION

Several Speaker Identification/Verification systems have been developed over the past few years. Most of them give good results when tested in laboratory conditions. However, when used in a real application they have to deal with many added problems. One of these problems (but not the only one) is the great variability shown by a speaker over a long period of time. That is, if the system is trained with an isolated recording session, although it performs very well with the same session, it is very likely to obtain much worse results as the time passes. One possible solution to this problem is to train the system with the voice of the speaker, recorded over a long period of time, for example a few months. But it is easy to see that this is not always possible, specially for a real application. We can imagine that a speaker that pays a service wants it to start working immediately, and not to waste months before the system is available to use.

Because of that all, a very common approach is to adapt the models to the speakers along the time, as the system is working in normal use, that is, on-line. As the speaker is making use of the system, the incoming voice is used to adapt the models.

In this paper we study two of the most promising techniques proposed for speaker adaptation in automatic speech recognition systems: Minimum Classification Error (MCE) training, and Maximum A Posteriori (MAP). Both of them have been shown to be effective not only as alternative training procedures to Maximum Likelihood (ML) training, but also as powerful adaptation techniques. The recognition scenario we will consider shows extreme conditions, in terms of distance between different recording sessions, in order to test the ability of the adaptation procedures.

The rest of the paper is organized as follows: in section 2, basic principles of MCE training are briefly described. Section 3 contains the basic MAP training formulation. Experimental results are presented in Section 4, and some conclusions are given in Section 5.

2. MCE TRAINING

Recently, different discriminative training algorithms have been proposed to perform a minimum classification error (MCE) training. In speaker recognition, discriminative training takes into account models from competing speakers providing a training criterion where the recognition errors of the training data are directly minimized [1]. In this work we will use a MCE training procedure based on the well formulated algorithm known as Generalized Probabilistic Descent (GPD), formulated to estimate HMM parameters [2]. However we will consider two major modifications, already proposed in [3], when using GPD for GMM-based speaker recognition systems. These modifications are:

1. The use of a misclassification measure based on an individual representation of competing speakers.
2. An empirical loss function will be included to control the training procedure. This produces a likelihood-based selection of correctly or incorrectly classified competing speakers providing different gradient weights for them.

The misclassification measure we use can be expressed as:

$$d_{kj}(X;\Lambda) = -g_k(X;\Lambda) + g_j(X;\Lambda) \qquad j \neq k \tag{1}$$

Where $g_j(X;\Lambda)$ is the score given by the j^{th} speaker's model for the observation X. So, not only 1 misclassification measure per utterance is calculated, but a set of them, one per each competing speaker. Then the empirical loss function $l(d_k)$ is applied to every of these misclassification measures.

$$l(d_k) = \begin{cases} = 0 & d_k < -Q_1 \\ = \dfrac{d_k^{\,2}}{2Q_1} + d_k + \dfrac{Q_1}{2} & -Q_1 < d_k < 0 \\ = d_k + 1 & d_k > 0 \end{cases} \quad (2)$$

Although this function may seem specially arbitrary it has been selected as is because of its first derivative (**Figure 1**), which is actually the function used in GPD.

$$\frac{\partial}{\partial d_k} l(d_k) = \begin{cases} = 0 & d_k < -Q_1 \\ = \dfrac{d_k}{Q_1} + 1 & -Q_1 < d_k < 0 \\ = 1 & d_k > 0 \end{cases} \quad (3)$$

Using such a loss function, all the errors are taken into account in training.

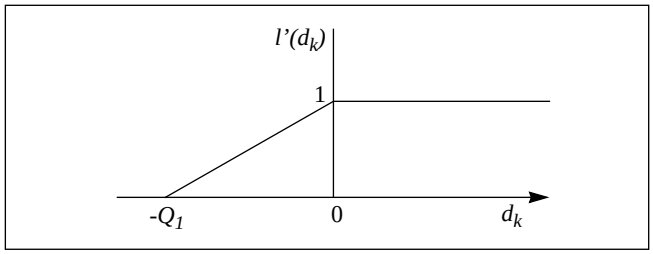

Figure 1: First derivative of empirical loss function.

Finally, parameters are updated according to the GPD formula:

$$\Lambda_{n+1} = \Lambda_n - \varepsilon_n \nabla(\sum l(d_k)) \quad (4)$$

The subscript n stands for the iteration number, Λ is the set of model parameters, ε_n is the adaptation step, and $\nabla(\cdot)$ represents the gradient of total loss function.

3. MAP TRAINING

Maximum A Posteriori (MAP) training provides a way of incorporating prior information of a previously trained system into the adaptation process. In our experimental conditions this should be particularly useful to deal with very distant recording sessions and sparse adaptation data. However, as stated in [4], unlike MCE, MAP estimation does not guarantee the highest performance for reducing the recognition errors. But, on the other hand, MAP avoids the classical problem of over-training presented in MCE.

The application of MAP estimation [5] to GMM is quite straight. In MAP estimation, based on the *a priori* density function *f(·)*, the parameter set λ_j of speaker *j* are obtained while $g_j(X;\lambda_j) f(\lambda_j)$ is locally maximized by the segmental MAP algorithm [5]. According to that, means and weights are updated as:

$$\hat{\mu}_{jk} = \frac{\tau_{jk}\mu_{jk} + \sum_{t=1}^{T} c_{jkt} x_t}{\tau_{jk} + \sum_{t=1}^{T} c_{jkt}} \quad (5)$$

$$\hat{w}_{jk} = \frac{\gamma_{jk} - 1 + \sum_{t=1}^{T} c_{jkt}}{\sum_{k=1}^{K} \left(\gamma_{jk} - 1 + \sum_{t=1}^{T} c_{jkt} \right)} \quad (6)$$

Where μ_{jk} and w_{jk} are the mean and weight for the k^{th} mixture of speaker *j*, c_{jkt} is the occupation probability of mixture *k* of speaker *j* at time *t* of a utterance. τ_{jk} are the initial coefficients which are used to control reliability of prior information relative to training data.

γ_{jk} initialization is done as:

$$\gamma_{jk} = \left(w_{jk} \sum_{k=1}^{K} \tau_{jk} \right) + 1 \quad (7)$$

and reestimation is performed according to:

$$\hat{\gamma}_{jk} = \gamma_{jk} + c_{jk} \quad (8)$$

Where the parameter γ_{jk} is reestimated in every iteration.

4. EXPERIMENTAL SETUP AND RESULTS

The database used in these experiments includes voice from 35 speakers, (20 male and 15 female), recorded in 4 different sessions over a period of 6 months. The distribution in time can be seen in **Figure 2**. This database has been extracted from CEUDEX Database [8]. The non-uniform time separation between sessions was chosen so as to have different extreme situations to test the variability of the voice characteristics of the speakers. Moreover we also designed different corpora for the different sessions: read sentences for session 1, isolated words for session 2, connected numbers for session 3 and sentences for session 4. The speakers were recorded in a quiet office environment, and the speech was parameterized to 10 LPC cepstral coefficients per 30 ms. frame every 10 ms. Each speaker was modeled by a Gaussian Mixture Model (GMM) with 32 mixtures.

Baseline system was trained with 80% of session 1, (which amounts approximately 35 seconds of speech per speaker) with Maximum Likelihood (ML). The baseline system recognition performance is shown in **Table 1**.

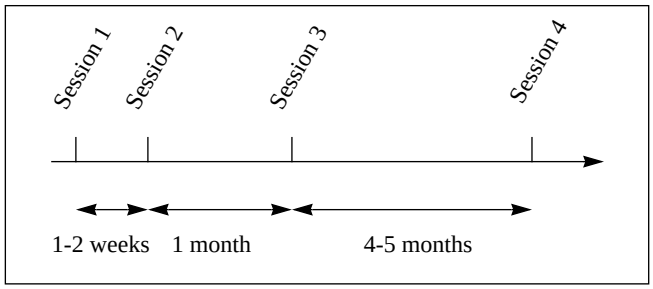

Figure 2: Distribution in time of recording sessions.

The voice from test utterances were extracted by an endpoint detector [6] to suppress silence and then split into 1 second segments, overlapped 0.5 seconds.

Test Set	Recognition Error Rate
20% of S1 not used in training	6.37%
Session 2	16.84%
Session 3	24.27%
Session 4	37.72%

Table 1: Baseline system performance.

It is clear from these results that, under the extreme test conditions, continuous adaptation to the speaker is needed. In order to test different adaptation techniques using only a small fraction of data from the different speakers, we selected 5 segments per session and speaker, that is, 2.5 seconds of speech from each speaker in each session. Adaptation to each session is always made from the GMM's obtained with session 1.

It is also important to remark that this is a supervised adaptation, in the sense that we have correctly labeled the segments we are using in training. This is equivalent to suppose that the system has a very good impostor rejection rate.

4.1. MCE Adaptation.

To apply MCE adaptation, 10 iterations of GPD training were made. In each iteration, a fixed number of 4 competing speakers was obtained from the segments selected to adapt, over all the speakers. Then, GPD training was made, using the modifications in misclassification measure and in loss function, presented above.

The results obtained are shown in **Table 2**. As can be seen, a great adaptation is achieved by using only 2.5 seconds of speech from each speaker. Although the rate for a distant session can not reach the rate for the training session, a great improvement is obtained.

The disadvantage of this technique is that data from all the speakers is needed to adapt, so the system can be adapted only if there is enough data from all the speakers. This may present a problem if a new speaker is added to the system, or if some speakers do not use the system frecuently enough.

Test Set	Baseline system	Adapted system
20% of Session 1 not used in training	6.37%	5.03%
Session 2	16.84%	12.40%
Session 3	24.27%	9.77%
Session 4	37.72%	20.65%

Table 2: MCE Adapted system performance.

4.2. MAP Adaptation.

Using MAP adaptation, we can see that a slight improvement is obtained in every session (see **Table 3**).

In this case, there is no need to use a smoothing technique such as Vector Field Smoothing [7], because we are using GMM's, so that all the data are used to retrain the whole model, and not only certain parts of it.

Test Set	Baseline system	Adapted system
20% of Session 1 not used in training	6.37%	6.08%
Session 2	16.84%	15.84%
Session 3	24.27%	19.84%
Session 4	37.72%	35.61%

Table 3: MAP Adapted system performance.

The advantage of MAP training is the simplicity and speed. Besides, only data from desired speaker is needed, so that we can retrain/adapt each model when enough data is available for it, no matter if there are utterances recorded from the others.

4.3. MAP+MCE Adaptation.

Some authors [4][7] have proposed the combination of MAP and MCE as a good method to adapt to different speakers in speech recognition tasks. We have tested it in our speaker identification task, by applying MCE to the models obtained from MAP adaptation. The results are shown in **Table 4**.

Test Set	Baseline system	Adapted system
20% of Session 1 not used in training	6.37%	5.69%
Session 2	16.84%	12.60%
Session 3	24.27%	9.73%
Session 4	37.72%	20.60%

Table 4: MAP+MCE Adapted system performance.

As we can see, unlike other results for speaker adaptation in speech recognition [4][7], our speaker recognition results show no important benefits when combining MAP and MCE techniques, compared to the results obtained only with MCE adaptation. Perhaps the high time separation between sessions and the small

quantity of adaptation data could be the reasons of the lack of improvement when using MAP.

5. CONCLUSION

It is clear that a speaker suffers from a great variability in his voice characteristics along the time. In our case, models that worked well for the same recording session for which they were trained with, perform much worse when testing with remote in time sessions, even though these other sessions were recorded in similar conditions. In our database, this is the case of most of the speakers, although there are some of them that still have good recognition rates even in fourth session, 6 months later. No special difference was found between male and female speakers in terms of degradation.

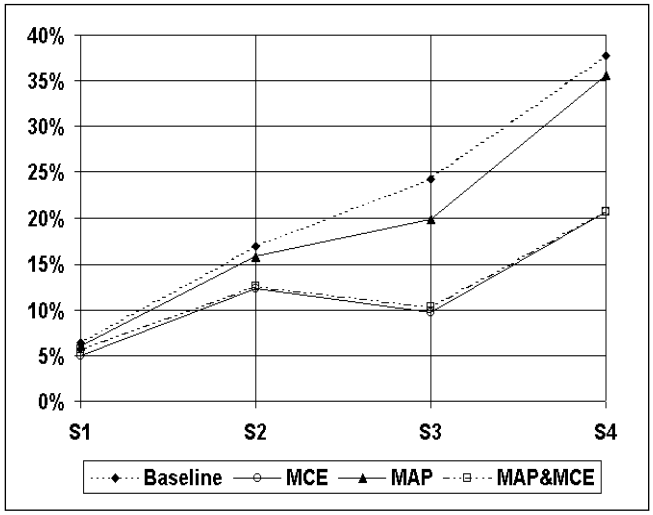

Figure 3: Summary of recognition error rates of adapted systems for different sessions.

However, it is shown that a great improvement in recognition rates can be achieved if speaker adaptation is applied. Among other methods of adaptation, MAP and MCE are becoming very popular.

We have found that MAP adaptation can obtain a significant reduction in recognition error (see **Figure 3**) in all the sessions, but it is important to remember that MAP does not guarantee any improvement in rates, because this method does not try to increment the discrimination capability. The points selected are those that give better results with the objective session, but preserve the rates with the previous sessions.

Results obtained with MCE adaptation are much better than baseline results (**Figure 3**), and also better than those obtained with MAP adaptation. This may be because MCE training focuses in reducing the recognition errors, instead of better modeling data. It is also found that the behavior is not homogeneous: some speakers improve more with MAP than with MCE, while some others do it in the opposite way.

When combining MAP and MCE adaptation, that is, applying MCE after having applied MAP (**Figure 3**), we found that the results are quite similar to those obtained when using only MCE adaptation. Although MCE is supposed to complement MAP adaptation, and to add its effects to MAP, is seems that MCE training eliminates much of the MAP training made before, predominating over it and making MAP adaptation unhelpful as a previous stage.

With regard to the recognition rates, it is important to note that the distribution among the different speakers appears very disperse. That is, while some speakers retain a moderate recognition error rate, others work much worse.

Anyway, none of both methods can obtain a recognition rate comparable to that of the training session, so we may think we need more data to adapt with, or more representative data, which is always difficult to establish.

6. REFERENCES

1. C-S. Liu et al., "A study on minimum error discriminative training for speaker recognition", *J. Acoust. Soc. Am. 97 (1), pp. 637-648, Jan. 1995.*
2. W. Chou et al., "A Minimum Error Rate Pattern Recognition Approach to Speech Recognition", *Intl. Journal of Pattern Recognition and Artificial Inteligence, Vol. 8, No. 1 (1994) 5-31.*
3. C. Martin del Alamo, F.J. Caminero Gil, C. de la Torre, L. Hernández Gómez, "Discriminative Training of GMM for Speaker Identification", *ICASSP'96. Atlanta, May 1996. Vol. I, pp. 89-92*
4. T. Matsui and S. Furui, "A study of speaker adaptation based on minimum error training", *in proc. EUROSPEECH'95, pp. 81-84.*
5. J.L. Gauvain and C-H Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observation of Markov chains", *IEEE Trans. on Speech and Audio Processing, Vol. 2, No. 2, pp. 291-298, 1994.*
6. A. Acero, C. Crespo, C. de la Torre, J.C. Torrecilla, "Robust HMM-Based Endpoint Detector", *EUROSPEECH'93, pp. 1551-1554.*
7. J. Takahashi, S. Sagayama, "Minimum Classification Error Training For A Small Amount of Data Enhanced by Vector-Field-Smoothed Bayesian Learning", *ICASSP'96. Atlanta. Vol. II, pp. 597-600.*
8. C. de la Torre et al., "CEUDEX: A Data Base Oriented to Context-Dependent Units Training in Spanish for Continuous Speech Recognition", *EUROSPEECH'95, pp. 845-848.*