MINIMUM CLASSIFICATION ERROR / EIGENVOICES TRAINING FOR SPEAKER IDENTIFICATION

Fabio Valente, Christian Wellekens

Institut Eurecom
Sophia-Antipolis, France
{*fabio.valente,christian.wellekens*}*@eurecom.fr*

ABSTRACT

This paper describes a new training approach based on two different techniques (Minimum Classification Error and eigenvoices) in order to achieve a better robustness when only poor training data is provided. In the first two sections of this paper we describe the MCE training and the eigenvoice approach. Then a unified MCE/eigenvoice training algorithm is proposed describing theoretical advantages. We compare the proposed method with classical ML/eigenvoice methods for a speaker identification task. The identification rate improvement is huge for sparse training data (up to 50% in the best case).

1. INTRODUCTION

It is well known in speech recognition that available training data is an important issue that can strongly affect performance. In fact when a large amount of data is provided, classical approaches like *Maximum Likelihood* methods yield quite satisfactory results. In [1], Reynolds obtains a speaker identification error rate of almost 0% using a gaussian mixture to model voice pdf and training those models with a Maximum Likelihood criterion. On the other side when a large amount of training data is not available, performances are strongly affected. We focused our attention on two different techniques that can improve robustness when only sparse data is available.

The first one is the eigenvoice method introduced in [2] for speaker adaptation purposes and used in [3] to perform speaker identification and verification. The eigenvoice approach consists in assuming that each speaker model can be represented in a reduced space (*the eigenspace*) as a linear combination of only a few eigenvectors (*the eigenvoices*). The weights of this combination are representative of a given speaker. As a consequence the data quantity needed to estimate speaker model is also reduced. In other words, a strong *a priori* knowledge is introduced i.e. the speaker belongs to a subspace of the space of all possible models. The price to pay, when using this technique is a model approximation that can be very far from reality. Until now, the eigenvoice framework was based on the *Maximum Likelihood* criterion.

It's well known that ML training is a suboptimal training approach that does not achieve good performance when training data is sparse. Many other training criteria were proposed and among them, the *Minimum Classification Error (MCE)* [4]. This method attempts to minimize directly the recognition error using parameters from all competing classes. In fact while ML strives to determine the best achievable model, MCE strives to optimize the classification task of all possible models.

Theoretically the use of MCE instead of ML in determining eigenvoices can improve results because MCE criterion is less affected than ML criterion by poor model approximation. So an algorithm that uses MCE training and eigenvoice technique should be more robust to lack of data (because of eigenvoices) and more robust to poor model approximation (because of MCE) typically caused by the construction of the eigenspace. We can figure out that the classic ML/eigenvoice approach aims at estimating a model $\Lambda_{eigenML}$ that is an approximation of a ML model Λ_{ML} that is itself an approximation of the speech signal. On the contrary the model constructed by MCE/eigenvoice, let say $\Lambda_{eigenMCE}$, aims directly at maximizing the classification task independently from the eigenspace construction and dimension (parameters that typically affects ML strongly). In section 2 we briefly review eigenvoices and in section 3 minimum classification error training, then in section 4 we describe our approach and in section 5, we illustrate some experiments.

2. EIGENVOICES APPROACH

The eigenvoice approach is based on space reduction methods applied to the vectorial space generated by the mean vectors of HMM output probability density functions.

The goal is to represent the speaker mean vectors $\hat{\mu}$ set as a linear combination of $[e(1), e(2), .., e(K)]$ eigenvoices i.e.

$$\hat{\mu} = \left[\hat{\mu}_1^{(1)}\, \hat{\mu}_2^{(1)} \ldots \hat{\mu}_m^{(s)} \ldots\right]^T = \sum_{j=1}^{K} w(j)e(j) \qquad (1)$$

So the representation of the speech (or the speaker) in the eigenspace is given by his K coefficients $w(i)$. Eigenvoices are computed using a large number of speaker dependent models and applying a space reduction technique like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) (see [2]). Once the eigenspace is defined, the problem is to find the speaker representation given some speech utterances $O = \{o_1, ..., o_t\}$. The problem was first solved by Nguyen in [5] proposing a *Maximum Likelihood Eigen Decomposition (MLED)*, but other alternative techniques were proposed (see [6]).

Another interesting point is the adaptation of the eigenspace to the speaker set in order to obtain a *Maximum Likelihood Eigen Space (MLES)* described in [7].

The performance of eigenvoice techniques strongly depends on the quality of the eigenspace. The choice of the eigenspace dimension K is a crucial point in the system design. Another main point is the way in which the eigenspace is constructed i.e. the

space reduction technique used (PCA,LDA,...), and the representativity of the speaker dependent models (see [2]).

An application of this method to speaker identification and verification can be found in [3]. Results from this paper show that when rich enrollment data is provided, classical ML/GMM methods outperform eigenvoice approach. In the case of sparse training data, we have the opposite situation because in this case the strong *a priori* constraints play a fundamental role.

Two different methods for identifying a speaker using eigenvoices are proposed:

1. Find the test speaker coordinates in the eigenspace using MLED, then find the distance between test speaker coordinates and client coordinates- this technique was named *eigendistance decoding*
2. Use a speaker model (in this case GMM) generated from client points in eigenspace to calculate the likelihood of test data- this technique was named *eigenGMM decoding*.

3. MINIMUM CLASSIFICATION ERROR

The MCE training is an alternative training approach that aims at minimizing directly the recognition error (see [4]). Classical ML training aims at realizing a model that represent the speech observations in the best sensible way. The problem is that this model is just an approximated one because the exact mathematical expression for the speech signal is unknown. The classification performance is based on the assumption that this model is accurate enough to distinguish between two different speech models. MCE takes advantages of the fact that the model can not be arbitrary accurate in some situations and tries to maximize the distance between all competing classes in order to facilitate the recognition task. Let's consider the main MCE training step.

Let $g_i(X,\Lambda)$, $i=1,...,M$ be a set of *class conditional likelihood functions*, where $\Lambda=\{\lambda^{(0)},...,\lambda^{(i)}\}$ is a parameter set. The decision rule C that associates a class C_i to an observation X is thus the following one:

$$C(X)=C_i, \text{ if } g_i(X;\Lambda)=max_j g_j(X;\Lambda). \quad (2)$$

Then a *class misclassification measure* is introduced:

$$d_i(X)=-g_i(X;\Lambda)+\log\left\{\frac{1}{M-1}\sum_{j,j\neq i} exp[g_j(X;\Lambda)\eta]\right\}^{1/\eta} \quad (3)$$

where η is a positive number. The misclassification measure is embedded into smoothed zero-one function $l_i(X;\Lambda)=l(d_i(X))$ where l is a sigmoid function of this type: $l(d)=1/(1+exp(-\gamma d+\theta))$. θ is normally set to zero and γ is a parameter that can make the convergence of the optimization method faster. γ is set to >1. The classifier performance is measured by the following function:

$$l(X;\Lambda)=\sum_{i=1}^{M} l_i(X;\Lambda)1(X\in C_i) \quad (4)$$

Equation (4) is minimized using a gradient like technique called *Gradient Probabilistic Descent (GPD)* method described in [4]. GPD is based on the idea that it is possible to achieve a local minimum of a loss function, using an iterative procedure based on individual loss i.e. the parameter update formula is:

$$\Lambda_{t+1}=\Lambda_t-\epsilon_t U_t \nabla l(X_t,\Lambda)|_{\Lambda=\Lambda_t} \quad (5)$$

where X_t is the training vector at time t and it can be demonstrated that when $t\to\infty$, a local minimum is achieved.

4. MCE/EIGENVOICES TRAINING

Gaussian mixture model (GMM) is a very efficient technique for speaker identification tasks (see [1]). Let's assume that the speech pdf can be modeled using a GMM i.e. speaker pdf is assumed to be of the form $p(o)=\sum_{m=1}^{M} c_m N(\mu_m,C_m)$ where o is an acoustic features vector and μ_m, C_m, c_m are respectively the mean vector, the covariance matrix and the weight of each multivariate gaussian $N(\mu_m,C_m)$. The approach we propose in this paper is based on the use of the MCE training, assuming that the parameter set Λ is constituted by $w_j^{(i)}$ with $i=0,...,K$ and $j=0,...,E$ i.e. the j-th weight of the i-th class referred to the j-th eigenvoice. In other words the class conditional likelihood function can be written as:

$$g_i(X;\Lambda)=P(X|\lambda^{(i)})=P(X|w_0^{(i)},...,w_E^{(i)}) \quad (6)$$

and equation (5) becomes:

$$w_j^{(i)}(t+1)=w_j^{(i)}(t)-\epsilon_t\frac{\partial l_i(X_n;\Lambda)}{\partial w_j^{(i)}} \quad (7)$$

$$\frac{\partial l_i}{\partial w_j^{(i)}}=\frac{\partial l_i}{\partial d_i}\frac{\partial d_i}{\partial w_j^{(i)}} \quad (8)$$

$$\frac{\partial l_i}{\partial d_i}=\gamma\, l_i(d_i)(1-l_i(d_i)) \quad (9)$$

$$\frac{\partial d_i}{\partial w_j^{(i)}}=\sum_m c_m(-e_j^m C_m^{-1}X_n+\mu_{im}C_m^{-1}e_j^m)N(\mu_{im},C_m) \quad (10)$$

where $\mu_{im}=\sum_p w_p^{(i)}e_p^m$ is the new estimated gaussian mean vector, assuming that e_p^m is the pth eigenvoice component referred to the mth gaussian distribution and $N(\mu_{im},C_m)$ is a gaussian distribution with mean μ_{im} and covariance matrix C_m. Thanks to eigenvoices, this approach should be more robust when poor training data is provided. Furthermore, because of MCE, modeling problems typically related to eigenvoices should have a reduced impact on recognition performance.

5. EXPERIMENTS

To compare MCE and ML eigenvoice training we carried out speaker identification experiments on the TIMIT database. Acoustic vectors consist in 16 MFCC (mel-cepestral coefficients). Each speaker is modeled using a 20 components GMM.

The eigenspace is built using the 462 speakers contained in the train set: 462 speaker independent models were generated using a Baum-Welch algorithm and then a PCA was applied to their mean vectors matrix as described in [6]. Experiments were run on subset of 11 speakers randomly selected (i.e. the dimension of the smallest TIMIT directory).

We studied the identification score of MCE and ML in relation with two parameters: quantity of data and eigenspace dimension using the eigenGMM decoding technique. For the first parameter,

we took inspiration from [1] where 8 of the 10 sentences provided for each test speaker were used for enrollment data and 2 for identification. In order to simulate a growing quantity of available data we achieved 3 sets of experiments with 2,5,7 training sentences with respectively 1,2,3 test sentences. On the other side we studied as well the performance for an eigenspace of dimension 10, 30 and 50; using more than 50 eigenvoices there's no significant improvement. The ML/eigenvoice training was achieved using a conjugate gradient algorithm while the MCE/eigenvoice training was done using a GPD algorithm with $\theta = 0$, $\gamma = 2$ and $\epsilon_t = \epsilon_0(1 - t/T)$. To reduce computational charge, eq (3) has been computed using the 3 best competitors instead of all competing classes following the approach described in [8]. We decided not to use GPD for ML training because the form of the function to minimize is easier than in the case of MCE and it is still possible to apply a classical gradient method instead of a stochastic gradient approximation. Results taken from [9] are illustrated in tables 1 and 2. Table 3 shows the relative gain of the MCE approach on the ML approach.

Error rate decreases when eigenspace dimension increases or data quantity increases for both training criteria.

Anyway MCE outperforms ML almost for all parameters values but the gain is not always huge. When a small quantity of training data is used, the MCE improves greatly the identification score (up to 50%). On the other hand when the model obtained using the ML training is good enough, there is no improvement at all. In fact in the case where we have a large amount of training data, and an eigenspace of dimension 30 (or 50) the ML/eigenvoice models is very precise (only 1.13% of error rate) and even applying MCE training there is no improvement.

It is interesting to notice that the identification rate of 1.13% represents the upper bound for both MCE and ML imposed from the space reduction; to increase the score, the system needs some information that cannot be represented in the eigenspace.

training/test sentences	10 eigenvoices	30 eigenvoices	50 eigenvoices
2-1	56.81%	23.86%	13.63%
5-2	36.36%	4.54%	2.27%
7-3	30.68%	1.13%	1.13%

Table 1. Speaker identification error rate using ML/eigenvoice training

training/test sentences	10 eigenvoices	30 eigenvoices	50 eigenvoices
2-1	48.86%	19.31%	12.49%
5-2	24.99%	3.4%	1.13%
7-3	22.72%	1.13%	1.13%

Table 2. Speaker identification error rate using MCE/eigenvoice training

5.1. MCE and ML inter-class distance

Fig. 1 is a plot of the position of 11 speakers in a bi-dimensional eigenspace after ML and MCE training. It is easy to notice that the MCE trained models have a bigger inter-class distance than the ML trained models and as consequence the recognition task is favorized.[1]

training/test sentences	10 eigenvoices	30 eigenvoices	50 eigenvoices
2-1	13.9%	19.04%	8.3%
5-2	31.25%	25%	50%
7-3	25.92%	0%	0%

Table 3. MCE/eigenvoice training relative gain on ML/eigenvoice training

Let's define the distance between two speakers as $d_{ij} = ||\overline{w^i} - \overline{w^j}||$ where the norm is the euclidean norm and $\overline{w^i}$, $\overline{w^j}$ are coefficient vectors that characterize the ith and jth speaker. To quantify the increased inter-speaker distance, we compute $D = E[(d_{MCE} - d_{ML})/d_{ML}]$ where d_{MCE} and d_{ML} are average inter-speaker distance computed on competing models obtained using MCE and ML training. The value of D we found for our tests set is 0.261 i.e. after MCE training the distance between all the competing speakers is increased of 26.1% compared to the previous ML trained models.

This distance is a physical distance in the eigenspace that not necessarily measure the error rate reduction; to quantify the effect of MCE training in the recognition task the KL-distance between different models can be used. Let's define

$$D_q = \int p(x|q) log \frac{p(x|q)}{p(x)} \, dx = D_q(p(x|q)||p(x)) \qquad (11)$$

where $p(x|q)$ is the likelihood of x given the model q and $p(x) = \sum_r p(r)p(x|r)$. The discriminability of a set of models can be defined as:

$$D = \sum_q p(q) D_q(p(x|q)||p(x)) \qquad (12)$$

To compute D, we used a numerical integration method as described in [10] (appendix C) using as models for $p(x|q)$ the 20 components GMM obtained using the ML and the MCE training. Before doing this computation we applied a space reduction transformation based on the KL-transform in order to reduce the computational charge of this task and we reduced the integration domain to $[min_i(\mu_{i,d} - 3\sigma_{i,d}), max_j(\mu_{j,d} + 3\sigma_{j,d})]$ where $\mu_{i,d}$ and $\sigma_{i,d}$ are the mean and the variance of the component i for dimension d. The average value of the gain $(D_{MCE} - D_{ML})/D_{ML}$ computed on our test set is 0.036. Obviously a significant gain of one of those distance measure does not correspond to a significant gain in the identification rate because the original ML models can be very accurate and the MCE training cannot improve them.

6. CONCLUSION AND FUTURE WORKS

Even if results are interesting we must outline two problems that we found using this approach and that may explain the gain differences for different parameters values.

First of all, there exists a proof that the GPD algorithm achieves a local minimum of the loss function when $N \to \infty$ where N is the number of observation vectors. In reality ∞ means large

[1] A demo on the effect of MCE training on speaker models can be found at this URL: *ecwww.eurecom.fr/~valentef/mce.html*

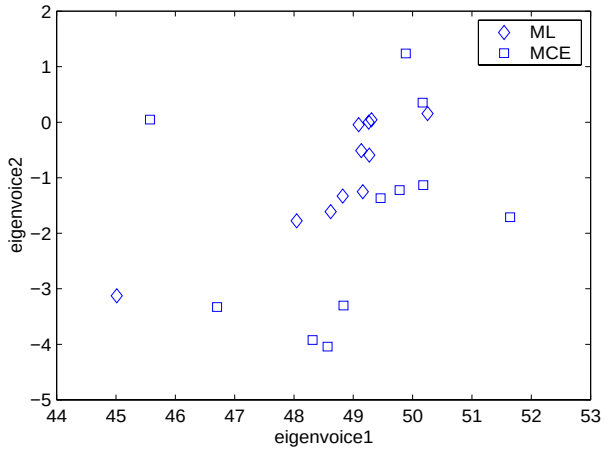

Fig. 1. Difference between ML and MCE training in a bi-dimensional eigenspace

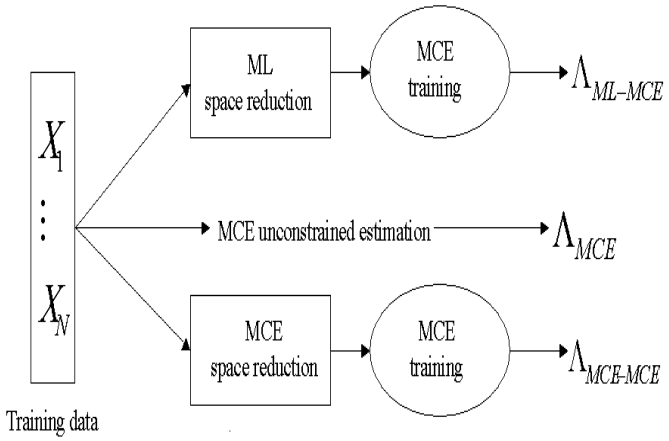

Fig. 2. MCE training and space reduction

amount of data but when only little training data is provided the stochastic approximation is no more valid, there's no guarantee of convergence; the GPD is not the best optimization technique when only sparse training data is available.

The second problem concerns the construction of the eigenspace. In previous sections we said that using MCE training is more robust than ML to poor eigenspace construction and experimental results seems to confirm our impression. Anyway there is no guaranty that the MCE optimal configuration is contained in the eigenspace. The problem is analogous to the MLES [7] which describes eigenvectors adaptation to build a ML eigenspace. Basically what we obtain with this approach is the projection of the MCE minimum in a ML eigenspace that can be really different from the MCE minimum in a MCE eigenspace or in an unconstrained space. This issue is described in fig. 2: starting from a training data set, MCE training can be used to obtain a Λ_{MCE} model (middle pattern). This task is not always achievable because of lack of data; for this reason a space reduction technique is applied before training the system with a MCE technique. In the upper pattern the space reduction is a space reduction optimal in the sense of ML that produces a model Λ_{ML-MCE} which can be very far from the ideal model Λ_{MCE}. For this reason the space reduction should be optimal from the MCE criterion point of view (lower pattern) and the model $\Lambda_{MCE-MCE}$ should be closer to the ideal model. The MCE space reduction problem has already been discussed in the pattern recognition framework. In [11], an algorithm based on GPD method called *Minimum Error Learning Subspace* which enables to directly pursue the minimum error recognition is described. In [12] a comparison between MCE subspace and LDA subspace is done.

Those algorithms work directly on the projection matrix that allows the space reduction. As outlined in [5], it is not always efficient to work on the projection matrix in speech recognition problems so future works should focus on the possibility of performing an alternative space reduction technique based on MCE.

7. REFERENCES

[1] Reynolds D.A., "Speaker identification and verification using gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[2] Kuhn R.; Junqua J.-C.; Nguyen P.; Niedzielski N., "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 695–707, Nov 2000.

[3] Thyes O.; Kuhn R.; Nguyen P.; Junqua J.-C.;, "Speaker identification and verification using eigenvoice," *ICSLP 2000, Beijing-China*, vol. 2, pp. 242–246, October 2000.

[4] Juang B.H.; Hou W.; Lee C.H., "Minimum classification error rate methods for speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.

[5] Nguyen P., "Fast speaker adaptation," M.S. thesis, Institut Eurecom, 1998.

[6] Westwood R., "Speaker adaptation using eigenvoice," M.S. thesis, Department of Engineering, Cambridge University, 1999.

[7] Nguyen P.; Wellekens C.; Junqua J.-C., "Maximum likelihood eigenspace and mllr for speech recognition in noisy enviroments," *Eurospeech, Budapest, Hungary*, vol. VI, pp. 2519–2522, 1999.

[8] Chou W.; Juang B.H.; Lee C.H., "Minimum error rate training based on n-best string models," *ICASSP-93*, vol. 2, pp. 652–655, 1993.

[9] Valente F., "Mce/eigenvoice training for speaker identification," M.S. thesis, UNSA, doctoral school STIC, DEA SIC image vision, 2002.

[10] Bilmes J., *Natural Statistical Models for Automatic Speech Recognition*, Ph.D. thesis, ICSI, Berkeley, 1999.

[11] Watanabe H.; Katagiri S., "Discriminative subspace method for minimum error pattern recognition," *IEEE NNSP*, vol. 5, pp. 77–86, 1995.

[12] Wang X.; Paliwal K.K., "Using minimum classification error training in dimensionality reduction," *NNSP X, 2000.*, vol. 1, pp. 338–345, 2000.