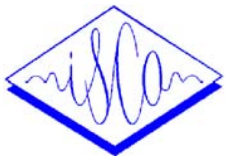

EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

USING SIMULATED ANNEALING EXPECTATION MAXIMIZATION ALGORITHM FOR HIDDEN MARKOV MODEL PARAMETERS ESTIMATION

J. Simonin & C. Mokbel

France Télécom - CNET - LAA/TSS/RCP
Technopole Anticipa
2, Avenue Pierre Marzin, 22307 Lannion - FRANCE
e-mail: simonin@lannion.cnet.fr

ABSTRACT

This paper presents the use of a simulated annealing technique during the parameters estimation of a Hidden Markov Model (HMM) in a speech recognition system. This technique allows to move out of a local optimum which characterizes a classical Expectation Maximization (EM) algorithm, and thus to achieve a better estimation with a limited amount of training data.

We choose here the Simulated Annealing Expectation Maximization (SAEM) algorithm introducing a simulated annealing technique in the EM method. The SAEM algorithm is compared to the classical EM algorithm, for both task-independent and task-dependent Viterbi training. The evaluation leads to significant improvement of recognition performances.

1. INTRODUCTION

In a speech recognition system, the acoustic model classically used is HMM-based. In order to achieve a reliable estimation of the HMM parameters, according to the Maximum Likelihood (ML) criterion, the EM algorithm is the most popular method used [1].

One of the EM algorithm's drawbacks is that it converges to a local optimum, especially if little amount of training data is available. An other solution is the use of the well-known simulated annealing technique to escape from local optima [2]. Simulated annealing is a randomized perturbation technique which enables to move out of a local optimum.

This technique may optimize the HMM structure, when applied to the parameter ties or the clustering of phonetic contexts in a speech recognition system [3]. Nevertheless this optimization concerns discrete parameters.

Another kind of parameters perturbation is a Gaussian density splitting, which consists in slightly perturbing the initial parameter set [4]. This perturbation is performed at the beginning of the training phase and concerns the Gaussian density mean vector.

In this paper, we choose a simulated annealing technique which may lead to improve the ML estimation algorithm. We adapt the SAEM algorithm [5] which uses a simulated annealing technique in a classical EM iterative procedure for the parameters estimation of a Continuous Mixture Densities HMM.

The first part of the paper describes the SAEM algorithm which includes a simulated annealing technique in each EM algorithm iteration. SAEM technique is described here in the specific matter of speech recognition system parameters estimation. The SAEM algorithm is adapted to the Viterbi parameters estimation.

In the second part, speech recognition experiments enable a comparison between the classical EM algorithm and the SAEM algorithm with task-independent or task-dependent training. Then we evaluate this technique with a limited amount of data used during a task-dependent training, where this kind of perturbation is most appropriate.

2. SAEM ALGORITHM

SAEM is adapted here to a classical HMM parameters estimation. This simulated annealing technique adds two steps to the classical EM.

2.1. HMM Parameters Estimation

In the proposed approach the considered model parameters $\lambda = (A, B, \pi)$ are given by:

- N, the number of states in the HMM and q_t, the occupied state at time t.

- A, the state transition probability matrix such as:

$$A = \{a_{ij}\} \text{ where } a_{ij} = Pr(q_{t+1} = j | q_t = i)$$

with $1 \le i, j \le N$ and $1 \le t \le T$.

- B, the set of sub-processes observation distributions, chosen to be a mixture of Gaussian with diagonal covariance matrices, given by:

$$B = \{b_i(o_t)\}$$

$$b_i(o_t) = p(o_t|q_t = i) = \sum_{1\le k\le NG} c_{ik}.N(o_t; \mu_{ik}, \Sigma_{ik})$$

which may have the following approximation:

$$b_i(o_t) = \max_{1\le k\le NG} \{c_{ik}.N(o_t; \mu_{ik}, \Sigma_{ik})\}$$

with $1\le i\le N$ and $1\le t\le T$,

and where $N(.;\mu_{ik}, \Sigma_{ik})$ is a Gaussian density with μ_{ik}, the mean vector and Σ_{ik}, the diagonal covariance matrix and where c_{ik} is the Gaussian component weight. NG is the number of Gaussian components of the Gaussian distribution mixture.

- π, the initial state distribution:

$$\pi = (\pi_i) \text{ et } \pi_i = Pr(q_0 = i)$$

with $1\le i\le N$.

For a sake of simplicity, we consider here only the case of mono-Gaussian density. The parameters re-estimation procedure leads to the following formulas (EM algorithm), where α and β are respectively the forward and the backward variables [6]:

$$\pi_i' = \frac{\alpha_0(i)\beta_0(i)}{\sum_{j=1}^{N} \alpha_T(j)} = \gamma_0(i)$$

$$a_{ij}' = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i)a_{ij}b_j(o_t)\beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)\beta_{t-1}(i)} = \frac{\sum_{t=1}^{T} \xi_{t-1}(i,j)}{\sum_{t=1}^{T} \gamma_{t-1}(i)}$$

$$\mu_i' = \frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i).o_t}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)} = \frac{\sum_{t=1}^{T} \gamma_t(i).o_t}{\sum_{t=1}^{T} \gamma_t(i)}$$

$$\Sigma_i' = \frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i).[o_t-\mu_i].[o_t-\mu_i]^T}{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)}$$

$$= \frac{\sum_{t=1}^{T} \gamma_t(i).[o_t-\mu_i].[o_t-\mu_i]^T}{\sum_{t=1}^{T} \gamma_t(i)}$$

with $1\le i, j\le N$ and $1\le t\le T$,

and where:

$$\alpha_t(i) = P(o_1,...,o_t, q_t = i / \lambda)$$

$$\beta_t(i) = P(o_{t+1},...,o_T / q_t = i, \lambda)$$

$$\xi_t(i,j) = P(q_t=i,q_{t+1}=j / O,\lambda)$$

$$\gamma_t(i)=P(q_t=i / O,\lambda)$$

It is to be noticed that $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$.

The problem is the optimum locality within the EM algorithm. In order to overcome this problem a simulated annealing technique, the SAEM algorithm, is introduced during the speech recognition model training.

2.2. SAEM Description

Between the step E, corresponding to the estimation of the auxiliary function, i.e. the conditional expectation of the logarithm of γ_t, and the step M, relative to the maximization of the auxiliary function, appearing in the classical EM algorithm, two steps are inserted.

These two steps, called S and A, are simulated annealing steps. The step S means a randomized perturbation of the estimated parameters for each observation. The step A enables to take into account these simulated annealing evaluations for the new estimation of the parameters. These steps applied to an HMM parameters training lead to the following A and S steps description.

During step S and for each couple of observation vectors (o_t,o_{t+1}) a random variable $r_t(i,j)$ is realized for each transition (i,j) in the model. $r_t(i,j)$ follows a multinomial distribution of order 1 with parameter $\xi_t(i,j)$. In step A, the probabilities of the association between the elements (o_t,o_{t+1}) and the states (i,j) are modified using the variable $r_t(i,j)$:

$$\tilde{\xi}_t(i,j) = \xi_t(i,j) + \rho_n [r_t(i,j) - \xi_t(i,j)]$$

where ρ_n is a decreasing temperature, n being the iteration number. Looking the preceding equation, it can be found that the expectation of $\tilde{\xi}$ is equal to the expectation of ξ, and its variance is equal to $\rho_n^2(\xi(1-\xi))$. Thus, the effects of the perturbations decrease with the iterations (since ρ_n decreases with the number n of iterations). Besides, the perturbation is more important for ξ close to 1/2.

The re-estimation formulas remain the same as for the EM algorithm, replacing the γ and ξ by their new estimated values.

We assume the SAEM algorithm convergence with a Continuous Mixture Densities HMM system as it is demonstrated in [2]. This convergence needs specific simulated annealing parameters values. That means an

initial perturbation amplitude and a temperature decreasing speed superior to 0 and below 1.

2.3. SAEM Application for Viterbi Estimation

During the HMM parameters estimation with a Viterbi algorithm, the previous re-estimation formulas are rewritten as described below.

With Viterbi training $\xi \in \{0,1\}$. For this reason, the parameter of the multinomial distribution of $r_t(i,j)$ variable should be chosen a priori. Here, we choose to fix this parameter to 1/2 corresponding to the maximum perturbation in the exact SAEM.

For each frame, during the n^{th} iteration, the step S is defined as:

If the drawing of lots, which is here a simple toss, is positive:

$$\tilde{\xi}_t(i,j) = (1 - \rho_n)\, \xi_t(i,j) \text{ for } \xi = 1 \text{ only,}$$

else

$$\tilde{\xi}_t(i,j) = \xi_t(i,j) \qquad \text{for } \xi = 1 \text{ only.}$$

In this case, the expectation of $\tilde{\xi}$ is equal to:

$$E[\tilde{\xi}_t(i,j)] = \xi_t(i,j) - \frac{\rho_n}{2}\xi_t(i,j) = 1 - \frac{\rho_n}{2}$$

and its variance

$$E[(\tilde{\xi}_t(i,j) - \xi_t(i,j))^2] = \frac{\rho_n^2}{2}(\xi_t(i,j))^2 = \frac{\rho_n^2}{2}$$

The preceding equations show that the perturbations decrease with the number of iterations. It is to be noticed that, for the optimal path, since $\xi \in \{0,1\}$:

$$\tilde{\gamma}_t(i) = \sum_{j=1}^{N} \tilde{\xi}_t(i,j) = \tilde{\xi}_t(i,j_0) = 1$$

j_0 being the state for which $\tilde{\xi}_t(i,j) = 1$. Then, the same perturbations affect directly the $\gamma_t(i)$ parameters, and the resulting $\tilde{\gamma}_t(i)$ are directly replaced in the re-estimation formulas.

The parameters perturbation advantage are illustrated with a HMM trained with a classical EM compared to the SAEM adaptation described above.

3. EXPERIMENTS

The algorithm is evaluated in a speech recognition system on three telephone databases. Experiments are performed with two different training; a task-independent training and a task-dependent training. These estimations follow a Viterbi algorithm.

3.1. Training and Test Databases

The training corpus for the task-independent system and for the initialization of every task-dependent system is made of about 700 short sentences recorded by hundred of speakers calling from different regions of France. This telephone database contains almost all the French diphones.

For evaluations and task-dependent system training, three isolated words telephone databases recorded by 800 speakers are used. Table 1 presents the records amount of these databases in the training step and in the test step.

Table 1: Characteristics of databases.

Database	Training Records	Tests Records
Digits (*0 to 9*)	3555	3622
Numbers (*00 to 99*)	7304	7288
Trégor (*36 words*)	12719	12842

3.2. Comparison Between SAEM and EM

Recognition performances relative to a SAEM Viterbi training are compared to a classical EM Viterbi training performances first with a task-independent training then with a task-dependent training.

Figure 1 presents error rates for the three corpora with a task-independent training in terms of error rates.

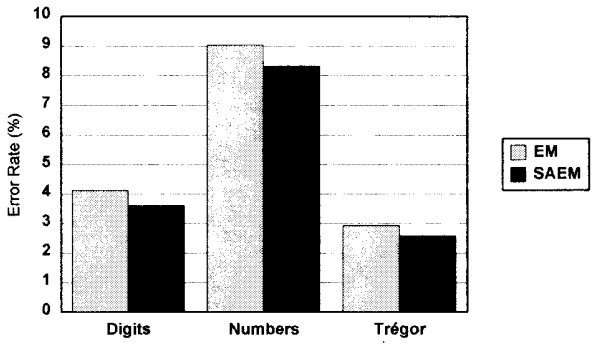

Figure 1: EM / SAEM comparison for task-independent system.

In the case of a task-independent system, the estimation of the HMM parameters is less reliable because a lack of appropriate data. In the optimal case, the use of the adapted SAEM yields to an error reduction from of 8% for the Numbers and of 12% for the Digits and Trégor compared with the results obtained with a classical EM training. This reduction is significant with respect to a 95% confident interval.

Figure 2 presents evaluation results with a task-dependent training.

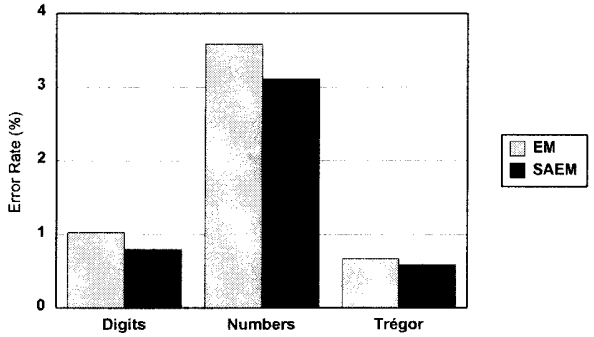

Figure 2: EM / SAEM comparison for task-dependent system.

In a task-dependent system, the corpus used during the training depends on the vocabulary application. The main result is that the error rate reduction is 12% for the Trégor, 13% for the Numbers and 22% for the Digits compared to the model estimated with a classical EM algorithm. The reduction with task-dependent training data is not significant for the Trégor evaluation (95% confident interval).

Results show better recognition performances for SAEM Viterbi training compared to a classical EM Viterbi training, whatever the corpus or the task dependence during the training may be. These improvements are more significant with a task-independent training than with a task-dependent training.

3.3. Limited Amount of Data

Simulated annealing technique is also evaluated with a limited amount of data. The following experiments involve two telephone databases which are continuous speech corpus described in Table 2 and in Table 3. These corpora consist of answers about date and time of the recording.

Table 2: Characteristics of continuous speech training corpus.

Database	Training Records
Dates (175 words)	1256 (5474 words)
Hours (215 words)	1280 (5391 words)

Models are evaluated with speaker-independent and speaker-dependent recognition tests.

Table 3: Characteristics of continuous speech test corpus.

Database	SI Tests Records	SD Tests Records
Dates (*175 words*)	275 (1175 words)	1330 (5782 words)
Hours (*215 words*)	267 (1170 words)	1334 (5430 words)

Speaker-independent and speaker-dependent recognition performances are reported in Figure 3. Relative rate of the adapted SAEM training error rate compared to the classical EM training error rate are described with these models.

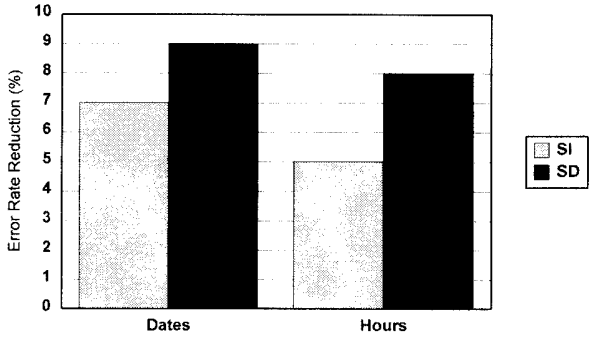

Figure 3: EM / SAEM error rate reduction for a limited amount of training data with speaker-independent recognition and speaker-dependent recognition.

For these corpus, the evaluations show significant error rate reduction respect to a 95% confident interval compared to a classical EM.

4. CONCLUSION

In this paper it is suggested to use SAEM algorithm in order to train HMM parameters. We adapt it to a Viterbi training in a speech recognition system.

This adaptation leads to an improvement of recognition performances compared to a classical EM approach in task-independent and in task-dependent training. Moreover, in the case of a limited amount of training data, using a perturbation technique leads to a decrease in the error rate.

5. REFERENCES

1. A.P. Dempster, N.M. Laird & D.B. Rubin, "Maximum Likelihood from incomplete data via the EM algorithm", *JRSS*, B39: 1-38, 1977.
2. S. Kirkpatrick, C.D. Gelatt & M.P. Vecchi, "Optimization by simulated annealing", *Science*, Vol. 220: 671-680, 1983.
3. R. De Mori, M. Galler & F. Brugnara, "Search and learning strategies for improving hidden Markov models", Computer Speech and Language, Vol. 9: 107-121, 1995.
4. J. Simonin, S. Bodin, D. Jouvet & K. Bartkova, "Parameter tying for flexible speech recognition", *Proc. of ICSLP*, Vol. 2: 1089-1092, 1996.
5. G. Celeux & J. Diebolt, "A simulated annealing type EM algorithm", *Rapport de recherche INRIA*, N°1123: 1-24, 1989.
6. L. Rabiner & B.H. Juang, "Fundamentals of speech recognition", *Prentice Hall Signal Processing Series*, 1993.