INCREMENTAL ENROLLMENT OF SPEECH RECOGNIZERS

C. Mokbel, O. Collin

France Télécom - CNET / DIH / DIPS
2, av. Pierre Marzin, 22307 Lannion cedex, France
email: {chafik.mokbel, olivier.collin}@cnet.francetelecom.fr

ABSTRACT

Classical adaptation approaches generally allow a reliably trained model to match a particular condition. In this paper, we define an incremental version of the segmental-EM algorithm. This method permits to incrementally enrich a model first trained with limited amount of data. Resource memory constraints allow only the initial data statistics to be stored. The proposed method uses these statistics by fixing, within the segmental EM algorithm applied on both initial and new data, the initial optimal paths in the model for the initial data. We proved theoretically that this is equivalent to the segmental MAP adaptation with specific choice of *priors*. Experimented on two speaker dependent telephone databases, the approach permitted to incrementally integrate new conditions of use. The performance was slightly less than that obtained with classical training over the whole data. As expected with the MAP interpretation of the algorithm, initial data characteristics influence largely the model evolution.

1. INTRODUCTION

We are currently involved in a telephone speech recognition system using simultaneously speaker independent and speaker dependent word modeling. The speaker independent part is constituted of the same twenty command words for each user. The dependent part is a set of words specific to each user that are real time dynamically added. These speaker dependent words are HMM models estimated, using a fixed variance strategy, with few data: only two or three speech utterances. Obviously, using a particular vocabulary by another speaker gives very high recognition error rates (about 70%). Although, the dependent part of our system is one speaker oriented, in several cases we need to adapt this model to make it more accurate, more robust to line conditions, or even to share it between a small set of users. As long as storage and computational constraints are required, we can't keep all utterances, or duplicate the models. Thus, we experimented an incremental enrollment which allows us to update the dependent models without additional resource requirement.

Classical HMM parameters adaptation techniques [2][1][4] permit to adjust the model parameters in order to better match a particular condition. In that case, the initial model is generally estimated using a large amount of data. This is not the case in our application, where a limited amount of data is first used to estimate the model parameters. This estimate must be adjusted incrementally in order to cover a larger set of conditions (more speakers, different environments, ...). In the following section, we propose an incremental version of the segmental EM algorithm. It appears that this algorithm is a particular case of the *Maximum A Posteriori* (MAP) adaptation algorithm [1] with adequate choice of the *priors* depending on the initial data set. This relation provides another view of the MAP estimation.

Section 3 describes the experiments conducted on two speaker dependent telephone databases in order to validate the approach. The results show that the incremental enrollment permits to a speaker dependent model to incorporate the characteristics of other speakers, and even to converge towards a speaker independent model. The importance of the initial speaker is highlighted and can be easily understood within the MAP framework. Another set of experiments proved the possibility of the incremental enrollment to increase the robustness of the model by integrating various environment conditions. Finally the conclusions are drawn in section 4 and a perspective of this work is proposed in order to reduce the influence of the initial data set.

2. INCREMENTAL ENROLLMENT

2.1 Classical Segmental EM Algorithm

To model speech signals in the feature vectors space of dimension p, "Hidden Markov Models" (HMM) are generally used. These models are composed by a Hidden Markovian automaton of Q states and by a set of output distributions associated to the states of the automaton. A first order ergodic HMM λ is characterized by the probabilities of occupation of the states at the initial time $\Pi = \{\pi_i; i = 1,\ldots,Q\}$, the probabilities of transitions between the states $\underline{\underline{A}} = \{a_{ij}; i, j = 1,\ldots,Q\}$ where a_{ij} is the probability of the transition from the i^{th} state to the j^{th} state, and the states' output distributions here supposed Gaussian with diagonal covariance matrix $\left\{N_i\left(.,\underline{\mu}_i,\underline{\underline{\Gamma}}_i\right); i = 1,\ldots,Q\right\}$.

The HMM parameters are not known *a priori* and are generally trained using speech databases. The training aims to determine the parameters' values such as the HMM describes reliably the distribution of the training speech signals. Since the sequences of states corresponding to the sequences of speech feature vectors are hidden and not observed, there is no direct analytic solution to the training problem. EM-based algorithms are generally used. Given a set of K training speech signals $\underline{\underline{X}} = \left\{\underline{\underline{X}}_1,\ldots,\underline{\underline{X}}_K\right\}$, the optimal HMM parameters λ^{opt} are in the MAP sense:

$$\lambda^{opt} = \arg\max_{\lambda}\left[p\left(\lambda / \underline{\underline{X}}\right)\right] = \arg\max_{\lambda}\left[p(\lambda)\prod_{k=1}^{K} p\left(\underline{\underline{X}}_k / \lambda\right)\right] \quad (1)$$

If no *a priori* is available for the model parameters, MAP criterion becomes equivalent to the Maximum Likelihood (ML) one which can be solved using the EM-algorithm:

$$\lambda^{opt} = \arg\max_{\lambda}\left[\prod_{k=1}^{K} p\left(\underline{\underline{X}}_k / \lambda\right)\right] \quad (2)$$

Here we are particularly interested in the segmental version of the EM algorithm [1], where the criterion to optimize differs slightly from that defined in Eqs. 1 and 2. For each observed sequence of speech frames $\underline{\underline{X}}_k$ different sequences of states S_k in the model can be associated. S_k belongs to a discrete space $\boldsymbol{S}_{(k)}$ which depends on the speech sequence length and on the number of states Q. Regarding these spaces of possible state sequences $\boldsymbol{S} = \{ \boldsymbol{S}_{(k)} \}$, Eqs. 1 and 2 can be written:

$$\lambda^{opt} = \arg\max_{\lambda}\left[\sum_{S\in\boldsymbol{S}} p\left(\lambda, S / \underline{\underline{X}}\right)\right] \quad (3)$$

$$\lambda^{opt} = \arg\max_{\lambda}\left[\prod_{k=1}^{K}\left\{\sum_{S_k\in\boldsymbol{S}_{(k)}} p\left(\underline{\underline{X}}_k, S_k / \lambda\right)\right\}\right] \quad (4)$$

The segmental version of the EM-algorithm replaces the summation in Eqs. 3 and 4 by a maximization, i.e.:

$$\lambda^{opt} = \arg\max_{(\lambda,S)\in\lambda\times\boldsymbol{S}}\left[p\left(\lambda, S / \underline{\underline{X}}\right)\right] = \arg\max_{\lambda\in\lambda}\left[\max_{S\in\boldsymbol{S}} p\left(\lambda, S / \underline{\underline{X}}\right)\right] \quad (5)$$

$$\lambda^{opt} = \arg\max_{\lambda}\left[\prod_{k=1}^{K}\left\{\max_{S_k\in\boldsymbol{S}_{(k)}} p\left(\underline{\underline{X}}_k, S_k / \lambda\right)\right\}\right] \quad (6)$$

Comparing Eqs. 5 and 6 to Eqs. 3 and 4 respectively, the segmental EM algorithm maximizes the joint distribution of the model parameters and the state sequences while the classical EM algorithm maximizes the mean of that distribution over the state sequence space. The segmental EM algorithm is an iterative algorithm where each iteration is decomposed into two steps: the Estimate step where the speech data are aligned on the model and, the Maximize step where the model parameters are re-estimated. The re-estimation equations for the iteration i are:

$$\begin{aligned} a_{lm}^{(i)} &= \frac{n_{lm}}{\sum_{p} n_{lp}} \quad ; \quad \underline{\mu}_l^{(i)} = \frac{1}{n_l}\sum_{k=1}^{K}\sum_{\underline{X}_k(t)\in q_l}\underline{X}_k(t) = \overline{\underline{X}_l} \\ \underline{\underline{\Gamma}}_l^{(i)} &= \frac{1}{n_l}\sum_{k=1}^{K}\sum_{\underline{X}_k(t)\in q_l}\underline{X}_k(t).\underline{X}_k(t)^T - \underline{\mu}_l^{(i)}.\underline{\mu}_l^{(i)T} \end{aligned} \quad (7)$$

where $\underline{X}_k(t) \in q_l$ designates that the t^{th} vectors of the k^{th} sequence is aligned during the Estimate step on the l^{th} state, n_l is the total number of those vectors and, n_{lp} is the total number of transitions from the l^{th} state to the p^{th} state.

2.2 Incremental Segmental EM Algorithm

If a model λ_I is already trained with an initial set of speech data $\underline{\underline{X}}^{(I)}$, and if some new data $\underline{\underline{X}}^{(N)}$ is available to enrich the model parameters, the classical segmental EM algorithm aims to:

$$\lambda^{opt} = \arg\max_{\lambda\in\lambda}\left[\max_{(S_I,S_N)\in\boldsymbol{S_I}\times\boldsymbol{S_N}} p\left(\lambda, S_I, S_N / \underline{\underline{X}}^{(I)}, \underline{\underline{X}}^{(N)}\right)\right] \quad (8)$$

The proposed incremental segmental EM algorithm optimizes only the model parameters and the new state sequences given the whole data. The optimal state sequences of the initial data are fixed the same as in the initially trained model. This means that the state sequences corresponding to the initial training data are considered to be always optimal. Eq. 8 can be written:

$$\lambda^{opt} = \arg\max_{\lambda\in\lambda}\left[\max_{S_N\in\boldsymbol{S_N}} p\left(\lambda, S_I^{opt(\lambda_I)}, S_N / \underline{\underline{X}}^{(I)}, \underline{\underline{X}}^{(N)}\right)\right] \quad (9)$$

Without *a priori* on the model parameters Eq. 9 becomes:

$$\lambda^{opt} = \arg\max_{\lambda\in\lambda}\left[\max_{S_N\in\boldsymbol{S_N}}\left(p\left(\underline{\underline{X}}^{(I)}, S_I^{opt(\lambda_I)} / \lambda\right).p\left(\underline{\underline{X}}^{(N)}, S_N / \lambda\right)\right)\right] \quad (10)$$

For this incremental version, the re-estimation equations for the iteration i can be easily obtained:

$$\begin{aligned} a_{lm}^{(i)} &= \frac{n_{lm}^I + n_{lm}^N}{\sum_p n_{lp}^I + \sum_p n_{lp}^N} \; ; \; \underline{\mu}_l^{(i)} = \frac{n_l^I.\underline{\mu}_l^I + n_l^N.\overline{\underline{X}_l}^N}{n_l^I + n_l^N} \\ \underline{\underline{\Gamma}}_l^{(i)} &= \left[n_l^I.\underline{\underline{\Gamma}}_l^I + n_l^N.\left(\overline{\underline{X}_l.\underline{X}_l^T}^N - \overline{\underline{X}_l}^N.\overline{\underline{X}_l}^{NT}\right) + \right. \\ &\quad \left.\frac{n_l^I.n_l^N}{n_l^I+n_l^N}\left(\underline{\mu}_l^I - \overline{\underline{X}_l}^N\right)\left(\underline{\mu}_l^I - \overline{\underline{X}_l}^N\right)^T\right].\frac{1}{n_l^I + n_l^N} \end{aligned} \quad (11)$$

where all the parameters of the initial model are constant and do not depend on the Estimate step of the iteration.

2.3 Comparison with MAP Estimation

In [1] MAP estimation of HMM parameters is described supposing that the *a priori* distribution of the model parameters is a product of Normal-Wishart and Dirichlet distributions for the Gaussian parameters and the transition probabilities respectively:

$$\begin{aligned} p(\lambda) \propto \prod_{q=1}^{Q}\left\{\prod_{l=1}^{Q} a_{ql}^{\nu_{ql}-1}\left\|\underline{\underline{R}}_q\right\|^{(\alpha_q-p)/2}.\exp\left[-\frac{1}{2}tr\left(\underline{\underline{U}}_q\underline{\underline{R}}_q\right)\right]\right. \\ \left.\exp\left[-\frac{\tau_q}{2}(\underline{\mu}_q - \underline{m}_q)^T\underline{\underline{R}}_q(\underline{\mu}_q - \underline{m}_q)\right]\right\} \end{aligned} \quad (12)$$

where $\underline{\underline{R}}_q$ represents the precision matrix of the q^{th} distribution ($\underline{\underline{\Gamma}}_q^{-1}$) and $\underline{\underline{U}}_q$ and α_q the parameters of its *a priori* distribution, p the dimension of the feature space, $\underline{m}_q$ and τ_q the mean and precision factor of the *a priori* distribution of the mean vector. Comparing the re-estimation equations in the MAP framework with those in Eq. 11, we conclude that the proposed approach is a particular case of the MAP adaptation with the *priors*:

$$\begin{aligned} &\tau_q = n_q^I \quad \alpha_q = n_q^I + p \quad \nu_{ql} = n_{ql}^I + 1 \\ &\underline{m}_q = \underline{\mu}_q^I \quad \underline{\underline{U}}_q = n_q^I.\underline{\underline{\Gamma}}_q^I \end{aligned} \quad (13)$$

In parallel, MAP adaptation can be seen as the proposed incremental enrollment, with the characteristics of the initial data used for training related to the *priors*.

3. EXPERIMENTS AND RESULTS

3.1 Databases and Modeling

Two telephone databases (bdd1 and bdd2) were used in the experiments involving simultaneous speaker dependent and speaker independent recognition. The vocabulary of each database is composed of two main parts: French command words and digits (speaker independent) and several collections of proper names. The first database, bdd1, was collected from 16 (11 male and 5 female) speakers over the fixed telephone network. The second database, bdd2, was collected from 40 (20 male and 20 female) speakers over both fixed and mobile telephone networks. Each speaker repeated the entire vocabulary 4 times for training issues over the fixed telephone network for bdd1 and over both fixed and cellular network for bdd2. In average, each speaker in the bdd1 database repeated, for evaluation, 4 times the vocabulary in home and office environments. For the test part of the bdd2 database, an average by speaker of 2 and 7 repetitions of the vocabulary were collected in the fixed and cellular networks respectively. 69% of the GSM test data were collected from a running car. These two databases helped us to develop the France Telecom - CNET voice activated dialing system where simultaneous speaker independent and dependent recognition (SSIDR) modes are combined [3].

Feature analysis consists in computing on 32ms frames shifted every 16ms, the energy on logarithmic scale, 8 MFCC coefficients and their first derivatives resulting in sequences of vectors with 18 coefficients. Those sequences of vectors were modeled using 20 state left-right HMMs. 13 command words were used in the speaker independent part of the model. These words were trained using other fixed and cellular telephone databases including several hundred of speakers. The speaker dependent part of the model is constructed using our databases with fixed variances. Training is done using 2 to 4 repetitions. The experiments were conducted in order to investigate if the incremental enrollment described previously can help to enrich the speaker dependent part of the model by including different conditions for the same speaker (for example PSN + GSM) or, other speakers towards speaker independent modeling. The influence of the first speaker on the incremental enrollment procedure is also studied.

3.2 Enriching Incrementally a Speaker Dependent Model with Other Speakers

These experiments aimed at investigate the possibility of evolution of a speaker dependent model to include other speakers. Two bounds can be measured for the incremental enrollment. The first bound is the performance of the speaker dependent system on other speakers. For the bdd1 database, 30 proper nouns were added on the independent model and were trained, using 2 repetitions of a specific speaker, and were evaluated on the repetition test by all the other speakers. The average error (substitution with the nouns or commands) rate was about 67%. The second bound is the one obtained with the classical training using all the data from the first and added speakers, 2 repetitions by speaker. The results are shown on Fig. 1 as well as the results with the incremental enrollment for adding one or two speakers. These results are given in average for all the combination of speakers in the database. Looking at Fig. 1, it turns out that incremental enrollment provides worse performance than direct training with grouped data. However, the obtained performance is acceptable compared to the starting 67% error rate, especially in regards to the practical advantages of the incremental enrollment: there is no need to conserve the initial data. Fig. 2 shows in average, for each speaker following its position, the performance after incremental enrollment. It can be seen that the performance for the first speaker is nearly constant and is better than that obtained with a classical training. However, the results for the added speakers are worse than those obtained with the classical training algorithm.

The same experiments where conducted on 20 (10 male and 10 female) speakers of the PSN part of the bdd2 database. While a first 10 (5 male and 5 female) speakers were used for incrementally enriching the speaker dependent model, the other 10 speakers were used for evaluating the obtained model in a speaker independent mode. Average results for the initial speaker, the added speakers, and the other speakers are shown on Fig. 3. These results show that the incremental enrollment seems to provide constant performance for the initial speaker and the added speakers, independently of the number of added speakers, while the performance in the speaker independent mode is improved with the number of added speakers.

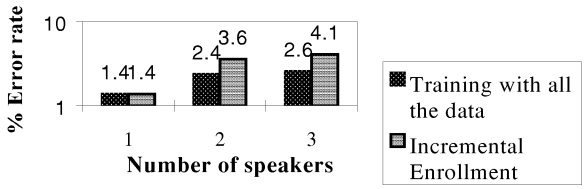

Figure 1. Results when adding other speakers to a speaker dependent model on the bdd1 database.

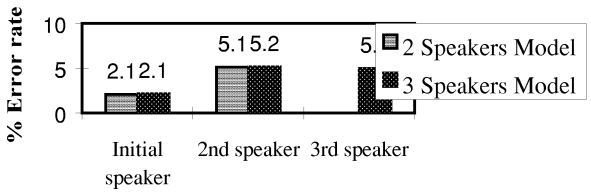

Figure 2. Results of incremental enrollment on bdd1 database function of the speaker position.

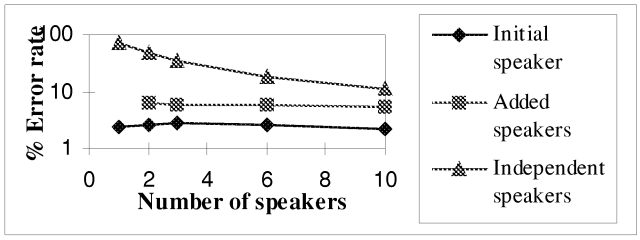

Figure 3. Results when adding other speakers to a speaker dependent model on the bdd2 database.

From the results of Figs. 2 and 3, we conclude that the initial speaker characteristics influence largely the incremental enrollment of the parameters. This can be explained with regards to section 2.3, since the proposed approach is seen as a particular case of the MAP estimation with an *a priori* distribution derived from the first speaker characteristics. Thereby, it is unlikely that the model's parameters go far away from the local optimum relative to the first speaker.

3.3 Enrich Incrementally a Speaker Dependent Model with Other Conditions

Experiments were conducted on the 20 previous speakers of the bdd2 database in order to study the incremental enrollment of their models to include different conditions. Fig. 4a shows the average results on the fixed telephone part of the database. The measured improvements with respect to the basic model are equivalent to a classical enrollment approach and, to the proposed incremental approach. Fig.4b shows the average results for the GSM conditions. In this figure the whole GSM and the quiet GSM data (not collected in a running car) are distinguished. This distinction is done since the training GSM repetitions are all from quiet conditions. According to the results, the incremental enrollment achieves performance as good as the classical enrollment with the whole data. The initial environment condition has a less influence in this case than in the previous application.

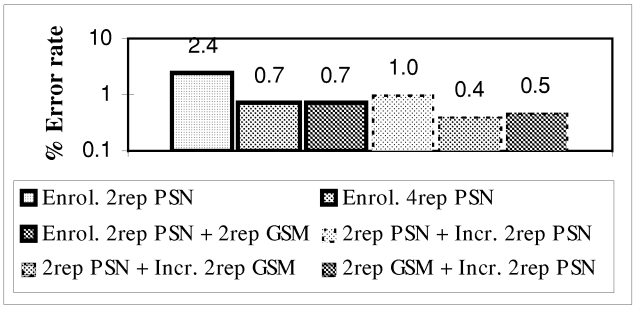

Figure 4a. Average PSN results when enriching the model with other training repetitions.

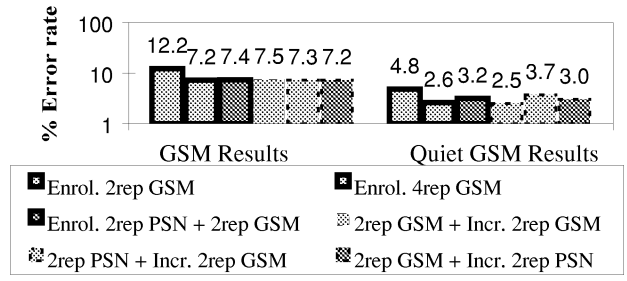

Figure 4b. Average GSM results when enriching the model with other training repetitions.

4. CONCLUSIONS

In this paper, an incremental enrollment procedure is proposed. Starting from an initial model trained with few initial data, for every set of new data the proposed incremental version of the segmental EM algorithm is used to update the model parameters by fixing, in the Estimate step, the optimal paths in the model corresponding the initial data, to their values on the initial model, and by optimizing the model parameters and the paths for the new data. We have theoretically shown that this approach is identical to the MAP adaptation approach with specific *priors* extracted from the initial data statistics. This permits to interpret the *priors* in the classical MAP framework as the statistics of some initial data used for training and for which the paths in the model are fixed to their optimal values on the *a priori* model.

The application of the proposed approach is slightly different from classical adaptation techniques. It aims at incrementally enrich a model initially trained with a reduced amount of data. For example, a personal directory (voice dialing) can be enriched for use in a family or in various conditions. In this context we have experimented the algorithm on the France Telecom - CNET personal directory system, where simultaneous speaker independent and dependent recognition is performed. The new algorithm was first experimented in order to enrich the speaker's voice labels by using labels from other speakers. The measured performance provides satisfaction since it is close to the performance of a system trained directly with the initial and new data. The results showed that the proposed algorithm is constrained by the initial data characteristics. This is explained by the fact that the initial speaker defines the *priors* when seeing the algorithm in the MAP framework. Another set of experiments has shown that by adding incrementally speakers' characteristics to the model, the proposed algorithm might converge the model to a speaker independent one. The results obtained when enriching incrementally the model by integrating different conditions of use (PSN+GSM) were also equivalent to those obtained with a classical training using the whole data.

The main perspective of this work is the reduction of the difference in performance between the first speaker and the added speakers. A simulated annealing version [5] of the proposed algorithm might be developed in order to allow the algorithm converge far from the local optimum defined by the initial training data.

5. REFERENCES

[1] Gauvain J.-L., Lee C.-H., ''Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,'' IEEE Trans. on SAP, V. 2, n° 2, pp. 291-298, Apr. 1994.

[2] Lee C.-H., Lin C.-H., Juang B.-H., ''A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models,'' IEEE Trans. on ASSP, V. 39, n° 4, pp. 806-814, Apr. 1991.

[3] Mokbel C., ''SSIDR: A software module to integrate speaker dependent recognition within PHIL90,'' (in French) France Télécom-CNET internal report, DT/379/LAA, n° 22, 1996.

[4] Mokbel C., Mauuary L., Karray L., Jouvet D., Monné J., Simonin J., Bartkova K., ''Towards Improving ASR Robustness for PSN and GSM Telephone Applications,'' Speech Communication, V. 23, n°1, pp. 141-159, Oct. 1997.

[5] Simonin J., Mokbel C., ''Using Simulated Annealing Expectation Maximization Algorithm for Hidden Markov Model Parameters Estimation,'' *EuroSpeech*, 1997, p. 449.