Bayesian Adaptation of Speech Recognizers to Field Speech Data

Carmelo Giammarco MIGLIETTA, Chafic MOKBEL, Denis JOUVET & Jean MONNÉ

France Télécom - CNET / LAA / TSS / RCP
2 av. Pierre Marzin, 22307 Lannion cedex, France
e-mail: mokbel@lannion.cnet.fr

ABSTRACT

This work studies a *Bayesian* (or *Maximum A Posteriori* MAP) approach to the adaptation of Continuous Density Hidden Markov Models (CDHMMs) to a specific condition of a speech recognition application. In order to improve the model robustness, CDHMMs formerly trained from laboratory data are then adapted using context dependent field utterances. Two specific problems have to be faced when using the MAP approach: the estimation of the *a priori* distribution parameters and the lack of field adaptation data for some distributions of the CDHMM.

To estimate the *a priori* distribution parameters, we need to identify different realizations of the model parameters. Three different solutions are proposed and evaluated. To overcome the lack of adaptation data, field acoustical training frames may be shared among similar distributions. This is performed using an acoustical tree, obtained by progressively clustering the model distributions.

Recognition results show that MAP adapted models significantly outperform those trained by *Maximum Likelihood* (ML), specifically when the field data set is small.

1. INTRODUCTION

The recognition of laboratory speech data has reached satisfactory results nowadays. When passing to practical applications, performances of speech recognizers often worsen in a considerable way [5], which impedes a large scale commercial use of speech recognition. In fact, field speech data are affected by *ambient noise*, and *telephone line distorsion* in case of telephone applications, and *hesitations* of the speaker and *coarticulation phenomena*, typical of natural speech. These factors make field speech data significantly different from laboratory ones, causing the loss of efficiency of speech recognizers trained using laboratory utterances.

In order to get more robust models for a specific task, field data collected in the real task condition (e.g. users' calls to a vocal server) should be integrated in the training set. A model prototype trained from laboratory utterances can be used to collect field data, that will be used to train a more robust model. The main drawback is that field data are often *sparse*: some words or sentences may be seldom uttered, hence there may not be enough data to train some parts of the model. Therefore, the completion of the training set with laboratory data is necessary.

A *Maximum Likelihood* (**ML**) solution consists in retraining the model using combined field and laboratory data as a single training set. This work studies a *Bayesian*, or *Maximum A Posteriori* (**MAP**) approach. The model, formerly trained from laboratory data, is adapted to the context using field data. The task considered is a French vocal server using a lexicon of 26 isolated words.

This paper is structured as follows. In Section 2, the Bayesian framework for model adaptation is considered and the main problems of the MAP approach are individuated. In Section 3, three strategies are proposed for the estimation of the prior distribution parameters. In Section 4, we propose the use of an acoustical tree to overcome the lack of adaptation data. In Section 5, recognition experiments and results obtained are described. Recognition performances of MAP trained models are compared to those obtained by ML trained models. Finally, in Section 6, the conclusions of this work and some prospects of Bayesian model adaptation are given.

2. MAP ESTIMATION AND PROBLEM DEFINITION

Let λ be a CDHMM model and θ its parameter vector. Let $\mathbf{X} = (x_1, \ldots, x_T)$ be a set of T observations (in the form of acoustical frames), which we consider to be the training set. The ML training approach considers the parameters θ to be *fixed*, and their *unknown values* are estimated from the observations $\mathbf{X}$ as follows:

$$\theta = \arg\max_{\theta} p(\mathbf{X}|\theta)$$

The MAP approach considers the parameters θ of the model λ to be *random values*. For a given condition the parameters θ are a *realization of a random variable* called Θ. Therefore, in this case the training problem consists in finding the most likely realization $\hat{\theta}$ of Θ, given the sample $\mathbf{X}$:

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{X}) \qquad (1)$$

The Bayesian approach is suitable to the adaptation of a model to a specific application condition: each context (the task, the speaker, the noisy or calm environment) is modeled by a proper realization of Θ. This approach allows the introduction of *a priori* information in the adaptation (*training*) process. In fact, the parameters θ of λ are distributed according to the law of Θ, called *prior distribution* $p_{\phi}(\theta)$. The prior distribution represents

the information we know about the model before any specific observation is made. This information can be obtained from known physical properties, from past experiences or it can be estimated from a database including a large set of independent contexts. This prior knowledge is useful especially when the adaptation data set is small. Equation 1 can be written as follows, according to the Bayes's theorem:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(\mathbf{X}|\theta)\, p_{\phi}(\theta) \tag{2}$$

The estimation problem is solved by evaluating the maximum point of the product $p(\mathbf{X}|\theta)p_{\phi}(\theta)$. In this work, we use the *segmental* version of the iterative *EM algorithm*. For each iteration of the algorithm, the field acoustical frames are Viterbi aligned on the model distributions, and then the current iteration parameters are evaluated. This approach is called *Segmental MAP* [4][1]. The first problem arisen is the choice of the *prior* distribution family. Since no physical knowledge is available to justify a particular choice, a *mathematical attractiveness* criterion is applied [1]. This means that the *prior* distribution family is chosen in order to simplify the determination of the solution of Equation 2.

In this work, an allophonic CDHMM, with single Gaussian distributions, is used [3]. The distributions have diagonal covariance matrix, hence we can consider each direction in the acoustical space independently. Only the adaptation of the means and variances of the distributions is made. In fact, former experiences show that the importance of the transition probabilities is minor.

Let m_i and r_i be respectively the mean and the precision matrix (inverse of covariance matrix) of the distribution *i*. We denote m_{il} and r_{il} respectively the *l*-th component of m_i and the *l*-th diagonal member of r_i. On the *l*-th dimension of the acoustical space, the mean and precision joint *prior* distribution is the following *normal-Wishart*:

$$p(m_{il}, r_{il} | \tau_{il}, \mu_{il}, \alpha_i, u_{il}) \propto r_{il}^{\frac{\alpha_i - 1}{2}} \cdot$$

$$\exp\left[-\frac{1}{2}(u_{il} \cdot r_{il})\right] \cdot \exp\left[-\frac{\tau_{il}}{2}(m_{il} - \mu_{il})^2 r_{il}\right] \tag{3}$$

where $\phi = (\tau_{il}, \mu_{il}, \alpha_i, u_{il}),\ i = 1 \ldots N,\ l = 1 \ldots p$ are the *prior* density parameters and *p* is the dimension of the acoustical space [1][4]. This choice yields the following MAP estimation formulae for the current parameters at each EM iteration [1]:

$$\tilde{m}_{il} = \frac{\tau_{il}\mu_{il} + n_i \bar{x}_{il}}{\tau_{il} + n_i} \tag{4}$$

$$\tilde{r}_{il}^{-1} = \frac{u_{il} + n_i S_{il} + \frac{\tau_{il} n_i}{\tau_{il} + n_i}(\mu_{il} - \bar{x}_{il})^2}{(\alpha_i - 1) + n_i} \tag{5}$$

where n_i is the number of the field acoustical frames Viterbi aligned on the distribution *i*, $\bar{x}_{il}$ and S_{il} are the *empirical* mean and variance, calculated for each component *l* using these acoustical frames.

According to equations 4 and 5, the distributions parameters are adapted to a specific context using adaptation data: i.e. some utterances recorded in the field condition. However, in order to get an accurate adaptation, the *prior* distribution parameters ϕ should be well estimated, and a sufficient amount of adaptation data is requested. These two points are detailed in the following.

3. ESTIMATION OF A PRIORI DENSITY PARAMETERS

Since any physical knowledge about the *prior* parameters ϕ is not available, ϕ is estimated from a large laboratory (hence, condition independent) database. This strategy is called *Empirical Bayes* (**EB**) approach [2]. In order to evaluate these *prior* distribution parameters ϕ, we have to consider *different realizations of the model parameters* Θ. These realizations cannot be directly observed, so we have to *estimate* them. Three different approaches are proposed in the following.

Let mean m_{ilq} and variance r_{ilq}^{-1} be the *q*-th realization of the *l*-th component of the *i*-th density. Referring to Equation 3, we can observe that μ_{il} is the mean of the *prior* distribution relative to m_{il}. So, it can be evaluated as a weighted mean of m_{ilq}. If we call σ_{il}^2 the variance estimate obtained from r_{ilq}^{-1} and m_{ilq}, and assign this value to the mode of the variance distribution, then it is easy to verify that: $u_{il} = (\alpha_i - 1)\sigma_{il}^2$. If α_i is fixed, then u_{il} is determined. Note that, as α_i increases, the Wishart distribution tightens around its mode. In the adaptation stage, the variance value will not change significantly from the prior mode if α_i has a high value, unless the adaptation training set is large. Since the estimation of the CDHMM distributions variances is critical, a high value for α_i is empirically chosen. Finally, the product $(\tau_{il} r_{il})^{-1}$ is the variance of the *prior* distribution of the mean m_{il}. Let Σ_{il} be the empirical variance of the mean, estimated from the realizations m_{ilq}. Then, τ_{il} is evaluated as $\tau_{il} = \sigma_{il}^2 / \Sigma_{il}$. Note that in this work a specific value of τ_{il} is estimated for each dimension of the acoustical space.

3.1. Associating several calls as a realization of the prior distribution

The laboratory database we use is composed of telephone calls made by speakers from different regions in France. They utter the lexicon words and some out-of-vocabulary words. These out-of-vocabulary words are used to train the garbage models.

Each phone call identifies a given speaker, a transmission channel and a specific ambient noise. Hence, it is reasonable to suppose that a single telephone call corresponds to one realization of the model parameters Θ. Therefore, we can use the speech data of a call to estimate a realization of the parameters. Unfortunately, a single call might not contain enough data for a reliable estimation of the model parameters.

A possible solution is to associate several calls to a single realization of Θ which provides more data to correctly estimate

the model parameters. The drawback is that calls associated to the same realization may not refer to the same speaker and ambient noise. Hence, a *tradeoff* between *modelling accuracy* and *quantity of estimation data available* is to be made.

3.2. Vector Quantization to Estimate the Realizations of the Prior Distribution

The parameters of the distributions are evaluated using the acoustical frames aligned on them at the last EM iteration. Hence, a set of acoustical frames is associated to each distribution. In each set, a vector quantization is performed, i.e. we divide the set into a given number of classes, using the LBG algorithm. Each class is supposed to be a realization of the *a priori* distribution to which the whole set is associated.

In [2], clustering was performed on speakers, in order to obtain speakers' groups, each group corresponding to a single realization. Here the clustering is directly operated on the acoustical frames. The advantage is that we can study each parameter independently instead than considering a general inter-speaker variability. In other words, we suppose that speakers classes may differ depending on the sound pronounced. This clustering is also suitable for noise classes since additive noise effects depend on the pronounced sounds.

3.3. Gaussian Mixtures to Estimate the Realizations of the Prior Distribution

The CDHMM used in this work has single Gaussian distributions. Starting from this model, a new one with Gaussian mixtures is obtained. Each mixture component is considered to be a realization of the *prior* distribution corresponding to the associated Gaussian distribution in the original model.

4. LACK OF ADAPTATION DATA AND ACOUSTICAL TREE

Model adaptation to the context is performed using field data collected when the recognition system is operating. To get a satisfactory quantity of field data may take a long time: in fact, in the case of a vocal server, several users' calls are needed. Moreover, field data are often unbalanced: some lexicon words might be seldom uttered, while other words might be frequent.

For those reasons, lack of adaptation data is quite common. In consequence of this, in the adaptation stage, some distributions of the model are not seen, i.e. the quantity of field frames associated to them is not enough for a correct estimation of $\bar{x}_{il}$ and S_{il} (Equations 4 and 5). Hence, the adaptation process can be performed only for some parts of the model, causing a lack of balance.

Supposing the adaptation function to be continuous in the acoustical space, it can be argued that acoustical frames may be shared among "similar" distributions, in order to increase the quantity of estimation data. An acoustical tree is built to associate "similar" distributions.

In the adaptation stage, when field data are lacking for a density, the tree is climbed and the adaptation frames of the close Gaussian densities are added to the training set of the initial density, until a satisfactory quantity of data is reached.

The acoustical tree is built by progressively coupling the model distributions, according to a minimal distance criterion. Let $\mathcal{N}_1(.,m_1,\Sigma_1)$ and $\mathcal{N}_2(.,m_2,\Sigma_2)$ be two Gaussian distributions to which n_1 and n_2 acoustical frames have been associated. This means that we suppose that the n_1 and n_2 frames are issued respectively by $\mathcal{N}_1$ and $\mathcal{N}_2$. To couple $\mathcal{N}_1$ and $\mathcal{N}_2$ means to replace them by a single Gaussian $\mathcal{N}_3(.,m_3,\Sigma_3)$, which issues the whole $n_3 = n_1 + n_2$ frames. It is easy to verify the following relations:

$$m_3 = \frac{n_1}{n_1+n_2} m_1 + \frac{n_2}{n_1+n_2} m_2$$

$$\Sigma_3 = \frac{n_1}{n_1+n_2}\Sigma_1 + \frac{n_2}{n_1+n_2}\Sigma_2 + \frac{n_1 n_2}{(n_1+n_2)^2}(m_1-m_2)(m_1-m_2)^T$$

This coupling operation yields a loss of precision in the modelling process. A distance between two distributions should express the loss of precision caused by coupling them. One possible distance is the log likelihood ratio between $\mathcal{N}_1$ & $\mathcal{N}_2$ and the new distribution $\mathcal{N}_3$ on the whole n_3 frames:

$$\log \frac{\|\Sigma_3\|^{\frac{n_1+n_2}{2}}}{\|\Sigma_1\|^{\frac{n_1}{2}}\|\Sigma_2\|^{\frac{n_2}{2}}}$$

The leaves of the acoustical tree are the original model distributions. Each step of the tree construction consists in choosing the two closest Gaussian densities and associating to their father node the new Gaussian density, that will replace both in the next distance evaluation step. This procedure is iterated until the root node is obtained.

5. RECOGNITION EXPERIMENTS AND RESULTS

The task model is composed of 26 allophonic word models and four garbage (word) models, to reject the out-of-vocabulary words and the ambient noises. Our laboratory database is composed of 7 hours 25 minutes of telephone data. The field database is composed of real users' telephone calls to a French vocal server. It is divided into equal parts for training and test. The training part contains 4 hours 25 minutes of signal. Field speech data are divided into four categories: correct (CORR), i.e. lexicon words correctly uttered; truncated (TRUN), i.e. lexicon words truncated by the speech-noise detector; out-of-vocabulary (OOV) words; ambient spurious noises (NOIS). In the recognition stage, we wish to correctly recognize correct and truncated words and to reject out-of-vocabulary words and ambient noise.

Two ML models have been trained: the *laboratory* one, trained using only laboratory data; the *mixed* one, using laboratory and field data as a single training set. Their error rates are reported in

Table 1. We observe that the mixed model outperforms the laboratory one, which is a confirmation of the necessity to use field data in the training. In the same table, we report the error rate reductions obtained by the MAP models referred to the mixed model. For each prior parameters estimation strategy, we have chosen the MAP models which perform the best, namely *8 calls*, *4 Gaussians mixtures* and *8 VQ classes* models. We can observe that MAP models largely outperform the mixed one, especialy for correct words recognition and noise rejection.

	CORR	TRUN	OOV	NOIS
ML laboratory	6.7	66.5	24.4	16.7
ML mixed	3.8	52.4	29.1	11.2

Table 1a: ML models error rates.

	CORR	TRUN	OOV	NOIS
8 calls	*13 %*	*2.3 %*	*4.8 %*	*57 %*
4 Gauss. mixtures	*14.5 %*	*1.1 %*	*0.2 %*	*54 %*
8 VQ classes	*12.6 %*	*1.6 %*	*3.6 %*	*57 %*

Table 1b: MAP models error rate reduction with respect to the ML mixed model.

ML mixed and MAP models in the former experiments were trained using 4 hours 24 minutes of field speech. In Figure 1, recognition results of models trained with a growing amount of minutes of field data are reported. The results represent the error rate reduction referred to the performances of the laboratory model, which is not trained from field data (see Table 1). We can observe that MAP models largely outperform the ML one, especially for correct words recognition and noise rejection. When the amount of field training data available is small, the use of the acoustical tree significantly improve the MAP model performances in correct data recognition. The acoustical tree worsens the performances in out-of-vocabulary noise rejection. We believe that this is due to a bad coupling of the garbage distributions with some lexicon allophone distributions in the tree construction stage. Using two acoustical trees for both vocabulary and garbage models would avoid this problem. Note that the MAP model adapted from 57 minutes of field speech is better that the mixed ML one trained from 4.5 hours.

6. CONCLUSIONS AND PROSPECTS

Three approaches to the estimation of the *a priori* distributions parameters are proposed and evaluated in the framework of MAP adaptation of CDHMM to specific field conditions. For the lack of the adaptation data, speech frames may be shared between "*similar*" distributions. A proposed acoustical tree permits an hierarchical clustering of the CDHMM densities.

The comparison between the performances of MAP and ML models, trained using the same quantity of adaptation data, shows that Bayesian models obtain significantly better results on field speech recognition. This result stands for the three approaches proposed for the estimation of the *prior* parameters. The gain obtained using MAP models is even more evident when the adaptation data set is small. With the acoustical tree, MAP task adaptation using 1 hour of field speech outperforms the ML approach using 4.5 hours of field speech. This means that the Bayesian adaptation reduces the size of the training field data set required by a factor 4.5. The use of the acoustical tree improves the performances of the Bayesian models, especially when the adaptation data set is small. Therefore, on-line model adaptation to the context seems to be realistic.

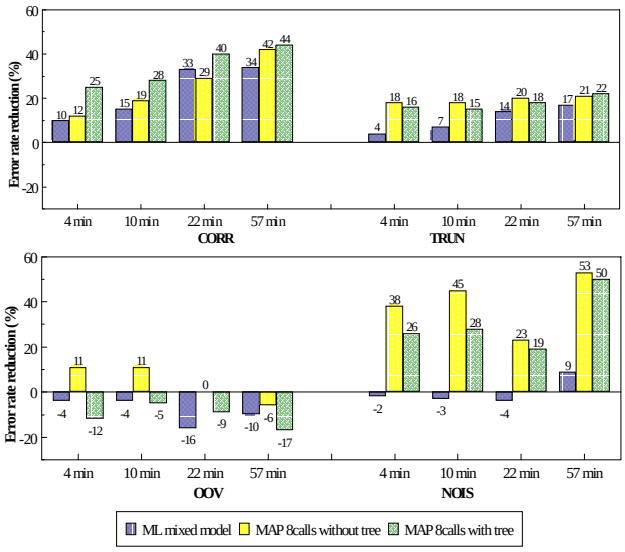

Figure 1: Error rate reductions obtained using ML mixed and MAP models with respect to the ML laboratory one.

8. REFERENCES

1. Gauvain, J.-L. and Lee, C.-H. "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains" *IEEE Trans. SAP*2: 291-298, april 1994

2. Huo, Q., Chan, C. and Lee, C.-H. "Bayesian Adaptive Learning of the Parameters of Hidden Markov Model for Speech Recognition" *IEEE Trans. SAP*3: 334-345, september 1995.

3. Jouvet, D., Bartkova, K. and Monné, J. "On the Modelization of Allophones in a HMM Based Speech Recognition System" *EUROSPEECH*: 923-926, 1991.

4. Lee, C.-H., Lin, C.-H. and Juang B.-H. "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models" *IEEE Trans. SP*39: 806-814, april 1991.

5. Mokbel, C. and Chollet, G. "Automatic Word Recognition in Cars" *IEEE Trans. SAP*3: 346-356, september 1995.