SPEECH RECOGNITION : NEW TECHNIQUES FOR SPEAKER ADAPTATION

Olivier Bellot,
Doctorant en informatique
olivier.bellot@lia.univ-avignon.fr + 33 4 90 84 35 58

Driss Matrouf,
Maître de conférences en informatique
driss.matrouf@lia.univ-avignon.fr + 33 4 90 84 35 26

Adresse professionnelle
Laboratoire Informatique d'Avignon (LIA), 339, chemin des Meinajaries Agroparc BP 1228 84911 AVIGNON Cedex 9 FRANCE

Résumé : Les systèmes de reconnaissance de la parole utilisant des modèles acoustiques dépendants du locuteur sont plus performants que ceux basés sur des modèles indépendants du locuteur. Le but des techniques d'adaptation est d'améliorer ces derniers modèles pour s'approcher des performances obtenues avec un modéle dépendant du locuteur. Dans cet article, nous proposons deux nouvelles méthodes d'adaptation. La première utilisant les données de test et d'apprentissage pour adapter les modèles indépendants du locuteur, la seconde étant une technique d'adaptation basée sur une classification hiérarchique des gaussiennes composant le modèle acoustique. Ces stratégies d'adaptation ont été évaluées sur le corpus de test de l'AUPELF, ARC B1. Ces deux techniques permettent respectivement un gain relatif par rapport au système initial de 15% pour la première technique et de 16% pour la seconde.

Summary : The speaker-dependent HMM-based recognizers gives lower word error rates in comparison with the corresponding speaker-independent recognizers. The aim of speaker adaptation techniques is to enhance the speaker-independent acoustic models to bring their recognition accuracy as close as possible to the one obtained with speaker-dependent models. In this paper, we propose two new method: the first method operates in two steps using test and training data and the second is a hierarchical These adaptations strategies were evaluated in a large vocabulary speech recognition task. The first method leads to a relative gain of 15 % with respect to the baseline system and 10 % with respect to the conventional MLLR adaptation, whereas the second leads to a relative gain of 16% with respect to the baseline system.

Mots clés : Communication Homme-Machine, Reconnaissance automatique de la parole, modèles acoustiques, adaptation au locuteur.

Key words : Human-computer communication, Automatic speech recognition, acoustic models, speaker adaptation.

Speech Recognition: New techniques for speaker adaptation

The speaker-dependent HMM-based recognizers have lower Word Error Rates (WER) than speaker-independent ones. In fact, modeling inter-speaker variability is usually performed by training acoustic models with as large as possible population of speakers. This training manner leads to a relative high variance in acoustic models and hence reduces discriminatory capabilities between different phonemes, especially in the context of larger perplexity tasks. Nevertheless, in the speaker-dependent case, the requirement of large amount of training data for each test speaker reduces the utility and portability of such systems.

The aim of speaker adaptation techniques is to enhance the speaker-independent acoustic models to bring their recognition accuracy as close as possible to the one obtained with speaker-dependent models.

In this paper, we will present two different approaches to increase the robustness of speech regnonizer with respect to the speaker acoustic variabilities. The first one is a method using test and training data for acoustic model adaptation . This method operates in two steps : the first one performs an *a priori* adaptation using the transcribed training data of the closest training speakers to the test speaker. This adaptation is done with MAP procedure allowing reduced variances in the acoustic models. The second one performs an *a posteriori* adaptation using the MLLR procedure on the test data, allowing mapping of Gaussians means to match the test speaker's acoustic space. This adaptation strategy was evaluated in a large vocabulary speech recognition task. Our method leads to a relative gain of 15% with respect to the baseline system and 10% with respect to the conventional MLLR adaptation.

The second method presented in this paper is based on tree structure. Within the framework of speaker-adaptation, a technique based on tree structure and the maximum a posteriori criterion was proposed (SMAP)[15]. In SMAP, the parameters estimation, at each node in the tree is based on the assumption that the mismatch between the training and adaptation data is a Gaussian PDF which parameters are estimated by using the Maximum Likelihood criterion. To avoid poor transformation parameters estimation accuracy due to an insufficiency of adaptation data in a node, we propose a new technique based on the maximum a posteriori approach and PDF Gaussians Merging. The basic idea behind this new technique is to estimate an affine transformations which bring the training acoustic models as close as possible to the test acoustic models rather than transformation maximizing the likelihood of the adaptation data. In this manner, even with very small amount of adaptation data, the parameters transformations are accurately estimated for means and variances. This adaptation strategy has shown a significant performance improvement in a large vocabulary speech recognition task, alone and combined with the MLLR adaptation.

1 – ADAPTATION USING TEST AND TRAINING DATA

1.1 – Introduction

To deal with inter-speaker variability, two classes of approaches have been studied. The first one consists in performing normalization in the feature space. This class contains the cepstral mean removal technique [1], the vocal track length normalization [2], a feature space normalization based on mixture density Hidden Model Markov (HMM) [3], and a signal bias removal estimated by Maximum Likelihood Estimation (MLE) [4].

The second class of approaches operates in acoustic model space. In this class a compact model for Speaker Adaptative Training (SAT) technique was introduced in [5]. This technique consists in modeling separatelly the speaker variation and removing its effect in the training data. Thus the variance of models is reduced and hence the overlap of the acoustic models. The most used techniques in the second class consist in adapting the speaker-independent models to a specific speaker so as to obtain a recognition accuracy as close as possible to the one obtained on speaker-dependent system. In this framework, many adaptation schemes have been proposed: in [6] Maximum *A Posteriori* (MAP) estimations

techniques were proposed. It attempts to obtain a Bayesian estimate of the model parameters using adaptation data available from the test speaker. In [7] the speaker-independent system is transformed to come closer to the test speaker by applying a linear transformation on the means of speaker-independent Gaussians. The transformation is estimated so as to maximize the likelihood of the test speaker's adaptation data.

Other adaptation schemes are based on the fact that the training data contains a number of training speakers, some of whom are acoustically closer, to the test speaker, than the others [8]. This technique uses the adaptation data to find a subset of the training speakers which are closer to the test speaker. And then, it compute and apply a linear transformation to map the acoustic space of each selected training speaker closer to the test speaker's acoustic space. The linear transformation is computed by using the MLLR procedure [9].

In this paper, we propose a method using training and test-data for acoustic model adaptation. There are two steps in this method: the first one performs an *a priori* adaptation using transcribed training data with MAP adaptation. The second one performs an *a posteriori* adaptation using test data with MLLR adaptation. Both modifications have different goals: the former allows a reduced variance in acoustic models whereas the latter allows a mapping of the acoustic models means to be closer to the test speaker's acoustic space.

In the next section we present the proposed adaptation method. We describe the goals and the strategies for the *a priori* and the *a posteriori* speaker adaptations. In section 1.3 we describe two strategies for training-speakers selection. Section 1.4 shows results for several recognition experiments in large vocabulary task framework.

1.2 – Adaptation process

Because of the inter-speaker variability modeling, the speaker-independent models have a relative large variance in comparison with the corresponding speaker-dependent models. By using the MLLR adaptation we only adapt the Gaussians means, so the resulting acoustic models still have a relative high variance and hence an high overlap among different speech units, resulting in reduced discriminatory capabilities. To reduce variances, one way is to use the MAP (Maximum *A Posteriori*) adaptation [6]. But this process requires a relative large amount of adaptation data to re-estimate all Gaussians variances.

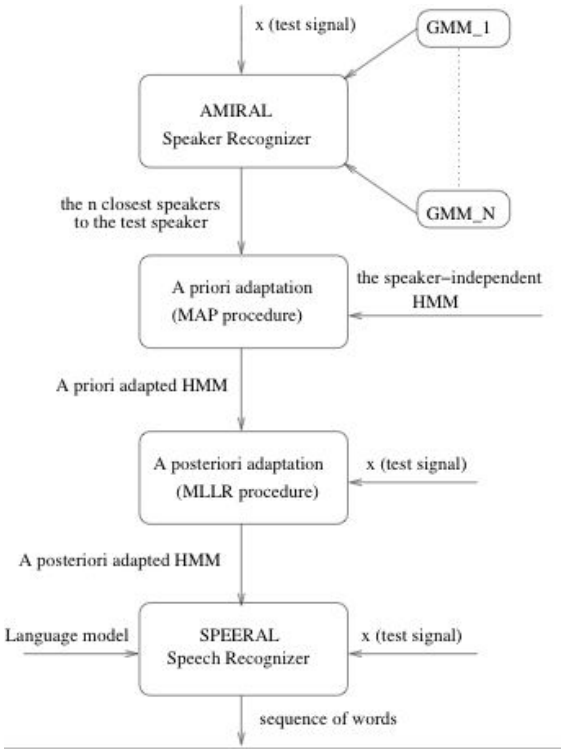

Figure 1: *Adaptation process*

In this paper, we propose a strategy resulting inadapted acoustic models with reduced variances. The adaptation is performed in two steps (see Figure 1). The first adaptation step, that we term the *a priori adaptation* is based on selecting a cluster of training speakers who "are" acoustically close to the test speaker. Then the speaker-independent acoustic models are adapted by using the transcribed training data corresponding to those selected speakers. This first adaptation is done with the MAP procedure, which transforms the means, variances and gains of Gaussians : let g a Gaussian having μ_g as mean and Σ_g as variance in the speaker-independent acoustic models. The new mean $\tilde{\mu}_g$ and variance $\tilde{\Sigma}_g$ of the Gaussian are given by:

$$\tilde{\mu}_g = \frac{\eta_g + \tau_g \mu_g}{c_g + \tau_g}$$

$$\Sigma_g = \frac{1}{c_g + \tau_g} + [\gamma_g + \tau_g[\Sigma_g + \mu_g \mu_g{}^{tr}]] - \tilde{\mu}_g \tilde{\mu}_g{}^{tr}$$

where :

$$c_g = \sum_t c_g(t)$$

$$\eta_g = \sum_t c_g(t) x_t$$

$$\gamma_g = \sum_t c_g(t) x_t x_t{}^{tr}$$

The parameter τ_g is usually chosen to be constant. $c_g(t)$ is the *a posteriori* probability of the Gaussian *g* at time *t* conditioned on all acoustic observations $x_t = 1...T$.

This first processing step adapts all Gaussians parameters, however, only variances and gains Gaussians adaptations can effectivelly improve the modeling capabililies of the system for a specific test speaker. In fact, really adapting the Gaussians means compared to a specific speaker, only relying on training speakers requires a very large population of speakers. The selected training speakers cluster is then accordingly closer to the test speaker. By using a not so large training speakers population (120 speakers), the spectral variation caused by the inter-speaker variability in each speech unit is reduced, but the Gaussians means remains unadapted to the test speaker.

The second processing step consists in adapting the Gaussians means of the acoustic models resulting from the first adaptation step. This last adaptation is done by using the MLLR procedure [9]: the test data is decoded by using the reduced variances acoustic models (the *a prior* adapted models). Then the resulting frame/state alignement is used to estimate a global linear tranformation, which is applied to the Gaussians means of the *a priori* adapted acoustic models. We term this second step adaptation the *a posteriori* adaptation.

1.3 – Speaker clustering

To perform the *a priori* adaptation we need to find a subset of the training speakers who are the closest to the test speaker. This is done with the LIA speaker recognizer, AMIRAL [10], based on Gaussian Mixture Models (GMM). For this task we used a GMM with 128 Gaussians for each training speaker. The system compares all the training speakers to the test speaker, and then order these speakers from nearest to farest. So, the transcribed training data of the nearest speakers are selected to adapt the speaker-independent acoustic models with the MAP procedure.

Another strategy to select the training speakers based on HMM instead of GMM was tested. Firstly, we constructed 120 training-speaker dependent HMMs. However, the data available from each training speaker are usually not sufficient to obtain robust estimations of the speaker dependent model parameters. So, we used MAP procedure [6] to adapt the speaker-independent models to each training speaker, and hence obtain 120 HMMs representing each of the training speakers. It was then required to find the subset of the closest training speakers to the test speaker. The test data are decoded using a speaker-independent system leading to frame/state alignment. Then the acoustic likelihood of test data, conditioned on this alignment, is computed using each of training speaker-dependent HMM. The top speakers are then selected as the acoustically closest training speakers to the test speaker.

The two strategies give always the same five first speakers. The experiments in this paper is performed using the GMM based strategy.

1.4 – Experimental results

In this section, we present the results of several recognition experiments. These experiments were conducted using SPEERAL [11], a large vocabulary speech recognition system, developped at the LIA. The lexicon size is about 20k words with 3.6% out-of-vocabulary words. This system uses a trigram language model. The baseline system is speaker and gender independent. The acoustic model contains 38 phonemes. Each phoneme is 3-state left-to-right context-independent CDHMM (Continuous Density HMM). Each state is a mixture of 64 Gaussians. The signal speech is parameterized using 13 coefficients, 12 mel-warped cepstral coefficients plus energy. The first and second order derivatives parameters are also used.

To estimate the acoustic and linguistic models, we have used a training data extracted from Bref [13], with 120 speakers (66 females and 54 males). The training data contain 66.5k sentences. The test data were provided for ARC B1 of AUPELF [12], with 20 speakers and 299 sentences. The sentences are articles published in the french newspaper “Le

Monde". The *a priori* adaptation is performed by using the 5 closest training speakers to the test speaker. This adaptation is done with the MAP procedure. In the Table 1, we term this adaptation *Adapt. 1*. The *a posteriori* adaptation is performed by using the test signal with the MLLR procedure. In the Table 1, we term this adaptation *Adapt 2*. Both adaptations will be compared to the MLLR-only one, applied on the test data. We term this last adaptation *Adapt. 3*. For MLLR, we used 1 global linear transformation with an *offset*.

	Baseline	**Adapt. 1**	**Adapt. 2**	**Adapt. 3**
WER	26.2	25.4	22.4	24.9

Table 1: *Word Error Rate (%) Comparisons: Adapt. 1:a priori adaptation with MAP, Adapt. 2: a posteriori adaptation with MLLR on models obtained in Adapt. 1, Adapt. 3: MLLR on the baseline acoustic models.*

We can see that the *a priori* adaptation using MAP (adapt. 1 in Table 1) doesn't improve significantly the word error rate (only 3% relative gain with respect to the baseline system). However, this step is important because its conjunction with the MLLR procedure leads to a relative gain of 15% with respect to the baseline system (compare adapt. 2 and adapt. 1). This fact is due to a smaller variance of acoustic models in *a priori* adaptation (MAP). Then the MLLR better maps the means of acoustic Gaussians model to match the test data.

The word error rates (WER) by speaker shows that there is a relative large variation of recognition improvements with respect to test speakers. For example, the WER for one speaker was 41.1% without adaptation, 40.5% after the *a priori* adaptation and finally 27.9% after the *a posteriori* adaptation (1.5% relative gain by *a priori* adaptation compared to 32% by *a priori* and *a posteriori* adaptation). For the complete test, the average relative gain is about 15%. And, if some speakers were not improved, the recognition for these speakers is not degraded. This gain variation between test speakers can be explained by the fact that some test speakers miss signicantly close training speakers, with respect to the whole training speakers set. This problem should be solved by using a larger population of speakers.

In our experiments, the relative gain obtained by using the MLLR (1 global transformation) with respect to the baseline system is about 5% (from 26.2% to 24.9%). This gain is 3 times smaller than the one obtaind by conjunction of the *a priori* and *a posteriori* adaptations. The relative gain obtained by the *a priori* and *a posteriori* adaptations with respect to the conventional MLLR is about 10%.

2 – STRUCTURAL ADAPTATION USING MAP AND GAUSSIANS MERGING TECHNIQUE ADAPTATION

In this section, the SMAPGM (Structural Adaptation using MAP and Gaussians Merging technique) will be presented. This technique uses a classification tree and a new adaptation method.

2.1 – Introduction

Due to complex inter-speaker variabilities, the performance of speaker-independent (SI) large vocabulary continuous speech recognition systems still lags behind that of speakerdependent (SD) systems. Speaker-independent systems are typically constructed using speech samples collected from an as large as possible population of speakers [14].

Nevertheless, in the speaker-dependent case, the large amount of required training data for each test speaker reduces the utility and portability of such systems.

The main difficulty in speaker adaptation techniques is to adapt a large number of parameters with only a relative small amount of data. The MAP adaptation approach allows accurate estimation of HMM parameters for which enough adaptation data is available [6], and the unseen parameters are still unchanged. In this manner, the MAP approach leads to too much local adaptation. Hence the MAP approach can't be effective with relative small amount of adaptation data especially in unsupervised mode.

In order to reduce this problem, Shinoda and Lee proposed a structural maximum a posteriori (SMAP) approach [15], in which a hierarchical structure (tree) in the parameter space is assumed. The parameters transformation for each node in the tree are

estimated by using the MAP approach in which the a priori parameters are given by the parent node. The resulting transformation parameter, corresponding to each HMM parameter, is a combination of the transformation parameters at all higher levels. The weights in this combination depend on the amount of adaptation data at each node and on a fixed parameter.

In SMAP, the parameters estimation at each node in the tree is based on the assumption that the mismatch between the training and adaptation data is a Gaussian PDF.

The mean and the variance of this Gaussian mismatch PDF are estimated directly from the adaptation data by using the Maximum Likelihood criterion. In this manner, the estimation accuracy of the transformation parameters depends on the amount of the adaptation data. To avoid poor transformation parameters estimation accuracy due to an insufficiency of adaptation data we propose a new technique based on maximum a posteriori approach [6] and PDF Gaussian Merging. The basic idea behind this new technique is to estimate transformations which make the training acoustic models as close as possible to the test acoustic models rather than transformation maximizing the likelihood of the adaptation data. The test acoustic models are estimated using the MAP approach [6]. In this manner, even with very small amount of adaptation data, the parameters transformations are accurately estimated.

In this paper, like in SMAP [15], we assume that the models parameters are organized in tree containing all the Gaussian distributions. Each node in that tree represents a cluster of Gaussians. All the Gaussian distributions of a given cluster/node share a simple common affine transformation (diagonal matrix plus *offset*) compensating the mismatch between training and test conditions.

To estimate this affine transformation, we propose a new technique based on a Gaussian distributions merging and the standard MAP adaptation. This new technique is very fast and allows a good adaptation for both means and variances even with small amount of adaptation data in unsupervised mode. At each node, the transformation is obtained by combining three kinds of information: the adaptation data, the parameters transformation at the parent node and the parent node adapted parameters.

Section 2.2 presents the whole adaptation process proposed in this work: the adaptation process in a given node in the tree, the combination of the mismatch information at different tree layers, the merging procedure, and the tree construction. Section 2.3 shows results for several recognition experiments in a large vocabulary task framework.

2.2 – Adaptation process

Gaussian distributions. The first step in the adaptation process is to build a classification tree structure representing the set of Gaussian distributions. Each node in the tree represents a subset of Gaussians and the root node represents the whole set. Let ν denote one node in the classification tree, and $G_\nu = \{g_{m_\nu}, m_\nu = 1...M_\nu\}$ be the subset of Gaussian distributions associated to the node $\nu : g_{m_\nu} = N(\mu_{m_\nu}, \Sigma_{m_\nu})$. In the following paragraphs, we describe the adaptation process for a node ν and show the strategy for combining information at different layers.

2.3 – Adaptation Process in a node

The goal of this work is to estimate for each node ν an affine transformation T_ν (diagonal matrix plus *offset*) shared by all Gaussian distributions in the subset G_ν. This affine transformation is then applied to only the distributions belonging to G_ν. Let $X = \{x_1, ..., x_t\}$ denote a given set of T observation vectors for parameters adaptation.

Let $\tilde{g}_{m_\nu} = N(\tilde{\mu}_{m_\nu}, \tilde{\Sigma}_{m_\nu})$ be the Gaussian obtained by adapting the Gaussian $g_{m_\nu} = N(\mu_{m_\nu}, \Sigma_{m_\nu})$ using the standard MAP adaptation (see Formulas).

Let $\tilde{G}_\nu$ be the subset of MAP adapted Gaussians in the node ν: $\tilde{G}_\nu = \{\tilde{g}_{m_\nu}, m_\nu = 1...M_\nu\}$. Let $\tilde{g}_\nu = N(\tilde{\mu}_\nu, \tilde{\Sigma}_\nu)$ and $g_\nu = N(\mu_\nu, \Sigma_\nu)$ be the two Gaussians obtained by merging into one all Gaussians in G_ν and $\tilde{G}_\nu$ respectively (see section 2.2).The affine transformation T_ν is then estimated as the one which matches the Gaussian g_ν to the Gaussian $\tilde{g}_\nu$. Each Gaussian $g_{m_\nu} = N(\mu_{m_\nu}, \Sigma_{m_\nu})$ is then adapted as follows :

$$\mu'_{m\nu} = \tilde{\Sigma}_{\nu}^{\frac{1}{2}} \Sigma_{\nu}^{-\frac{1}{2}} (\mu_{m\nu} - \mu_{\nu}) + \tilde{\mu}_{\nu}$$

$$\Sigma'_{m\nu} = \tilde{\Sigma}_{\nu} \Sigma_{\nu}^{-1} \Sigma_{m\nu}$$

Where $\mu'_{m\nu}$ and $\Sigma'_{m\nu}$ are the adapted parameters of $\mu_{m\nu}$ and $\Sigma_{m\nu}$ respectively. This adaptation procedure can be performed iteratively. We have shown experimentally that the likelihood of adaptation data increases at each iteration.

2.4 – Merging Process

The merging process is based on the merging of pairs of Gaussian distributions until we obtain a single Gaussian. In this work the merging of two Gaussians uses the minimum loss likelihood criterion. Let $G = \{g_1,...,g_n\}$ denote a set of Gaussians to be merged into one representing the set G. Firstly, we choose two Gaussians $g_i = N(\mu_i, \Sigma_i)$ and $g_j = N(\mu_j, \Sigma_j)$ in G. Let c_i and c_j denote their associated counts. The Gaussian $g = N(\mu, \Sigma)$ obtained by merging g_i and g_j is given by the classic formula:

$$\mu = \frac{c_i \mu_i + c_j \mu_j}{c_i + c_j}$$

$$\Sigma = \frac{c_i \Sigma_i + c_j \Sigma_j + \frac{c_i \times c_j}{c_i + c_j} (\mu_i - \mu_j)(\mu_i - \mu_j)^{tr}}{c_i + c_j}$$

The count c associated with the new Gaussian g is the sum of the two counts c_i and c_j associated with the two Gaussians g_i and g_j. The two Gaussians g_i and g_j in G are then replaced by the Gaussian g. We repeat this merging procedure until we obtain one Gaussian representing the set G. The initial count $c_{m\nu}$ associated to a Gaussian $g_{m\nu}$ is the sum over all observation vectors of the a posteriori probabilities: $c_m = \Sigma_t \gamma_{mvt}$.

2.5 – Adaptation Using Hierarchical Priors

In section 2.1. we have treated the problem of estimating an affine transformation T_ν associated to the node ν. The estimation of T_ν was based only on the Gaussians belonging to this node and their associated observation vectors. To estimate the transformation T_ν by using all Gaussians in the CDHMM and their associated observation vectors we use the adaptation with hierarchical priors.

Let $p(\nu)$ denote the parent node of ν. Let g_ν and $g_{p(\nu)}$ be the two Gaussians obtained by merging into one all the Gaussians in G_ν and $g_{p(\nu)}$ respectively (the original Gaussians in the node ν and $p(\nu)$). In the manner, let $\tilde{g}_\nu$ and $\tilde{g}_{p(\nu)}$ denote the Gaussians obtained by merging into one all the Gaussians in $\tilde{G}_\nu$ and $\tilde{G}_{p(\nu)}$ respectively (the MAP adapted Gaussians in the node ν and $p(\nu)$) (see section 2.1).

On one hand we merge the Gaussians g_ν and $g_{p(\nu)}$ to obtain one Gaussian $g_\nu^{p(\nu)} = N(\mu_\nu^{p(\nu)}, \Sigma_\nu^{p(\nu)})$ and on the other hand we merge the Gaussians $\tilde{g}_\nu$ and $\tilde{g}_{p(\nu)}$ to obtain one Gaussian $\tilde{g}_\nu^{p(\nu)} = N(\tilde{\mu}_\nu^{p(\nu)}, \tilde{\Sigma}_\nu^{p(\nu)})$. In this merging process the count associated to the Gaussians in the parent node $p(\nu)$ is a fixed parameter, and the count associated to the Gaussians in the node ν is the sum of the counts associated to all Gaussians in that node ($\Sigma_m c_{m\nu} = \Sigma_m \Sigma_t \gamma_{mvt}$). The affine transformation T_ν is then estimated as the one which matches the Gaussian $g_\nu^{p(\nu)}$ to the Gaussian $\tilde{g}_\nu^{p(\nu)}$. Each Gaussian $g_{m\nu} = N(\mu_{m\nu}, \Sigma_{m\nu})$ is then adapted as follows :

$$\mu'_{m\nu} = (\tilde{\Sigma}_\nu^{p(\nu)})^{1/2} (\Sigma_\nu^{p(\nu)})^{-1/2} (\mu_{m\nu} - \mu_\nu^{p(\nu)}) + \tilde{\mu}_\nu^{p(\nu)}$$

$$\Sigma'_{m\nu} = (\tilde{\Sigma}_\nu^{p(\nu)})(\Sigma_\nu^{p(\nu)})^{-1} \Sigma_m$$

Where $\mu'_{m\nu}$ and $\Sigma'_{m\nu}$ are the adapted parameters of $\mu_{m\nu}$ and $\Sigma_{m\nu}$ respectively. These adaptation formula are then used instead of equations 1 and 2. In this manner the resulting transformation parameter, corresponding to each parameter, is a combination of mismatch information at all levels. In this combination the weight for each level changes autonomously according to the amount of adaptation data.

2.6 – Construction of the tree structure

The use of the tree structure has been largely studied in the contextual acoustic units estimation framework [16].

In this work we have used a binary tree. We assumed that all Gaussians in a state of the CDHMM belong to the same class and the tree leaves represent the CDHMM states. Each node in the tree is a collection of states which are collections of Gaussians. For classification, each state is represented by one Gaussian obtained by merging allGaussians in that state. Hence, we construct a state classification tree using the loss likelihood minimization criterion for clustering. We used the *up to down* strategy as classification tree algorithm. Our classification tree algorithm is not optimal because, at each node with n states, we don't explore the 2^{n-1} two-cluster splits possible. Instead, we use an iterative procedure like k-means clustering with two centers.

2.7 – Experimental results

In this section, we present the results of several speech recognition experiments. These experiments were conducted using the same experimental set than in the previous section, expect that now the baseline system is gender-dependent with 3-states left-to-right context-dependent unit acoustic models.

In these experiments, we used two binary trees with six layers: one for the male acoustic models and the other for the female acoustic models. These classification trees are built once before the adaptation process. In the experiments, both mean vectors and covariances were adapted. All adaptation procedures were performed speaker per speaker in unsupervised mode.

We will call the proposed technique SMAPGM (Structural Adaptation usingMAP and Gaussians Merging technique).

In Table 1 we can see that the SMAPGM technique gives an average relative gain about 16% with respect to the baseline system. It should be noted that part of the improvements of MLLR and SMAPGM can be cumulated.

In fact, by performing SMAPGM after MLLR the relative cumulated gain is about 18% with respect to the baseline system and by performing MLLR after SMAPGM the relative cumulated gain is about 19.5%. In these experiments, we have noted that the effect of the proposed method is more significant for speakers with higher word error rates.

	Male	**Female**	**Average**
Base	21.2	21.0	21.1
SMAPGM	18.0	17.7	17.8
SMAPGM+MLLR	16.6	17.4	17.0
MLLR+SMAPGM	17.1	17.5	17.3

Table 1: *Word Error Rate (%) for gender-dependent speech recognizer with different speaker adaptation techniques. SMAPGM designates the proposed technique: structural adaptation using MAP and Gaussians Merging technique.*

We have performed the same experiments with a better lexicon and language model. The baseline word error rate becomes 19%. After SMAPGM adaptation, the word error rate was 16.3% (a relative gain of 14% with respect to the baseline system, instead of 16% with the first system). When SMAPGM is performed after MLLR, the word error rate comes down to 15.9% (a relative gain of 16% with respect of baseline system, instead of 19.5% with the first system). The relative gain obtained by using SMAPGM seems to be larger for the baseline system with higher word error rate. In order to compare SMAPGM with SMAP [8], we realized experiments under the same conditions (with the same tree with six layers). The SMAP adaptation leads to a word error rate of 17.3% (a relative gain of 9%, instead of 14% for SMAPGM adaptation, see Table 2).

	Average	**Relative gain**
Baseline	19.0	
SMAPGM	16.3	14.2
SMAP	17.3	8.9

Table 2: *Word Error Rate (%) and relative gain in regard of baseline system for gender-dependent speech recognizer with SMAP and SMAPGM adaptations*

3 – CONCLUSION

In this paper, we have presented two new method for speaker adaptation. Their effectiveness was confirmed by experiments in a large vocabulary speech recognition task: a relative gain of 15% with regard to the baseline system was obtained in the case of the first technique and a relative gain of 16% in the case of the second technique.

BIBLIOGRAPHIE

[1] T. Anastaskos, F. Kubala, J. Makhoul and R. Schwartz, “Adaptation to new microphones using tied-mixture normalization”, in *Proc. IEEE Int. Conf. Acoustics, Speech ans Signal Processing*, pp. 433-436, 1994.

[2] H. Eide and H. Gish, “A parametric approach to vocal tract lenght normalization”, in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 346-349, 1996.

[3] Y. Zhao, “An Acoustic-phonetic-based Speaker Adaptation Technique Improving Speaker independent Continuous Speech Recongnition”, *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 380-394, July 1994.

[4] M. Rahim and B-H. Juang, “Signal bias removal by maximum likelihood estimation for robust telephone speech recognition”, in *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 1, pp. 19-30, January 1996.

[5] T. Anastaskos, J. McDonough, R. Schwartz and J. Makhoul, ”A Compact Model for Speaker- Adaptive Training”, *Proc. ICSLP’96*, pp. 1137- 1140, Philadelphia, 1996.

[6] J.-L. Gauvain and C.-H. Lee, « Maximum *A Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains », in *IEEE Trans. on Speech and Audio Processing*, 2(2):291-298, April 1994.

[7] J. R. Bellegarda, P. V. de Souza, A. Nadas, D. Nahamoo, M. A. Picheny and L. R. Bahl, ” The Metamorphic Algorithm: A Speaker Mapping Approach to Data Augmentation”, *IEEE Trans. Speech andAudio Processing*, vol. 2, no. 3, pp. 413-420, July 1994.

[8] M. Padmanabhan, L. R. Bahl, D. Nahamoo and M.A. Picheny, ”Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition Systems”, in *IEEE Transactions on Speech and Signal Processing*, vol. 6, no. 1, pp. 71-77, January 1998.

[9] C.J. Leggetter and P.C.Woodland, ”Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models”, in *Computer Speech and Language*, pp. 171-185, 1995.

[10] C. Fredouille, J.-F. Bonastre and T. Merlin, ”AMIRAL: A Block-Segmental Multirecognizer Architecture for Automattic Speaker Recognition”, in *Digital Signal Processing*, 2000.

[11] P. Nocera, G. Linares, D. Massonié, L. Lefort, « Phoneme lattice based A* search algorithm for speech recognition », Sept. 2002, Brno, TSD2002

[12] J. Dolmazon, F. Bimbot, G. Adda, J. Caerou, J. Zeiliger, M. Adda-Decker, “Première campagne AUPELF d’évaluation des systèmes de Dictée Vocale”, *Ressources et ´evaluation en ingénierie des langues*, pp. 279-307, 2000

[13] L. F. Lamel et al., “BREF, a Large Vocabulary Spoken Corpus for French”, in *EuroSpeech’91*, Genoa, Sept. 1991.

[14] D. Matrouf, O. Bellot, P. Nocera, G. Linares, J.F. Bonastre, « A Priori and a posteriori Transformations for Speaker Adaptation in Large Vocabulary Speech Recognition Systems », *7th European Conference on Speech Communication and Technology*, Vol. II, p. 1245-1248, Aalborg DENMARK, Sept. 2001.

[15] K. Shinoda and C.-H. Lee, « Unsupervised adaptation using structural Bayes approach », *In Proc IEEE ICASSP*, Seattle, Washington, USA, 1998.

[16] R. Singh, B. Raj and R. M. Stern, « Automatic Clustering and Generation of Contextual Questions for Tied States in Hidden Markov Models », *In Proc IEEE ICASSP*, Phoenix, Arizona, USA, 1999.