STRUCTURAL SPEAKER ADAPTATION USING MAXIMUM A POSTERIORI APPROACH AND A GAUSSIAN DISTRIBUTIONS MERGING TECHNIQUE

Olivier Bellot, Driss Matrouf, Pascal Nocera, Georges Linares, Jean-Francois Bonastre

Laboratoire d'Informatique d'Avignon
LIA, Avignon, France
{olivier.bellot,driss.matrouf,pascal.nocera} @lia.univ-avignon.fr
{georges.linares, jean-francois.bonastre} @lia.univ-avignon.fr

ABSTRACT

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. Recently, a technique based on hierarchical structure and the maximum a posteriori criterion was proposed (SMAP). In this paper, like in SMAP, we assume that the acoustic models parameters are organized in a tree containing all the Gaussian distributions. Each node in that tree represents a cluster of Gaussian distributions sharing a common affine transformation representing the mismatch between training and test conditions. To estimate this affine transformation, we propose a new technique based on merging Gaussians and the standard MAP adaptation. This new technique is very fast and allows a good unsupervised adaptation for both means and variances even with small amount adaptation data. This adaptation strategy has shown a significant performance improvement in a large vocabulary speech recognition task, alone and combined with the MLLR adaptation.

1. INTRODUCTION

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 an as large as possible population of speakers [1]. 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.

Speaker adaptation techniques try to obtain near speaker-dependent performance with relatively small amounts of test-speaker specific data [1, 2, 3, 4, 5, 6]. The main difficulty in speaker adaptation techniques is to adapt a large number of parameters with only a small amount of data. In this framework, two approaches are commonly used, the maximum likelihood linear regression (MLLR) [4] and the maximum a posteriori (MAP) [7]. On one hand, the MLLR adaptation gives significant improvement with a relatively small amount of adaptation data, but the recognition performance does not improve much when the amount of data increases. This is because the transformation is too global (the number of free parameters is too small).

On the other hand the MAP adaptation approach allows accurate estimation for the acoustic-unit models for which adaptation data is available [7]. However, as unseen units are still unchanged, MAP 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 [8], in which a hierarchical structure (tree) in the parameter space is assumed. The parameters transformation for each node in the tree is 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 this paper, like in SMAP, 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 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, the tree construction and a comparison between the proposed technique and SMAP. Section 3 shows results for several recognition experiments in a large vocabulary task framework. The last section (4) is dedicated to some conclusions, comments and perspectives concerning our new acoustic model adaptation technique.

2. ADAPTATION PROCESS

In this work, we address the problem of parameters adaptation in continuous-density HMM (CDHMM) based speech recognizers. Here we focus on the adaptation of the CDHMM 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 $G_\nu = \{g_{m_\nu}, m_\nu = 1 \dots M_\nu\}$ be the subset of Gaussian distributions associated to the node ν: $g_{m_\nu} = N(\mu_{m_\nu}, \Sigma_{m_\nu})$. In the following paragraphs, we describe the adaptation process for a node ν, and show the strategy for combining information at different layers.

2.1. 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 $X = \{x_1, x_2, \dots, x_T\}$ 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:

$$\begin{aligned} \tilde{\mu}_{m_\nu} &= \frac{a_{m_\nu} + \tau_{m_\nu}\mu_{m_\nu}}{b_{m_\nu} + \tau_{m_\nu}} \\ \tilde{\Sigma}_{m_\nu} &= \frac{c_{m_\nu} + \tau_{m_\nu}(\Sigma_{m_\nu} + \mu_{m_\nu}\mu_{m_\nu}^{tr})}{b_{m_\nu} + \tau_{m_\nu}} - \tilde{\mu}_{m_\nu}\tilde{\mu}_{m_\nu}^{tr} \end{aligned}$$

where,

$$\begin{aligned} a_{m_\nu} &= \sum_t \gamma_{m_\nu t} \\ b_{m_\nu} &= \sum_t \gamma_{m_\nu t} x_t \\ c_{m_\nu} &= \sum_t \gamma_{m_\nu t} x_t x_t^{tr}, \end{aligned}$$

and $\gamma_{m_\nu t}$ is the a posteriori probability of the Gaussian g_{m_ν} at time t, conditioned on all acoustic observations $x_{t=1\dots T}$. This probability is obtained by using the speech recognizer based on the original acoustic models. The parameter τ_{m_ν} is usually chosen to be constant.

Let $\tilde{G}_\nu$ be the subset of MAP adapted Gaussians in the node ν: $\tilde{G}_\nu = \{\tilde{g}_{m_\nu}, m_\nu = 1 \dots M_\nu\}$. Let $g_\nu = N(\mu_\nu, \Sigma_\nu)$ and $\tilde{g}_\nu = N(\tilde{\mu}_\nu, \tilde{\Sigma}_\nu)$ be the two Gaussians obtained by merging into one all Gaussians in G_ν and $\tilde{G}_\nu$ respectivelly (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,

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

$$\Sigma'_{m_\nu} = \tilde{\Sigma}_\nu \Sigma_\nu^{-1} \Sigma_{m_\nu} \quad (2)$$

Where μ'_{m_ν} and Σ'_{m_ν} are the adapted parameters of μ_{m_ν} and Σ_{m_ν} respectively.

This adaptation procedure can be performed iteratively. We have shown experimentally that the likelihood of adaptation data increases at each iteration.

2.2. 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 $G = \{g_1, g_2, \dots, g_n\}$ 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:

$$\begin{aligned} \mu &= \frac{c_i\mu_i + c_j\mu_j}{c_i + c_j} \\ \Sigma &= \frac{c_i\Sigma_i + c_j\Sigma_j + \frac{c_i * c_j}{c_i + cj}(\mu_i - \mu_j)(\mu_i - \mu_j)^{tr}}{c_i + c_j} \end{aligned}$$

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_ν} associated to a Gaussian g_{m_ν} is the sum over all observation vectors of the a posteriori probabilities: $c_{m_\nu} = \sum_t \gamma_{m_\nu t}$.

2.3. 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 tranformation 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 ($\sum_m c_{m_\nu} = \sum_m \sum_t \gamma_{m_\nu t}$). 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,

$$\begin{aligned} \mu'_{m_\nu} &= (\tilde{\Sigma}_\nu^{p(\nu)})^{\frac{1}{2}} (\Sigma_\nu^{p(\nu)})^{-\frac{1}{2}} (\mu_{m_\nu} - \mu_\nu^{p(\nu)}) + \tilde{\mu}_\nu^{p(\nu)} \\ \Sigma'_{m_\nu} &= (\tilde{\Sigma}_\nu^{p(\nu)}) (\Sigma_\nu^{p(\nu)})^{-1} \Sigma_{m_\nu} \end{aligned}$$

Where μ'_{m_ν} and Σ'_{m_ν} are the adapted parameters of μ_{m_ν} and Σ_{m_ν} respectively. These adaptation formula are then used instead of equations 1 and 2. In this manner the resulting transformation parameter, corresponding to each paramater, 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.

2.4. Construction of the tree structure

The use of the tree structure has been largely studied in the contextual acoustic units estimation framework [9]. 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 all Gaussians in that state. Hence, we construct a state classification tree using the loss likelihood minimisation 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.

2.5. Comparison with SMAP

The common features between the proposed technique in this work and the SMAP method are first, the use of a tree structure to classify the distributions and second, the assumption of a simple transformation in a given node. The main difference between the two techniques is that in the proposed technique the standard MAP approach is used in HMM parameters space in a given node whereas in SMAP the MAP paradigm is used in transformation space using the parent node transformation as an a priori. Another difference is that our approach uses two fixed combination weight parameters, one between different layers and the other for standard MAP in a given node, whereas SMAP uses only one combination weight parameter between different layers. We have not yet implement the SMAP technique, but experimental comparison between the two approaches will be interesting.

3. EXPERIMENTAL RESULTS

In this section, we present the results of several speech recognition experiments. These experiments were conducted using SPEERAL [12], a large vocabulary speech recognition system, developed 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 gender dependent with 3-state left-to-right context-dependent unit acoustic models. Each state is a mixture of 64 Gaussians. The speech signal is parameterized using 39 coefficients: 12-mel warped cepstral coefficients plus energy and their first and second order derivative parameters. The cepstral mean removal and the normalization of the variance have been performed sentence by sentence.

To estimate the acoustic models we have used a training data extracted from Bref [10], with 120 male and female speakers. These acoustic models are then adapted by using the MAP algorithm with the data from 54 males to create speaker-independent male models, and with the data from 66 females to create speaker-independent female models. These gender-dependent acoustic models are used as initial models for adaptation. A diagonal matrix covariance was used for each mixture Gaussian component. The test data comes from ARC B1 of AUPELF, with 20 speakers and 299 sentences [11].

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 Structural Adaptation using MAP and Gaussians Merging techniques: SMAPGM.

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.

Adaptation techniques	Word Error (%)		
	Male	Female	Avg
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 recogniser 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 16.6% (a relative gain of 13% 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.

4. CONCLUSION

We have presented a new unsupervised acoustic model adaptation technique based on MAP adaptation and merging Gaussian distributions. Its effectiveness was confirmed by experiments in a large vocabulary speech recognition task: a relative gain of 16% with regard to the baseline system was obtained. The conjunction of the proposed method with MLLR leads to a relative gain of 19.5%.

Several problems remain to be investigated. First, the depth of the classification tree used for adaptation, which can be estimated depending on the amount of adaptation data (depth was fixed in the experiments presented in this paper). Second, the combination weights used to combine mismatch information between a given node and its parent. Third, the fixed combination parameter used for MAP adaptation in a given node. At last, making a tree structure that well represents the embedded structure in the acoustic space should be further studied. In a next work, we will evaluate and compare the proposed method with the SMAP adaptation technique.

5. REFERENCES

[1] 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.

[2] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition",*Speech Communication*, 25:29-47, 1998.

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

[4] 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.

[5] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined tranformation and Bayesian methods", *IEEE Trans. on Speech and Audio Processing*, 4(4), July 1996.

[6] O. Siohan, T. A. Myrvoll and C.-H Lee, "Structural Maximum a Posteriori Linear Regression for Fast HMM Adaptation", *Workshop on automatic speech recognition: challenges for new millenium*, Paris, Sept 2000.

[7] 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.

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

[9] 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.

[10] L. F. Lamel et al., "BREF, a Large Vocabulary Spoken Corpus for French", in *EuroSpeech'91*, Vol. 2, pp.505-508, Italy, Sept. 1991

[11] 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 évaluation en ingénierie des langues*, pp. 279-307, 2000

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