ADAPTING PSN RECOGNITION MODELS TO THE GSM ENVIRONMENT BY USING SPECTRAL TRANSFORMATION

Thierry Soulas, Chafic Mokbel, Denis Jouvet and Jean Monné
France Télécom - CNET / LAA / TSS / RCP
2 av. Pierre Marzin, 22307 Lannion cedex, France
e-mail : mokbel@lannion.cnet.fr

ABSTRACT

In this work, environment adaptation is studied in order to transform PSN speaker independent isolated words HMM to the GSM environment. LMR transformations associated with groups of HMM densities are used to adapt the densities. Both mean vectors and covariance matrices of the densities are adapted.

It has been shown that few amount of GSM data are sufficient to transform the PSN HMM in order to match the GSM environment and to achieve performances equivalent to those of an HMM trained with large amount of GSM data. The number of groups of Gaussian densities seems to have small influence on the results. However, the minimum number of groups depends on the vocabulary size.

Finally, this technique is compared to the Bayesian adaptation and the results show that similar performances can be obtained with both methods.

1. INTRODUCTION

The mismatch between test and training environments seriously degrades the speech recognition performances. A microphone replacement or a different transmission environment are typical cases of situations where an adaptation of the model parameters is necessary to match the new recognition conditions. Environment adaptation methods have been widely studied in recent years [1],[2].

Spectral transformation and/or Bayesian adaptation techniques may be used to get the new estimate of the model parameters, based on few amount of data from the new (target) environment. Bayesian adaptation [3] consists in re-estimating the model parameters, within the classical "Estimate Maximize" EM algorithm, using a priori distributions of these parameters and the few amount of the adaptation data. The adaptation may also be achieved by a spectral transformation of the original environment acoustic space. This could be done in two ways: by transforming the original training data and then retraining the model with the adapted data [4], or by directly applying the transformation on the model parameters [5]. It is obvious that applying the spectral transformation to directly adapt the model parameters is more attractive. However, this supposes that some hypotheses must be assumed on the HMM sub-processes in the target space.

In order to use spectral transformation to adapt the HMM parameters, two problems remain : the idecntification of the transformation function and the estimation of the function parameters. Several functions have already been used to perform spectral transformation. "Linear Multiple Regression" (LMR) is a popular one [1][2][6][7]. Several LMR functions can be used, each one been associated to a region of the acoustical space [6][7].

Two main approaches exist to estimate the HMM parameters. The first approach is based on the "*Minimum Mean Square Error*" (MMSE) criterion between aligned vectors from the original and target spaces. The second approach makes use of the initial HMM model in the adaptation procedure [6][7][8][9]. For speaker adaptation, interesting results are obtained in [6] and [7], where the LMR is used to update mean vectors of the original model Gaussian densities. The LMR parameters may be estimated, within the segmental-EM [10] algorithm framework, in order to get the "*Maximum Likelihood*" (ML) of the adapted model on the target data. Moreover, to improve the effect of the transformation, several regression matrices are applied, each one being associated with several state distributions grouped by groups [6][7]. The set of groups is generated, by merging HMM's Gaussian densities.

In this work, the preceding approach is extended in order to adapt both of Gaussian mean vectors and covariance matrices using LMR transformations. The Gaussian densities of the original model are also grouped and a LMR is associated with each group. In order to define the groups, a binary tree is constructed by merging couples of closest Gaussian densities [11] going from the leaves to the root. This method is used to adapt a speaker independent ASR system trained in the PSN network environment to the "Global System Mobile" (GSM) network environment, using a few amount of GSM data.

This paper is organized as follows. In section 2, the adaptation algorithm is detailed. In section 3, experiments and results on a PSN/GSM database are reported. Section 4 presents the conclusions.

2. ADAPTATION ALGORITHM

Using spectral transformation to adapt the model parameters to a new environment supposes that the original model represents the main acoustical information relative the vocabulary words, and that few amount of data from the new environment are sufficient to estimate the parameters of a function that transforms the model parameters to better match the new environment.

Supposing that LMR is a suitable transformation, two different approaches exist to estimate its parameters, as noted in the introduction. Given some data from the target environment and the reference environment training data, feature vectors may be coupled in order to directly estimate the transformation parameters based on the MMSE criterion. Coupling speech frames is generally a source of error, especially when dealing with speaker independent data. The direct estimation of LMR parameters using the reference HMM, in the framework of the EM algorithm, offers a smart solution to the problem. Actually, all the reference conditions are merged in the HMM parameters and the coupling between adaptation frames and the HMM densities is seen as an incomplete data problem.

In the proposed adaptation scheme, groups of HMM Gaussian densities are defined. All the densities of a group share the same LMR transformation. This provides robustness to lack of adaptation data for some densities and a smoothing of the global adaptation model [6][7]. To do this sharing, the densities of the HMM, are recursively merged in a binary tree structure, the closest two densities being merged at each iteration. The distance criterion is based on the likelihood loss after the merge operation [11]. The desired number of groups of densities, is arbitrary chosen, it determines which level is to be considered in the tree. Once the groups of densities and the associated LMRs are defined, two problems remain.

1. How to determine the parameters of the adapted distributions given the LMRs parameters?
2. In the "Maximize" step of the segmental EM algorithm, how to estimate the LMRs parameters ?

2.1 Adapted Distributions Parameters

In order to get the adapted Gaussian parameters the first and second order moments must be computed. Given a PSN model distribution with μ and Σ as mean vector and covariance matrix, let us consider the LMR associated with this distribution. This LMR is defined with the regression matrix **A** and the bias vector **b**. After adaptation, the new mean vector and covariance matrix are obtained by :

$$\hat{\mu} = \mathbf{A} \,.\, \mu + \mathbf{b}$$

$$\hat{\Sigma} = \mathbf{A} \,.\, \Sigma \,.\, \mathbf{A}^{T}$$

Finally, by substituting the parameters, the adapted pdf of the sub-process corresponding to the state e, becomes :

$$p(\mathbf{y}/e)=(2\pi)^{-q/2}.\|\mathbf{A}.\Sigma.\mathbf{A}^{T}\|^{-1/2}.e^{-\frac{1}{2}(\mathbf{y}-\mathbf{A}.\mu-\mathbf{b})^{T}.\mathbf{A}^{-T}.\Sigma^{-1}.\mathbf{A}^{-1}.(\mathbf{y}-\mathbf{A}.\mu-\mathbf{b})}$$

where **y** is a feature vector in the new environment and q is the dimension of the feature space.

Looking to the resulting pdf, it can be seen that the result is similar to the application of the (inverse) LMR on the feature vector **y** (i.e. $\mathbf{A}^{-1}.\mathbf{y} - \mathbf{A}^{-1}.\mathbf{b}$).

2.2 Estimation of the LMR Parameters

At each iteration of the segmental EM algorithm, the adaptation data vectors are aligned on the adapted model densities, given the LMRs parameters obtained at the end of the preceding iteration. From this alignment, the vectors, associated with a group of densities, help to estimate the corresponding LMR parameters using the ML criterion. The logarithm of the likelihood of the adaptation data, relative to a group C, given the LMR parameters **A** et **b** is defined as follows :

$$\log[p_{\lambda,c}(\{\mathbf{y}\})]=\underline{cte\ }+\sum_{l=1}^{L}\sum_{y_t\in d_l}\left\{-\frac{1}{2}\log\|\mathbf{A}.\Sigma_l.\mathbf{A}^{T}\|\right.$$
$$\left.-\frac{1}{2}(\mathbf{y}_t-\mathbf{A}.\mu_l-\mathbf{b})^{T}.\mathbf{A}^{-T}.\Sigma_l^{-1}.\mathbf{A}^{-1}.(\mathbf{y}_t-\mathbf{A}.\mu_l-\mathbf{b})\right\}$$

where λ is the HMM, and d_l is the l-th density of the group C with parameters μ_l and Σ_l.

Deriving this equation with respect to $\mathbf{A}^{-1}$ and **b** (covariance matrix being defined positive), leads to the following equations :

$$\mathbf{b} = \mathbf{A}.\left[\sum_{l=1}^{L} n_l.\Sigma_l^{-1}\right]^{-1}.\sum_{l=1}^{L} n_l.\Sigma_l^{-1}.[\mathbf{A}^{-1}.\overline{\mathbf{y}}_l-\mu_l]$$

and,

$$\sum_{l=1}^{L} n_l\left\{\mathbf{A} - \frac{1}{n_l}\Sigma_l^{-1}\mathbf{A}^{-1}\sum_{y_t\in d_l}\mathbf{y}_t\mathbf{y}_t^{T} + \Sigma_l^{-1}\left[\sum_{k=1}^{L} n_k.\Sigma_k^{-1}\right]^{-1}\sum_{p=1}^{L} n_p.\Sigma_p^{-1}\mathbf{A}^{-1}\overline{\mathbf{y}}_p\overline{\mathbf{y}}_l^{T}\right.$$
$$\left.-\ \Sigma_l^{-1}.\left[\sum_{k=1}^{L} n_k.\Sigma_k^{-1}\right]^{-1}.\ \sum_{p=1}^{L} n_p.\Sigma_p^{-1}\mu_p\overline{\mathbf{y}}_l^{T} + \Sigma_l^{-1}.\ \mu_l\overline{\mathbf{y}}_l^{T}\right\} = 0$$

where n_l is the number of vectors associated with the l-th density and $\overline{\mathbf{y}}_l$ represents the mean of these vectors.

Unfortunately no direct solution can be easily derived for these equations if the covariance matrices are non-diagonal. In the CNET HMM-based system (PHIL90), the sub-processes have Gaussian pdfs with diagonal covariance matrices. Making the hypothesis of diagonal regression matrix **A**, a diagonal element a can be obtained by solving the quadratic equation (in 1/a) :

$$\frac{1}{a^2}\cdot\left[\frac{\left(\sum_{l=1}^{L} n_l\bar{y}_l/\sigma_l^2\right)^2}{\sum_{l=1}^{L} n_l/\sigma_l^2} - \sum_{l=1}^{L}\sum_{y_t\in d_l} y_t^2/\sigma_l^2\right]$$

$$+\frac{1}{a}\cdot\left[\sum_{l=1}^{L} n_l\mu_l\bar{y}_l/\sigma_l^2 - \frac{\left(\sum_{l=1}^{L} n_l\bar{y}_l/\sigma_l^2\right)\left(\sum_{k=1}^{L} n_k\bar{y}_k/\sigma_k^2\right)}{\sum_{l=1}^{L} n_l/\sigma_l^2}\right] + N = 0$$

where N is the total number of vectors for the group C, σ_l and μ_l are the mean and standard deviation for the dimension considered and $\bar{y}_l$ is the element, corresponding to the dimension considered, of the mean of the vectors associated with the l- th density.

Given the diagonal element, the corresponding bias can be easily found using :

$$b = \frac{\sum_{l=1}^{L} n_l[\bar{y}_l - a.\mu_l]/\sigma_l^2}{\sum_{l=1}^{L} n_l/\sigma_l^2}$$

Using these equation, the LMRs parameters can be estimated within the segmental-EM algorithm as well as the adapted model parameters.

3. EXPERIMENTS AND RESULTS

This approach is experimented on a speaker independent isolated word database collected in both the PSN (≈1000 calls) and GSM (≈1300 calls) network environments. The vocabulary is formed of 10 digits and 40 command words. The automatically detected vocabulary words were validated by a human listener. Only the correctly detected words are used in these experiments. Two sets of experiments are conducted function of the vocabulary : one on the digit vocabulary, and the other on the whole vocabulary of 50 words.

Left-right HMMs with 30 states are used to model the vocabulary words, and silence models are placed on both sides of the vocabulary models to avoid precise detection of the words to recognize. A simple Gaussian pdf with a diagonal covariance matrix is associated to each HMM state.

Two parts are distinguished for both PSN and GSM databases. For the PSN database, one part is used for training, the other one for testing. For the GSM database, the target environment, the first part constitutes the adaptation data and the second part corresponds to the testing data.

Figure 1 shows the recognition error rates obtained on GSM and PSN test databases for the digit vocabulary, function of the number of groups of densities used for the spectral transformation. A GSM error rate reduction of 45% with respect to a training with PSN data is observed. The performances obtained when training in the GSM network condition are nearly reached. Based on figure 1, it seems that the choice of the number of groups is not critical. 200 groups of densities is a good compromise for the digits database.

Figure 2 shows the recognition error rates on the GSM speech data as a function of the number of density groups and of the amount of GSM data used for the adaptation. The results show that less adaptation data can be used without deteriorating the recognition performances. Actually, the same error rates are obtained when only 1/16 from the GSM training database is used for the adaptation.

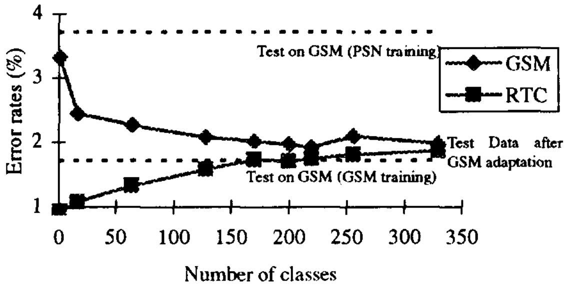

Fig. 1 - GSM adaptation using regression groups.

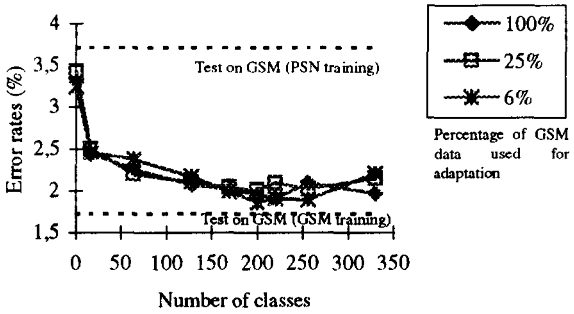

Fig. 2 - GSM adaptation with little amount of data.

Figure 3 shows the results obtained with experiments on the whole vocabulary (50 words) when using different amount of GSM data to adapt the PSN model. These results are compared to the ML training of a HMM with the same amount of GSM training data alone or combined with the whole set of PSN training data. Looking to these results it appears clearly that for large amount of GSM data, GSM training, combined PSN & GSM training and adaptation with linear regression (number of groups sufficiently high to consider all the conditions in the adaptation data) produce equivalent performances. For few amount of GSM adaptation data, the adaptation by LMR achieves high performances with respect to the classical training techniques. LMR adaptation is compared to the Bayesian adaptation technique described in [3]. It seems that equivalent performances are achieved with both techniques for this task.

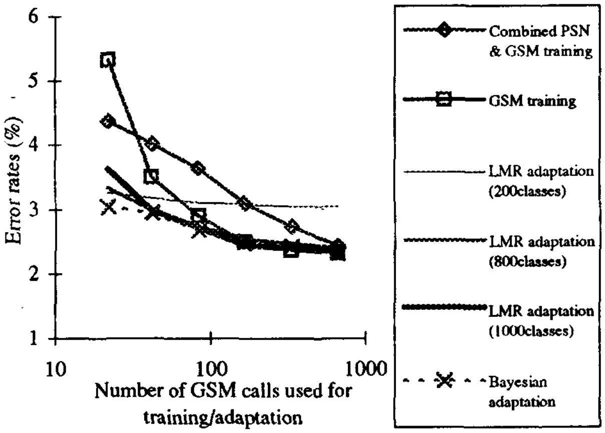

Fig. 3 - Comparing LMR adaptation, Bayesian adaptation and classical training on the whole vocabulary.

4. CONCLUSION

This paper presents an environment adaptation technique based on LMR. Linear multiple regressions are used in order to adapt the parameters of the HMM densities, both mean vectors and covariance matrices. HMM densities are grouped and all the densities of a given group share the same LMR. On the basis of a few amount of adaptation data, the transformation parameters and the adapted mean vectors and covariance matrices are estimated in a *Maximum Likelihood* sense within the framework of the segmental EM algorithm.

Experiments are conducted in order to adapt an isolated words speaker independent HMM trained with PSN data to the GSM network environment. The results show that this adaptation allows to reach the performances of a GSM training. Choosing the number of the groups of densities (and associated transformations) to be great or equal to the 1/3 of the number of densities in the model seems to be a good compromise in our experiments.

It has been proved that, with this technique, a little amount of adaptation data (~50 calls) is sufficient to get a reliable adapted model. The adaptation technique keeps, from the original HMM, all the speaker independent acoustical information relative to the vocabulary words and transform it to better match the target environment on the basis of few amount of adaptation data. Finally, when compared to Bayesian adaptation similar results are obtained.

References

[1] K. Takagi, H. Hattori and T. Watanabe, "Rapid Environment Adaptation For Robust Speech Recognition," Proc. ICASSP 95, Vol. 1, pp 149-152, May 1995.

[2] J. Takahashi, S. Sagayama: "Telephone Line Characteristic Adaptation using Vector Field smoothing Technique," Proc. ICSLP 94, Vol. 3, S18-2, pp 991-994, September 1994.

[3] C.G. Miglietta, C. Mokbel, D. Jouvet and J. Monné (1996), ''Bayesian Adaptation of Speech Recognizers to Field Speech Data,'' Proc. ICSLP96, pp. 917-920, 1996.

[4] C. Mokbel and G. Chollet, "Word Recognition in the Car: Speech Enhancement / Spectral Transformations," Proc. ICASSP 91, S14.12, 1991.

[5] C. Mokbel and G. Chollet, ''Automatic Word Recognition in Cars,'' IEEE Trans. on SAP, Vol. 3, n° 5, September 1995, pp. 346-356.

[6] C. J. Legetter, P. C. Woodland: "Speaker Adaptation of Continuous Density HMMs using Multivariate Linear Regression," Proc. ICSLP 94, Vol. 2, pp 451-454, September 1994.

[7] C. J. Legetter, P. C. Woodland: "Speaker Adaptation of HMMs using Linear Regression," Technical Report CUED/F-INFENG/TR.181, Cambridge University Engineering Department, June 1994.

[8] K. Shinoda, K. Iso, T. Watanabe: "Speaker Adaptation for Demi-Syllabe Based Continuous Density HMM," Proc. ICASSP 91, S13.7, pp 857-860, 1991.

[9] K. Ohkura, H. Ohnishi, M. Iida: "Speaker Adaptation Based On Transfer vectors of Multiple Reference Speakers," Proc. ICSLP 94, Vol 2, S09-7, pp 455-458, September 1994.

[10] A. Dempster, N. Laird, D. Rubin: "Maximum Likelihood from Incomplete Data via the EM algorithm," J. Roy. Stat. Soc., Vol. 39, no. 1, pp. 1-38, 1977.

[11] D. Jouvet, L. Mauuary and J. Monné, "Automatic Adjustments of the Structure of Markov Models for Speech Recognition Applications," Proc. EuroSpeech 91, pp 927-930, 1991.