PARAMETER TYING FOR FLEXIBLE SPEECH RECOGNITION

J. Simonin, S. Bodin, D. Jouvet & K. Bartkova

France Télécom - CNET - LAA/TSS/RCP
Technopole Anticipa
2, Avenue Pierre Marzin, 22307 Lannion - FRANCE
e-mail: simonin@lannion.cnet.fr

ABSTRACT

This paper presents two parameter tying techniques which enable a trade-off between computational cost and recognition performances of a speaker independent flexible speech recognition system working over the telephone network. Parameter tying is conducted at phonetic and acoustic levels.

At the phonetic level, allophone and triphone based phonetic modeling are used simultaneously to achieve the best trade-off between computational cost and recognition performances. This decreases error rate with a controlled computational cost as compared to an allophone modeling.

At the acoustic level, the tying is performed by clustering the Gaussian densities of mixture distributions. After clustering, a particular density may be use by several distribution. This allows the total number of Gaussian densities to be divided by two while improving the recognition performances.

1. INTRODUCTION

Accurate contextual representation produces an efficient speaker independent flexible speech recognition system. This flexible, i.e. task independent, system requires an important number of parameters. Parameter tying techniques may be a solution where a good trade-off between recognition performances and number of parameters is obtained.

Efficient speech recognition techniques have to take into account acoustic variations due to phonetic contexts. At the phonetic level, a well-known contextual representation may be achieved with a triphone modeling. However, a reliable estimate of some context-dependent units may be difficult. Indeed, some units insufficiently represented in the training corpus are badly estimated. Here, parameter tying aims at robust context-dependent models with a reduced number of parameters [1]. To do this, allophone modeling merges contexts which have similar effect on the acoustic realization of a given sound [2].

At the acoustic level, the parameter tying applies to emission distributions parameters. The system may use tied mixtures leading to a Semi-Continuous Hidden Markov Model (SCHMM) system [3][4]. Another way for tying acoustic parameters is to cluster the Gaussian components of multi-Gaussian distributions[5][6]. Increasing the number of densities improves the recognition performances if sufficient training data are available to estimate the density parameters. However, it also increases the computational cost as more parameters are used. So, Gaussian functions merging methods with clustering technique [7] are useful to optimize the amount of parameters regarding both computational cost and the amount of the available training data.

This study aims to employ parameter tying techniques to control the amount of model parameters without penalizing the word error rate. Two parameter tying techniques are developed and evaluated in the frame of a speaker independent flexible speech recognition system over the telephone network. At the phonetic level, allophones and triphones are combined to obtain a reliable estimate while limiting the number of parameters. At the acoustic level, a parameter tying is applied to optimize the acoustic space description when multi-Gaussian models are used.

2. PHONETIC LEVEL

At the phonetic level, considering the contextual effect allows a more relevant phoneme based acoustic modeling [8].

2.1. Allophones and Triphones

A triphone is a contextual modeling of a phoneme which depends on its right and left contexts. Every triphone model has its own parameter set. An allophone holds all the contextual realization of a phoneme [9]. Each sound is represented by a single model with several entry and exit states where tying of the probability density function is applied. Parameter tying allows to reduce total number of parameters (of all the units). In this model, a path from an entry state to an exit state represents an allophone. This results in a better estimate of these parameters when limited amount of training data is available. Unlike triphone, the allophone uses an a priori parameter tying.

Because of the tying of the parameters in the allophone modeling, its estimate needs less training data than for triphone modeling. The difference between allophone and triphone is based on a trade-off between a correct model parameters estimate and a detailed modeling. To achieve the compromise, combination of allophone and triphone modeling is proposed.

2.2. Combining Allophones and Triphones

Combination between allophone and triphone, in a given phonetic context, consists in choosing the triphone whenever it is reliably estimated and the allophone otherwise. The reliability criterion is the number of triphone occurences in the training corpus.

More precisely, the combination between allophone and triphone uses a threshold on the number of occurences used to estimate phonetic models during the training. With a low threshold, the combined model uses too many triphones that may be badly estimated, which decreases the recognition performances. A high threshold results in a more reliable estimate of the phonetic units parameters. However, the obtained models are less detailed since a large amount of allophones is used. The threshold should be optimised in order to get a good compromise between an accurate modeling and a reliable estimate of the parameters. Experimental results are reported in section 4.2..

3. ACOUSTIC LEVEL

Increasing the number of components of multi-Gaussian distributions improves the recognition performances if sufficient training data are available. However, it also increases the computational cost. To overcome this problem, merging strategy is applied to Gaussian densities.

3.1. Gaussian Density Splitting

In the proposed approach, the considered multi-Gaussian distributions are caracterised by the number of used Gaussian densities. Consider two models λ_{n1} with n_1 Gaussian densities and λ_{n2} with n_2 Gaussian densities. If $n_2>n_1$, model λ_{n2} generally provides better recognition performances than λ_{n1}, especially in a case of no over-training.

The HMM emission distribution for an observation X[τ] related to a transition is given by:

$$B\ (X[\tau]) = \max_{1\leq k\leq NG} \{c_k.N(X[\tau];\ \mu_k,\ \Sigma_k)\}$$

where $N(.;\mu_k,\ \Sigma_k)$ is a Gaussian density with μ_k, the mean vector, Σ_k, the diagonal covariance matrix and c_k, the Gaussian component weight. NG is the number of Gaussian components of the multi-Gaussian distribution B.

To increase the number of Gaussian densities, a splitting is performed. Splitting consists in slightly perturbing the initial parameter set for those components which were estimated from a sufficient number of examples. This splitting is performed at the beginning of the training phase where parameters are reestimated. Thus, each Gaussian function with parameters, c_k, μ_k and Σ_k, is replaced by two Gaussian functions with new parameters: μ'_k and μ''_k, mean vectors, Σ'_k and Σ''_kcovariance matrices, c'_k and c''_k, component weights, such as:

$$\mu'_k = \mu_k + \varepsilon\ .\ (\Sigma_k)$$
$$\mu''_k = \mu_k - \varepsilon\ .\ (\Sigma_k)$$
$$\Sigma'_k = \Sigma''_k = \Sigma_k$$
$$c'_k = c''_k = c_k\ /\ 2$$

where ε is the mean perturbation factor and (Σ_k) the diagonal parameters vector of Σ_k. The number of Gaussian densities is approximately multiplied by 2 after every splitting. So, to reduce the number of parameters a merging technique is applied.

3.2. Gaussian Density Merging

Let assume that model λ_{n2} is obtained by splitting from model λ_{n1} and that merging of the densities of the model λ_{n2} is performed to obtain the same number of parameters as in the first one, λ_{n1}. $\lambda_{n2\rightarrow n1}$ denotes the parameter set of the resulting model.

We should have in principle:

$$P(\{X\}/\lambda_{n2\rightarrow n1}) \leq P(\{X\}/\lambda_{n2})$$

The error rate reduction obtained after a splitting between λ_{n2} and λ_{n1} will be compared in the section 4.3. with the error rate reduction after a splitting and a merging between $\lambda_{n2\rightarrow n1}$ and λ_{n1}.

The main idea is that after merging the Gaussian densities, each multi-Gaussian still has the same number of components, which are shared with other multi-Gaussian of the model. So, if the merging process does produce reliable merged Gaussian functions, the merged model is as accurate as the original one. Moreover, this merging may avoid an over-training phenomenon when the amount of training data is limited. So, the merged model may provide better description of the acoustical space than the original model. Here, a clustering strategy is applied to merge Gaussian densities.

The closeness of some Gaussian functions leads to merge these functions. Denote $N(.;\ \mu_1,\ \Sigma_1)$ and $N(.;\ \mu_2,\ \Sigma_2)$ two Gaussian functions to which n_1 and n_2 acoustical frames have been associated. The distance between these two functions is measured as the decrease in the likelihood of the corresponding training set observation after merging [7]. Denote d the acoustic space dimension, the distance D is given by:

$$D = -n_1\ .\sum_{i=1}^{d}\log(\sigma_{1i}) - n_2\ .\sum_{i=1}^{d}\log(\sigma_{2i}) + (n_1 + n_2)\ .\sum_{i=1}^{d}\log(\sigma_i)$$

where $(\Sigma_1) = (\sigma^2_{1i})_{(1\leq i\leq d)\}}$, $(\Sigma_2) = (\sigma^2_{2i})_{(1\leq i\leq d)}$. $(\Sigma) = (\sigma^2_i)_{(1\leq i\leq d)}$ is the diagonal parameters vector of the covariance matrix resulting of the merging.

If these two Gaussian functions are merged, the resulting function has a number of frames equal to the sum of the number of frames associated to the functions that are merged. Its parameters μ_i and σ^2_i after a weight normalisation, are estimated by:

$$n'_1 = n_1\ /\ (n_1 + n_2)$$
$$n'_2 = n_2\ /\ (n_1 + n_2)$$
$$\mu_i = n'_1\ .\ \mu_{1i} + n'_2\ .\ \mu_{2i}$$
$$\sigma^2_i = n'_1\ .\ \sigma^2_{1i} + n'_2\ .\ \sigma^2_{2i} + n'_1\ .\ n'_2\ .\ (\mu_{1i} - \mu_{2i})^2$$

4. EXPERIMENTS

Parameter tying techniques are evaluated in a flexible speech recognition system on several telephone databases.

4.1. Training and Test Databases

The training database is made of about 700 short sentences recorded by hundreds of speakers calling from different regions of France. This telephone database contains almost all the French diphones. For evaluation of recognition performances three speaker independent isolated words telephone databases are used. Table 1 presents the characteristics of these databases.

Database	**Number of Speakers**	**Number of Records**
Digits *(0 to 9)*	335	3622
Tregor *(36 words)*	380	12844
Numbers *(00 to 99)*	385	7288

Table 1: Characteristics of test databases.

Table 2 enables to evaluate the triphone covering of the test databases and shows that almost all the triphones used to model the digits are present in the training set.

Database	**Total amount of triphones units**	**Amount of triphones units for which the number of training occurences is less than**		
		20	**30**	**50**
Digits	*57*	1	2	9
Tregor	*724*	66	80	218
Numbers	*7942*	929	1419	3145

Table 2: Total number of triphones used in the modeling as a function of the number of occurences in the training set.

4.2. Phonetic Tying Tests

The Figure 1 describe the variation of error rate as a function of the ratio between the combined model Gaussian density number and the allophone based model Gaussian density number. Results are presented with different occurrence thresholds **S** (see section 2.2.). Occurrence thresholds represent the minimum number of triphone occurences in the training set. The Gaussian density number is a function of this threshold.

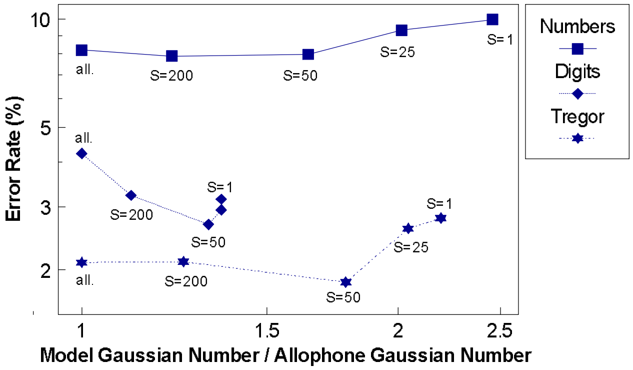

Figure 1: Combination of allophones and triphones in Numbers, Digits and Tregor models.

The first point is that for all databases an optimal occurrence threshold **S** exists for which the error rate is lower than for the allophone based modeling and the combined modeling with **S**=1. Combination of allophone and triphone modeling leads to a maximum error rate reduction of 37% for the Digits, 12% for the Tregor and 6% for the Numbers compared with the reference allophone model. The error rate improvement is linked to the triphone covering described in the Table 2.

The increasing of the amount of Gaussian densities compared with the allophone model in the case of a maximum error rate reduction is 32% for Digits, 78% for Tregor and 22% for Numbers. The trade-off between computational cost and recognition performances is clearly linked to the phonetic covering of the test databases.

4.3. Acoustic Tying Tests

The current experiments involve an initial allophone based modeling. The reference model is a multi-Gaussian model with 1 Gaussian component. Three successive splittings are applied on the reference model and after each splitting the corresponding model is trained. At each step, a merging is applied in order to obtain the same number of Gaussian densities as for the model before splitting. Figure 2, 3 and 4 show the error rate obtained in these two different cases, the splitting only case and the splitting and merging case.

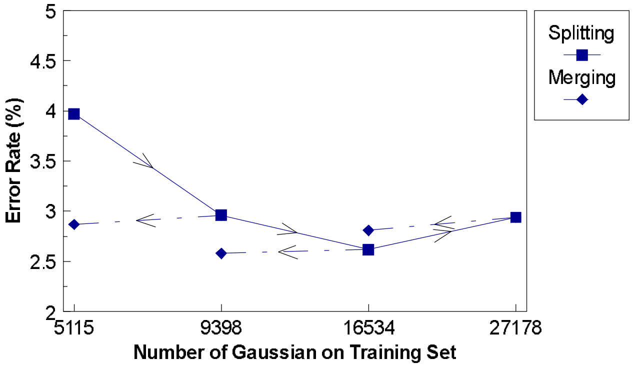

Figure 2: Splitting and merging in the Digits Model.

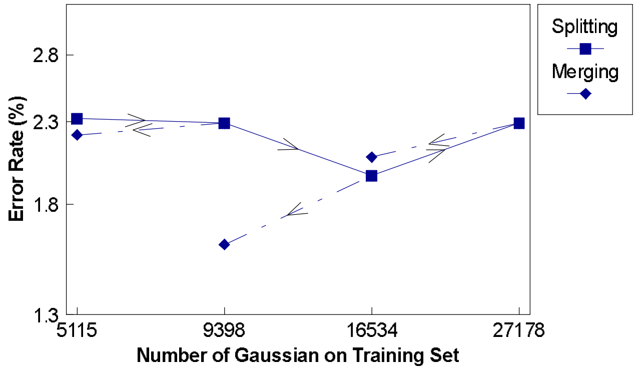

Figure 3: Splitting and merging in the Tregor Model.

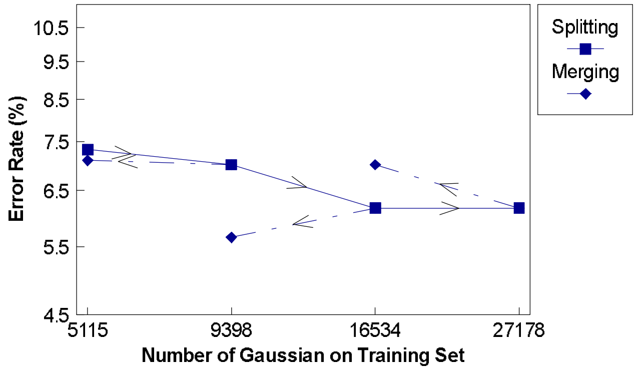

Figure 4: Splitting and merging in the Numbers Model.

These results with the 8 Gaussian densities per distribution model, i.e. 27178 Gaussian densities on the training set, illustrate the over-training problem.

With the Numbers model, the error rate is twice or three times the one obtained on the Digits and the Tregor models. The error rate obtained with the 8 Gaussian densities per distribution model is equivalent to the one obtained with the 4 Gaussian densities model. This means a poor modeling of some relative part of the acoustic space model. Then, when over-training occurs, the merging degrades more the acoustic space modeling with the Numbers corpus than with the Digits or the Tregor corpus.

In the optimal case, the model obtained after splitting and merging yields a 28% error rate reduction for the Digits, 30% for the Tregor and 21% for the Numbers, as compared to the reference model having the same total amount of Gaussian densities.

Results show a correlation between the error rate reduction provided by splitting only and the error rate reduction provided by splitting and merging on the same initial model. Moreover, the best error rate improvement after splitting only is associated with the optimal improvement after splitting and merging of the reference model. In this case, the merging provides a better acoustic modeling than the reference model. Experiments confirm thus the efficiency of this procedure of splitting and merging to optimize the acoustic space modeling.

5. CONCLUSION

The combination of allophone and triphone modeling leads to a better trade-off between computational cost and recognition performances as compared to a reference allophone modeling. This is particularly the case for flexible speech recognition when the triphones present in the training corpus accurately cover the tested corpus.

At the acoustic level, the proposed splitting and merging procedure selects the most appropriate acoustic model. The optimization of the acoustic model proposed in the procedure is a key parameter to solve the flexible system problem, i.e. an efficient covering of the acoustic space and a relatively low computational cost.

6. REFERENCES

1. S.J. Young "The General Use of Tying in Phoneme-based HMM Speech Recognisers", *Proc. of ICASSP*, Vol.3: 569-572, 1992.
2. K. Bartkova, D. Jouvet "Modelization of Allophones in a Speech Recognition System", *Proc. of ICPhS*, Vol. 4: 474-477, 1991.
3. X.D. Huang, M.A. Jack "Semi-Continuous Hidden Markov Models for Speech Signal", *Computer Speech and Language*, Vol. 3: 239-252, 1989.
4. X.D. Huang "Phoneme Classification Using Semicontinuous Hidden Markov Models", *IEEE Trans. on Signal Processing* 40: 1062-1067, 1992.
5. M.Y. Hwang, X. Huang "Shared-Distribution Hidden Markov Models for Speech Recognition", *IEEE Trans. Speech and Audio Processing* 1: 414-420, 1993.
6. J.R. Bellegarda, D. Nahamoo "Tied Mixtures Continuous Parameter Modeling for Speech Recognition", *IEEE Trans. ASSP* 38: 2033-2045, 1990.
7. D. Jouvet, L. Mauuary, J. Monné "Automatic Adjustments of the Structure of Markov Models for Speech Recognition Applications", *Proc. of EuroSpeech*, Vol.2: 923-927, 1991.
8. R. Schwartz, Y. Chow, O. Kimball, S. Roucos, M. Krasner and J. Makhoul "Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech", *Proc. of ICASSP*, Vol. 3: 1205-1208, 1985.
9. D. Jouvet, K. Bartkova and A. Stouff "Structure of Allophonic Models and Reliable Estimation of the Contextual Parameters", *Proc. of ICSLP*, Vol.1: 283-286, 1994.