DATA-DRIVEN PMC AND BAYESIAN LEARNING INTEGRATION FOR FAST MODEL ADAPTATION IN NOISY CONDITIONS

Stefano Crafa, Luciano Fissore◇ and Claudio Vair◇*

◇ CSELT - Centro Studi E Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - 10148 Torino, ITALY
* CSELT Consultant
e-mail: {vair, fissore}@cselt.it

ABSTRACT

In this paper, we present an integration of Data Driven Parallel Model Combination (DPMC) and Bayesian Learning into a fast and accurate framework which can be easily integrated in standard training and recognition systems.

The original DPMC technique has been enhanced to avoid any modification of the acoustic models, as required by the original method. The Bayesian Learning estimation has been used in order to specialize a general noisy speech model (the a priori model) to the target acoustic environment, where the DPMC-generated observations are used as adaptation data.

Thanks to these innovations, the proposed method can achieve better performance than the original DPMC, while consuming far less computational resources.

1. INTRODUCTION

A practical speech recognition system used over the telephone network has to deal with a different noisy environment from call to call. Differences come both in noise type (stationary, non-stationary, impulsive) and in SNR. In order to cope with these difficulties, a noise compensation method must be found, which is both accurate and fast. A solution which requires as little modification as possible of the standard training and test environments is also desirable.

Data-driven Parallel Model Combination (DPMC) [1] yields good recognition performance in noisy environments, but it has two major drawbacks. First, DPMC is not very fast, mostly because a great number of observations has to be generated, processed and combined in order to obtain sufficiently accurate models. Moreover, it requires the use of nonstandard information: the knowledge of the 0-th cepstral parameter is essential for this technique, but its value is not usually included in HMM models, since it is replaced by frame energy. DPMC needs also the delta and delta-delta parameters to be calculated as simple differences [2], instead of linear regressions as in state of the art recognizers.

The solution proposed in this paper provides a method which eliminates both problems, while leaving unchanged DPMC's good recognition performance. In particular, the use of Bayesian Learning model estimation makes it possible to obtain accurate speech models even if a reduced number of artificial observations is used, enhancing the speed of the whole process.

2. DATA-DRIVEN PMC

DPMC [1] faces the problem of recognition in a noisy environment in a simple but effective way. The corrupted speech model used for recognition is calculated for each particular noisy condition, starting from the clean speech model and the specific noise model. The combination process closely follows what happens in the physical world: first a number of observations is generated from the two models (speech and noise), in order to make explicit the data contained in them; after that, the observations are converted from cepstral domain to linear spectrum and then additively combined. At this point we have obtained corrupted speech observations, that can be used to train the corrupted speech model.

Since information is stored in the HMM models in the form of gaussian (or linear combination of gaussian) probability density functions, the generation of artificial observations can be achieved using an extension of Box-Muller's method [3]. The generation of artificial observations has several advantages when compared to direct model combination: the combination process is made easier, while requiring less computational power [2]. The number of observations which have to be generated depends on the desired accuracy of the resulting model, but it is far lower than the number of observations generally contained in a training database. That is possible because the HMMs (and the observations generated from them) include explicit information about the multimodality of the speech signal, thus removing the redundancy present in the training database. A qualitative example of this fact is shown in figure 1. In order to plot it, two HMM models have been trained with an increasing number of observations, artificial and real respectively. The distances (calculated as Kullback-Leibler number) between the original model and the ones just trained have been calculated and plotted in figure 1. It can be seen that using the same number of observations, the model trained on the artificial samples is more accurate than the other one. The trend reverses only when the number of samples used approaches the total number of observations used to compute the reference model: that is because the model computed on the real samples is exactly equal to the one it is compared to, so the distance rapidly goes to zero.

The generated observations are not equally distributed among the HMM's states. Each state is assigned a number

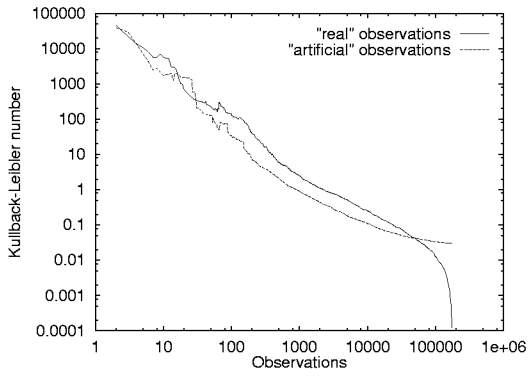

Figure 1: The relative "speed" with which real and artificial observations carry informations about speech.

of observations proportional to its number of gaussians. This proportionality parameter will be referred to in the following as PPG (points per gaussian).

2.1. The 0-th cepstral parameter

The HMM models generally adopted use the following observation vector: $\{E, C_1, C_2 \dots C_n\}$, where E is the frame energy and C_i is the i-th cepstral parameter. Unfortunately, the knowledge of the C_0 parameter is essential in order to convert the cepstral vector into the power spectrum domain. It is possible to compute the value of this parameter from that of the the others, instead of training a new model which includes C_0.

Given all cepstral parameters, from C_0 to C_n, the energy of the Mel bands on the log-spectrum can be computed as

$$\begin{bmatrix} E_0^l \\ E_1^l \\ \vdots \\ E_n^l \end{bmatrix} = \mathbf{D^{-1}} \cdot \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_n \end{bmatrix} \tag{1}$$

where $\mathbf{D}$ is the DCT matrix. The conversion of the energy to the linear spectrum is then determined by computing

$$E_i = 10^{E_i^l} \qquad 0 \le i \le n \tag{2}$$

By combining (1) and (2), we obtain a system of $n+1$ equations in $n+1$ variables, but in our case we do not know the value of C_0. To solve it we have to rely on another information: the sum of the energies of the Mel bands equals the total frame energy

$$\sum_{i=0}^{n} E_i = E \tag{3}$$

So, after easy calculations, we find that the desired value for C_0 is

$$C_0 = \log_{10} E - \log_{10} \sum_{i=0}^{n} 10^{\sum_{k=1}^{n} D_{ik} \cdot C_k} \tag{4}$$

The described process takes place for each artificial observation that has been generated from the HMM model. Once the value of C_0 is known, the compensation process can continue as in standard DPMC.

2.2. Delta and Delta-delta parameters

In the original DPMC method [2], dynamic parameters compensation is carried out only under the hypothesis that delta parameters are computed as a *difference* of static parameters:

$$\Delta \mathbf{O}^c(\tau) = \mathbf{O}^c(\tau + k) - \mathbf{O}^c(\tau - k) \tag{5}$$

on the contrary, in most recognition systems delta parameters are obtained by the following *regression* formula:

$$\Delta \mathbf{O}^c(\tau) = \alpha \cdot \frac{\sum_{k=-M}^{M} k \cdot \mathbf{O}^c(\tau + k)}{\sum_{k=-M}^{M} k^2} \tag{6}$$

Moreover, the classical DPMC requires the knowledge of additional data, such as a model of the probability distribution of $\mathbf{O}^c(\tau - k)$, in order to achieve good compensation of the dynamic (delta and delta-delta) parameters. These statistics should be computed explicitly for this noise compensation method, as they are not included into the HMM models normally used.

It is however possible to achieve good compensation without changing anything in the way dynamic parameters are computed by the front-end. Our procedure is illustrated in figure 2: each artificial frame of both clean speech and noise is composed by static and dynamic parameters. We expand the information contained in dynamic parameters at time τ into a sequence of static parameters at times $\tau + k$, where k ranges from $-M$ to M in equation (6). These newly obtained clean speech and noise static parameters are then combined together to obtain a sequence of corrupted speech static parameters, from which the corresponding dynamic parameter can be computed.

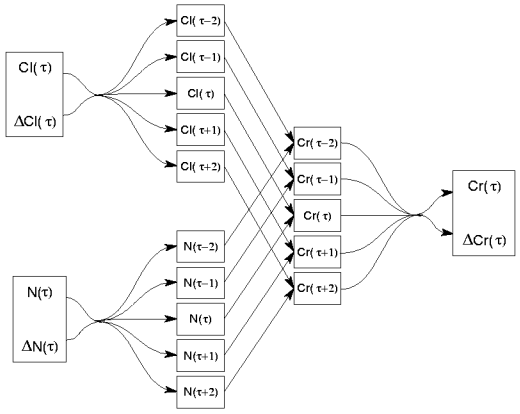

Figure 2: The delta parameter combination process. The following abbreviations have been used: 'Cl' for clean speech; 'N' for noise and 'Cr' for corrupted speech.

Equation (6) computes, according to the least square method, the slope of the line which interpolates the $2M+1$

samples at times $\tau+k$. Thus the parameter trend in time can be locally approximated as:

$$\mathbf{O}(\tau+k) = \mathbf{O}(\tau) + k\ \Delta\mathbf{O}(\tau) \tag{7}$$

This is the best observation sequence approximation, given the information we know.

If we also know the delta-delta parameter, we can reconstruct an even better approximation of the observations by using a partial expansion of the Taylor series:

$$\mathbf{O}(\tau+k) = \mathbf{O}(\tau) + k\ \Delta\mathbf{O}(\tau) + \frac{1}{2}k^2\Delta\Delta\mathbf{O}(\tau) \tag{8}$$

After clean speech and noise static parameters have been combined, the corrupted speech samples at times $\tau+k$ are known, so the delta parameter can be computed back using (6). For the delta-delta a similar equation is used.

3. BAYESIAN LEARNING

In standard DPMC method, the main way to set a compromise between computational load and resulting model accuracy is the choice of the number of artificial observations to be generated. Unfortunately, since a great amount of data is required to estimate accurate models, this tradeoff is always unbalanced towards an high number of observations, because the corrupted speech model is trained from scratch for each instance.

A great reduction of computational load can be obtained by using a different approach to model estimation. The unknown noisy environment of a particular phone call can be viewed as a specific instance of a more general noise model: this is the approach to the problem taken by the Bayesian Learning estimation method [4]. The general information about corrupted speech is taken from an HMM – the a priori model – computed only once over a training database which has been corrupted with different noises at various SNRs. The adaptation samples are obtained by the combination of artificial speech and noise observations created by DPMC. A priori and adaptation data are combined as follows, for the means:

$$\tilde{\mu}_k = (1-\lambda_k)\,\mu_k^{AP} + \lambda_k\,\mu_k^{AD} \tag{9}$$

and for the covariance matrices:

$$\begin{aligned}\tilde{\mathbf{\Sigma}}_k \quad = \quad & (1-\lambda_k)\,\mathbf{\Sigma}_k^{AP} + \lambda_k\,\mathbf{\Sigma}_k^{AD} + \\ & \lambda_k(1-\lambda_k)(\mu_k^{AP}-\mu_k^{AD})(\mu_k^{AP}-\mu_k^{AD})^T\end{aligned} \tag{10}$$

where tilde parameters are the ones estimated by Bayesian Learning, the ones with 'AP' superscript come from the a-priori models and those with 'AD' are computed on the adaptation data. The k subscript refers to each gaussian in the model. The weighting factor λ_k is defined as

$$\lambda_k = \frac{\zeta_k^{AD}}{\zeta_k^{AD}+\tau} \tag{11}$$

where ζ_k^{AD} is the average number of observations associated to the k-th gaussian, and τ is a weighting factor, which has to be empirically determined [5].

It is thus possible to reduce the number of observations generated and processed by the DPMC step, thanks to the Bayesian Learning's ability to yield accurate results even using a low amount of adaptation data. Furthermore, the use of a general-noise a priori model gives the opportunity to obtain better recognition performances than those obtained with a greater number of observations and a more traditional model estimation method.

4. RECOGNITION EXPERIMENTS

The above mentioned method has been tested by using a speaker-independent isolated-word telephone speech recognizer with a vocabulary of 475 city names. The vocabulary words are transcribed using 391 sub-word units [6], including context-independent and diphone transition units. The recognizer is based on Continuous Density HMMs with a mix of a maximum of 32 gaussians for each state. The models are differentiated for males and females speakers. The parameter vector is represented by $\{E, C_1, C_2, \ldots, C_n\}$ and by the corresponding delta and delta-delta parameters.

To simulate a noisy environment, real noise samples, collected in an open telephone box, have been added to the telephone speech signal with suitable normalization to get an SNR of 15dB. The tests are performed on a set of 14400 utterances collected from 1050 speakers.

The noise model has been computed on a one-minute registration made inside an open telephone box and consists in a single-state, single-gaussian HMM model. In order to compute the general noisy-speech model – the Bayesian a priori model – the training database has been corrupted with several noise types (excluded that of the open telephone box) at different SNR levels.

4.1. Experimental results

Table 1 presents the expected upper and lower bounds of the error rate for the proposed technique. The upper bound has been computed performing a recognition in the presence of noise using clean speech models; the degradation in performance is noticeable, if compared with the 2.92% error rate scored by clean models in a clean (31dB SNR) environment. In order to obtain the lower bound, a corrupted speech model has been computed on a version of the training database, corrupted with the same noise present in the test. The remaining entry of the table shows the performance obtained using for recognition the general noise a priori model that has been used for Bayesian Learning model estimation.

Table 1: Expected variability range of the Bayesian-DPMC algorithm.

	Model	ER %
Lower bound	Corrupted speech	6.30%
	Multi-noise	9.04%
Upper bound	Clean speech	13.60%

The τ parameter in equation (11) has to be chosen empirically [5], so a set of experiments has been done in order to determine its optimal value, as shown in table 2. The value of the PPG parameter in these experiments has been set to 10. Since DPMC makes use of pseudo-random data generation, its performance depends on the

Table 2: Error rates obtained with different τ weights given to the a-priori model. The PPG parameter has been set to 10.

τ	Mean	Std.Dev.	min	MAX
0	10.15%	0.47	9.30	10.58
10	9.78%	0.43	8.90	10.44
20	9.44%	0.45	8.70	10.05
50	9.08%	0.39	8.37	9.71
100	8.78%	0.34	8.47	9.24
200	8.52%	0.26	8.10	8.84
400	8.45%	0.25	8.03	8.84
500	8.42%	0.26	0.03	8.77
600	8.40%	0.13	8.23	8.57
700	8.40%	0.12	8.30	8.63
800	8.44%	0.12	8.23	8.63
1000	8.47%	0.22	8.30	8.84

seed the number generator is initialized with. For this reason the figures in table 2 are averaged over 10 different recognition tests. The recognition results show a clear improvement when compared with the performance of clean speech models. That is true even in the case of τ equal to zero, which corresponds to using the adaptation data only and excluding the a priori model. Due to the low number of observations generated, the performance for $\tau = 0$ is worse than that of the multi-noise model. As the weight of the a priori models grows, additional informations can be used, so both the error rate and its standard deviation decrease. The use of Bayesian Learning to combine the two sources of information yields better results than using only one of them, ending up in a reduction of the error rate ($\tau = 700$) of

$$\Delta ER\% = \frac{13,60 - 8,40}{13,60 - 6.30} \times 100 = 71,23\%$$

Table 3 reports the experiments performed with the PPG parameter set to 50. The higher number of observations generated yields better performance if compared to the case of PPG=10, but this improvement is somewhat small when it must be traded with an increase of five times of the running time.

Table 3: Error rates obtained with different weights given to the a-priori model. The PPG parameter has been set to 50.

τ	Mean	Std.Dev.	min	MAX
0	9.39%	0.28	8.90	9.77
50	8.94%	0.40	8.30	9.44
400	8.54%	0.23	8.10	8.84
600	8.34%	0.28	7.90	8.77
700	8.32%	0.26	7.90	8.63
800	8.32%	0.31	7.83	8.70
1000	8.31%	0.24	7.83	8.57
2000	8.30%	0.16	8.03	8.57
3000	8.32%	0.17	8.03	8.57
4000	8.32%	0.15	8.17	8.57

Comparison with standard training technique such as Segmental K-Means (SKM) is done in table 4. The advantage of the integration of DPMC with Bayesian Learning is evident: just by comparing lines 3 and 4 in table 4 it is obvious how such a combination is capable of increasing the recognition performance, while reducing of a whole order of magnitude the amount of data which requires being processed.

Table 4: Error rates obtained with different numbers of observations generated by DPMC.

Method	PPG	Mean	Std.Dev.	min	MAX
SKM	10	12.07%	0.42	11.22	11.52
SKM	50	10.07%	0.26	9.71	10.37
SKM	100	9.32%	0.24	8.04	9.71
Bayes.	10	8.40%	0.12	8.30	8.63
Bayes.	50	8.30%	0.16	8.03	8.57

5. CONCLUSIONS

We have addressed some problems related to speech recognition in a noisy environment. In particular we have developed a procedure which can be easily integrated in existing recognizers and brings considerable advantages in both speed and accuracy over standard DPMC technique. These improvements are accomplished through the use of Bayesian Learning model estimation and the exploitation of general information about speech in a noisy environment. The result of this is that it is possible to obtain better performances than by using standard Segmental k-Means training. At the same time computational load is reduced of an order of magnitude.

6. REFERENCES

[1] M.J.F.Gales, S.J.Young, "A Fast and Flexible Implementation of Parallel Model Combination", *Proceedings of ICASSP*, 1995.

[2] M. J. F. Gales, *Model-based Techniques for Noise Robust Speech Recognition*, Ph. D. Thesis, University of Cambridge, Sep 1995.

[3] G. E. M. Box, M. E. Muller. "A note on the generation of random normal deviates", *Annals Math. Stat.*, Vol 29, 610-611, 1958.

[4] J. L. Gauvain, C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", *IEEE Transactions on Speech and Audio Processing*, Vol 2, No 2, 291-298, Apr 1994.

[5] Q. Huo, C. H. Lee. "On-Line adaptative Learning of the Continuous Density Hidden Markov Model Based on Approximate Recursive Bayes Estimate" *IEEE Transactions on Speech and Audio Processing*, Vol 5, No 2, 161-172, March 1997.

[6] L. Fissore, F. Ravera, P. Laface. "Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition", *Eurospeech*, 1995