FRAME-SYNCHRONOUS STOCHASTIC MATCHING BASED ON THE KULLBACK-LEIBLER INFORMATION

Lionel Delphin-Poulat and Chafic Mokbel

France Télécom CNET/DIH/RCP
Technopole Anticipa
2, avenue Pierre Marzin
22307 Lannion Cedex, France
e-mail: Lionel.Delphin-Poulat@cnet.francetelecom.fr
e-mail: Chafic.Mokbel@cnet.francetelecom.fr

Jérôme Idier

Laboratoire des Signaux et Systèmes
Supélec
Plateau de Moulon
91192 Gif-sur-Yvette Cedex, France
e-mail: Jerome.Idier@lss.supelec.fr

ABSTRACT

An acoustic mismatch between a given utterance and a model degrades the performance of the speech recognition process. We choose to model speech by Hidden Markov Models (HMMs) in the cepstrum domain and the mismatch by a parametric function. In order to reduce the mismatch, one has to estimate the parameters of this function. In this paper, we present a frame synchronous estimation of these parameters. We show that the parameters can be computed recursively. Thanks to such methods, parameters variations can be tracked. We give general equations and study the particular case of an affine transform. Finally, we report recognition experiments carried out over both PSTN and cellular telephone network to show the efficiency of the method in a real context.

1. INTRODUCTION

In recent years, the problem of robustness of automatic speech recognition has been a great subject of interest. An efficient way to deal with robustness is to take into account the HMMs used to perform recognition and to model disturbances by a mismatch function. The observation associated with the HMM are the cepstrum coefficients since they efficiently extract short-term information from the speech signal. As stated in [6], this mismatch function can be viewed in the signal space, in feature space or in model space. Thus, we can define a different transform in each space but there exists a strong interaction between those different mismatch functions. For instance, within a Gaussian framework (*i.e.* Gaussian state dependent densities in the HMM), it can easily be seen that adapting data thanks to a linear transformation is equivalent to adapt both mean and variances in a constrained way. Up to now, studies have essentially focused on off-line methods relying on the well known expectation-maximization (EM) algorithm. Among them, linear regression is a popular technique: in [4] the model means are adapted thanks to linear regression and in [7] both means and variances are adapted thanks to the same technique. The main drawback of this method is the necessity to collect data on the new environment. However, on-line methods are becoming available such as in [2]. But all this methods are feasible or interesting if data are stacked in blocks to perform estimation.
In this work, we process data in a frame-synchronous way. This allow to track the variations of the mismatch parameters. A way to achieve this goal was formerly presented in [5]. It was proposed in [3] and in [9] to base the estimation on a different criterion (*i.e* maximizing the Kullback-Leibler Information) and on a stochastic algorithm to design a frame-synchronous algorithm. In [9], the algorithm was applied to identify a FIR (finite impulse response) system and in [3] to perform on-line estimation of the parameters of an HMM. Here we use this framework to estimate the parameters of the mismatch function. We show from a theoretical point of view that we can deal with various kinds of functions (especially non linear functions) and that the parameters of the functions are updated recursively after each frame. On the contrary to [5], there is no need to solve an equation at each step of the algorithm.
In section 2, we present the theoretical framework. Then, in section 3, we show how this general framework can be applied to the case of an affine transform. In section 4, we report convergence measures of the proposed algorithm and results of recognition experiments. Finally in section 5, we give some conclusions and prospects.

2. THEORETICAL FRAMEWORK

2.1. Global Framework

Let $Y_t = y_0, \ldots, y_\tau, \ldots, y_t$ be a sequence of noisy observations. The equalization function is defined by $x_t = f_\theta(y_t)$ where the sequence $X_t = x_0, \ldots, x_\tau, \ldots, x_t$ is distributed according to an HMM λ, which has N states, and θ is a vector of size p containing the equalization function's parameters. x_τ and y_τ are assumed to be scalar. Let $S_t = s_0, \ldots, s_\tau, \ldots, s_t$ denotes a partial state sequence. The state dependent distributions are assumed to be Gaussian with mean μ_i and variance σ_i^2 in state i. The extension of the proposed method to HMMs with Gaussian mixtures is straightforward. The Kullback-Leibler Information is defined by:

$$J(\theta) = E\{\log(\mathrm{p}(Y_t|\theta)|\theta^0)\} \tag{1}$$

where θ^0 denotes the exact parameter of the mismatch function. It is shown in [8] and in [3] that it is possible to generate a sequence $\{\theta_t\}$ that increases the Kullback-Leibler information; the sequence $\{\theta_t\}$ converges to the true parameter θ^0. These values are generated in the following way:

$$\theta_{t+1} = \arg\max_\theta Q_{t+1}(\Theta_t, \theta) \tag{2}$$

with $\Theta_t = (\theta_0, \ldots, \theta_t)$ and:

$$Q_{t+1}(\Theta_t, \theta) = E\{\log(\mathrm{p}(Y_{t+1}, S_{t+1}|\theta)|Y_{t+1}, \Theta_t\} \quad (3)$$

2.2. Stochastic Algorithm

We define the sequential variables:

$$\alpha_{\tau|\Theta_{\tau-1}}(i) = \mathrm{p}(Y_\tau, s_\tau = i|\Theta_{\tau-1}) \quad (4)$$

$$\beta_{\tau|t,\Theta_{\tau-1}}(i) = \mathrm{p}(Y_{\tau+1}^t|s_\tau = i, \Theta_{\tau-1}) \quad (5)$$

$$\gamma_{\tau|t,\Theta_{\tau-1}}(i) = \mathrm{P}(s_\tau = i|Y_t, \Theta_{\tau-1}) \quad (6)$$

with $Y_{\tau+1}^t = y_\tau, \ldots, y_t$ and $\tau \leq t$. The computation of those variable is detailed in [3]. Thanks to those variables we are going to compute $Q_{t+1}(\Theta_t, \theta)$ (In the following equations, we will omit additive terms that do not depend on θ). It can easily be shown that:

$$Q_{t+1}(\Theta_t, \theta) = \sum_{\tau=1}^{t+1} \mathcal{L}_{\tau|t+1}(\Theta_{\tau-1}, \theta) \quad (7)$$

with:

$$\mathcal{L}_{\tau|t+1}(\Theta_{\tau-1}, \theta) = -\frac{1}{2}\sum_{i=1}^{N} \gamma_{\tau|t+1,\Theta_{\tau-1}}(i)\frac{(f_\theta(y_t) - \mu_i)^2}{\sigma_i^2} + \log(|f'_\theta(y_t)|) \quad (8)$$

where $f'_\theta(y_t) = \frac{\partial f_\theta(y)}{\partial y}(y_t)$. θ is computed according to the following stochastic algorithm:

$$\theta_{t+1} = \theta_t + (I_{t+1}(\theta_t))^{-1} S(\theta_t, y_{t+1}) \quad (9)$$

with :

$$S(\theta_t, y_{t+1}) = \left.\frac{\partial \mathcal{L}_{t+1|t+1}(\Theta_t, \theta)}{\partial \theta}\right|_{\theta=\theta_t} \quad (10)$$

and with the Fisher Information Matrix :

$$I_{t+1}(\theta_t) = \left. -\frac{\partial^2 Q_{t+1}(\Theta_t, \theta)}{\partial \theta^2}\right|_{\theta=\theta_t} \quad (11)$$

Let us compute these two values:

$$S(\theta_t, y_{t+1}) = \left. -m_{t+1|t+1}(\Theta_t, \theta)\frac{\partial f_\theta(y_t)}{\partial \theta}\right|_{\theta=\theta_t} + \left.\frac{1}{f'_\theta(y_\tau)}\frac{\partial f'_\theta(y_\tau)}{\partial \theta}\right|_{\theta=\theta_t} \quad (12)$$

with $m_{\tau|t+1}(\Theta_{\tau-1}, \theta) = \sum_{i=1}^{N} \gamma_{\tau|t+1,\Theta_{\tau-1}}(i)\frac{(f_\theta(y_\tau) - \mu_i)}{\sigma_i^2}$. In the expression of $S(\theta_t, y_{t+1})$, it should be noted that $\frac{f_{\theta_t}(y_{t+1}) - \mu_i}{\sigma_i^2}$ represents the difference between the modified current frame the mean of Gaussian i multiplied by the corresponding precision. The frame is modified according to the estimated value of θ at time t. $m_{t+1|t+1}(\Theta_t, \theta)$ is the conditional expectation of the previous value with respect to the estimated MAP state probability density at time t. This term is a weight of the gradient vector $\left.\frac{\partial f_\theta(y_t)}{\partial \theta}\right|_{\theta=\theta_t}$. Therefore the first term in equation 12 tends to approach the filtered frame at time t towards the mean of the most probable state. The second term in 12 prevents the algorithm from compressing the whole space by setting $f_\theta(y_t)$ to the mean of the most probable state at time t. Moreover:

$$\begin{aligned} I_{t+1}(\theta_t) = \sum_{\tau=1}^{t+1} \Bigg\{ & m_{\tau|t+1}(\Theta_{\tau-1}, \theta) \left.\frac{\partial^2 f_\theta(y_\tau)}{\partial \theta^2}\right|_{\theta=\theta_t} \\ & + n_{\tau|t+1}(\Theta_{\tau-1}, \theta_t) \left.\frac{\partial f_\theta(y_\tau)}{\partial \theta}\right|_{\theta=\theta_t} \left.\frac{\partial f_\theta(y_\tau)}{\partial \theta}\right|_{\theta=\theta_t}^T \\ & - \frac{1}{(f'_{\theta_t}(y_\tau))^2} \left.\frac{\partial f'_\theta(y_\tau)}{\partial \theta}\right|_{\theta=\theta_t} \left.\frac{\partial f'_\theta(y_\tau)}{\partial \theta}\right|_{\theta=\theta_t}^T \\ & + \frac{1}{f'_{\theta_t}(y_\tau)} \left.\frac{\partial^2 f'_\theta(y_\tau)}{\partial \theta^2}\right|_{\theta=\theta_t} \Bigg\} \end{aligned} \quad (13)$$

with $n_{\tau|t+1}(\Theta_{\tau-1}, \theta) = \sum_{i=1}^{N} \frac{\gamma_{\tau|t+1,\Theta_{\tau-1}}(i)}{\sigma_i^2}$ and T denoting the transpose operator.

2.3. Approximations

The computation of $I_{t+1}(\theta_t)$ is very difficult and cannot be achieved in practice without making some approximations. First, we cannot compute $\gamma_{\tau|t,\Theta_{\tau-1}}(i)$ for every $t \geq \tau$. In [3], $\gamma_{\tau|t,\Theta_{\tau-1}}(i)$ is replaced by the fixed-lag variable $\gamma_{\tau|\tau+\Delta,\Theta_{\tau-1}}(i)$ where Δ represents the future frames taken into account to estimate the *a posteriori* distributions of the HMMs' states. In our particular case we set $\Delta = 0$ which is the filtered Markov state estimate. One should also notice that after doing this, $I_{t+1}(\theta_t)$ is still a sum over τ of terms of the kind $g(\theta_t, y_\tau)$. Therefore for each new value of θ_t, one should compute the entire sum from the beginning. One solution might be to replace $g(\theta_t, y_\tau)$ by $g(\theta_\tau, y_\tau)$. This formula can also be simplified if $g(\theta_t, y_\tau)$ is separable *i. e.* $g(\theta_t, y_\tau) = g_\theta(\theta_t)g_y(y_\tau)$, $g_\theta(\theta_t)$ can be factorized in the sum. We will see in section 3 that in the case with an affine transform, all functions are separable. Finally, as stated in [3], $(I_{t+1}(\theta_t))^{-1}$ in equation 9 can be replaced by $K_t I_p$. I_p denotes the identity matrix of order p and K_t is any sequence of positive numbers that satisfies:

$$\lim K_t = 0 \quad \sum_{t=1}^{\infty} K_t = \infty \quad \sum_{t=1}^{\infty} K_t^2 < M < \infty \quad (14)$$

Moreover, we will not compute the forward variable: we will make the Viterbi assumption and replace the summation by taking a maximum. By doing this, we also avoid numerical problems that appear in the computation of the forward variables. To compensate for the absence of knowledge of the future, we will maintain several estimations of θ along each path as it was proposed in [5].

3. APPLICATION TO AN AFFINE TRANSFORM

After developing the general framework, we show how this method can be applied to the case of an affine transform which was reported to be of interest in [4] and in [7]. Here the mismatch function is $f_\theta(y_t) = ay_t + b$, with $\theta = [b, a]^T$. Hence, we have $f'_\theta(y_t) = a$. We have to compute the following derivatives: $\frac{\partial f_\theta(y_t)}{\partial \theta} = [1, y_t]^T$, $\frac{\partial f'_\theta(y_t)}{\partial \theta} = [0, 1]^T$, $\frac{\partial^2 f_\theta(y_t)}{\partial \theta^2} = 0_{2,2}$,

$\frac{\partial^2 f'_\theta(y_t)}{\partial \theta^2} = 0_{2,2}$. $0_{2,2}$ is the null matrix of order 2. We then deduce:

$$S(\theta_t, y_{t+1}) = -\begin{bmatrix} m_{t+1|t+1}(\Theta_t, \theta) \\ m_{t+1|t+1}(\Theta_t, \theta) y_{t+1} - \frac{1}{a_t} \end{bmatrix} \quad (15)$$

$m_{t+1|t+1}(\Theta_t, \theta) = \sum_{i=1}^{N} \gamma_{t+1|t+1,\Theta_t}(i) \frac{a_t y_{t+1} + b_t - \mu_i}{\sigma_i^2}$ in this particular case. Let us denote $C_\tau = \sum_{i=1}^{N} \gamma_{\tau|t+1,\Theta_{\tau-1}}(i) \frac{1}{\sigma_i^2}$ (we have $C_\tau > 0$). We can rewrite the Fisher Information Matrix:

$$I_{t+1}(\theta_t) = \begin{bmatrix} \sum_{\tau=1}^{t+1} C_\tau & \sum_{\tau=1}^{t+1} C_\tau y_\tau \\ \sum_{\tau=1}^{t+1} C_\tau y_\tau & \sum_{\tau=1}^{t+1} C_\tau y_\tau^2 + \frac{t+1}{a_t^2} \end{bmatrix} \quad (16)$$

The matrix determinant is:

$$\delta_{t+1} = \sum_{1 \le \tau < \tau' \le t+1} C_\tau C_{\tau'} (y_\tau - y_{\tau'})^2 + \sum_{\tau=1}^{t+1} C_\tau \frac{t+1}{a_t^2} \quad (17)$$

This determinant is non-negative. Thus the matrix is always invertible and since it is of order 2, it can be expressed as follow:

$$I_{t+1}(\theta_t)^{-1} = \frac{1}{\delta_{t+1}} \begin{bmatrix} \sum_{\tau=1}^{k+1} C_\tau y_\tau^2 + \frac{t+1}{a_t^2} & -\sum_{\tau=1}^{t+1} C_\tau y_\tau \\ -\sum_{\tau=1}^{t+1} C_\tau y_\tau & \sum_{\tau=1}^{t+1} C_\tau \end{bmatrix} \quad (18)$$

Recursion on $\theta_t = [b_t, a_t]^T$ can be easily written. If we take a simpler equalization function, $x_t = f_\theta(y_t) = y_t + b$ with $\theta = b$ and if we make the Viterbi assumption as proposed in 2.3 it can easily be shown that along the path S_{t+1}:

$$b_{t+1} = b_t - \frac{1}{\sum_{\tau=1}^{t+1} \frac{1}{\sigma_{s_\tau}^2}} \frac{y_{t+1} + b_t - \mu_{s_{t+1}}}{\sigma_{s_{t+1}}^2} \quad (19)$$

This expression is equivalent to:

$$b_t = -\frac{\sum_{\tau=1}^{t} \frac{y_t - \mu_{s_\tau}}{\sigma_{s_\tau}^2}}{\sum_{\tau=1}^{t} \sigma_{s_\tau}^2} \quad (20)$$

It is to be noticed that we have the same expression for b as the one given in [5].

4. EXPERIMENTAL RESULTS

4.1. Convergence Measures

We have verified on an example the convergence properties of the proposed method. The HMM chosen is in fact a mixture of two Gaussians with equal weight $\frac{1}{2}$. The first one has a zero mean and a standard deviation of 2. The second one has a mean of 3 and a standard deviation of 1. The true values of the affine transform are $a = 3$ and $b = -1$. The results obtained thanks to the proposed algorithm (referred to KL Estimation on the figure) are compared to the results obtained by EM computation of the parameters (referred to EM Estimation on the figure) thanks the algorithm proposed in [6]. They are also compared to the results (referred as MUSE on the figure) obtained by the technique proposed in [5]. For the batch EM algorithm, we perform 5 iterations of the EM algorithm every 10 observations (we take into account all observations from the beginning). In the proposed method and for the exact Maximum Likelihood estimation along each path (MUSE), we update the parameters for the 100 most likely paths and we plot the parameters corresponding to the most likely path at time t. All results appear on figure 1. For the three methods, we make the Viterbi approximation, which leads to biased estimated values. We can see that the two methods converge to the same point. On the figure, the distinction between the curves corresponding to the two frame-synchronous method cannot be made except at the beginning of the estimation. On the studied sample, the KL Estimation is smoother than the MUSE Estimation.

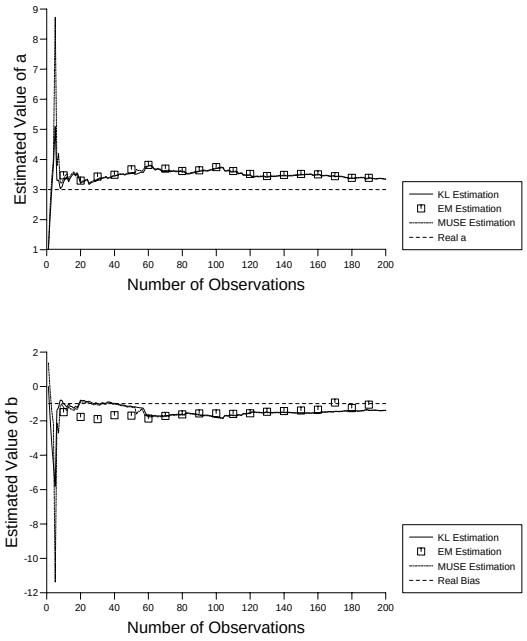

Figure 1: Convergence of Parameters of the Function.

4.2. Models and Database

The technique was tested on a digit database. This database was recorded on both PSTN and GSM network (European cellular telephone network). This database contains hundreds of call made by different speakers from different regions of France. The whole database contains thousands of utterances. For cellular telephone network recordings, we distinguish three conditions:

- GSM1: indoors and stopped car GSM recordings.
- GSM2: running car GSM recordings.
- GSM3: outdoors GSM recordings.

The model used are 30-state HMMs with Gaussian distributions. Feature vectors are composed of the first 8 cepstral coefficients, energy and their first and second order derivatives, thus the size of the feature vector is 27. The covariance matrices are diagonal, therefore the previous framework can be applied on each di-

mension of the feature vector separately. The system works in a speaker-independent mode.

4.3. Speech Recognition Results

In the results presented below, the model training is performed on half of the data recorded over PSTN and recognition experiments are performed on the other half of PSTN data and the GSM data. On figure 2, we plotted the recognition error rate versus the different testing conditions. We compared the results obtained thanks to the proposed method (referred as KL Linear Regression on the figure) to the baseline results (*i. e.* without adapting data). We also compared these results to those obtained thanks to a previous frame-synchronous linear adaptation presented in [1] (referred as MUSE linear regression on the figure). This frame-synchronous linear adaptation was done thanks to the method proposed in [5]. We can see that both frame-synchronous affine transforms lead to large recognition improvements. Those improvements are almost the same for these methods. Besides, in current implementations, the computational cost for both methods is the same.

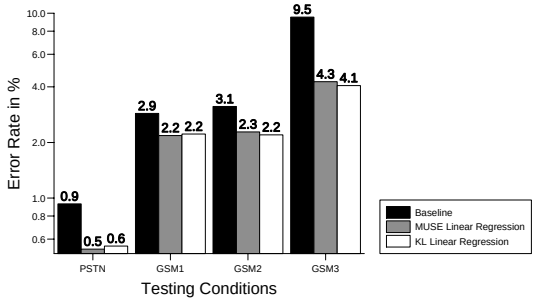

Figure 2: Error rate on the digit vocabulary with a PSTN-trained model.

5. CONCLUSION

In this article we presented a new way to perform frame-synchronous adaptation of data. We gave a general theoretical framework. We developed these equations in the case of a linear transform. Future works may include the study of non-linear functions, the ability of the proposed method to track perturbation. This work can also be easily extended to the case of parameters that depend on the state, this allow to have different linear transform and thus approximating a non-linear function by a piece-wise linear function. In this case, clustering should be performed to reliably estimate those parameters. We show the efficiency of the proposed method to deal with acoustic mismatch between GSM and PSTN recordings in automatic speech recognition. Moreover appropriately chosen coefficient K_t defined in equation 14 may lead to similar results with a lower computational cost.

6. REFERENCES

[1] Delphin-Poulat, L., Mokbel, C., "Frame-Synchronous Adaptation of Cepstrum by Linear Regression", to appear in Proc. ASRU'97, Santa-Barbara, USA, 1997.

[2] Digalakis, V., "On-Line Adaptation of Hidden Markov Models Using Incremental Estimation Algorithms", EUROSPEECH'97, pp. 1859-1862, Rhodes, Greece, 1997.

[3] Krishnamurthy, V., Moore, J.B., "On-Line Estimation of Hidden Markov Model Parameters Based on the Kullback-Leibler Information Measure", IEEE Trans. on Signal Processing, pp. 2557-2573, vol. 41, n. 8, August 1993.

[4] Legetter, C. J., Woodland, P.C., "Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression", Proc. ICSLP'94, pp. 451-454, Yokohama, Japan, 1994.

[5] Mokbel, C., "MUSE : MUlti-Path Stochastic Equalization A theoretical framework to combine equalization and stochastic modeling", Proc. ESCA workshop on Robust Speech Recognition, pp. 211-214, Pont-à-Mousson, France, 1997.

[6] Sankar, A., Lee, C.-H., "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", IEEE Trans. on Speech and Audio Processing, vol. 4, n. 3, pp. 190-202, May 1996.

[7] Soulas, T., Mokbel, C., Jouvet, D., Monné, J., "Adapting PSN Recognition Models to the GSM Environment by Using Spectral Transformation", Proc. ICASSP'97, pp. 1003-1006, Munich, Germany, 1997.

[8] Titterington, D.M.,"Recursive Parameter Estimation using Incomplete Data", J.R. Statist. Soc. B, vol. 46, n. 2, pp.257-267, 1984.

[9] Weinstein, E., Feder, M., Oppenheim, A.V.,"Sequential Algorithms for Parameter Estimation Based on the Kullback-Leibler Information Measure", IEEE Trans. on Speech and Audio Processing, vol. 38, n. 9, pp. 1652-1654, September 1990.