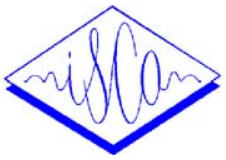

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

PATH-DEPENDENT KALMAN ESTIMATION OF A CEPSTRAL BIAS

Lionel Delphin-Poulat

France Télécom CNET/DIH/DIPS
Technopole Anticipa
2, avenue Pierre Marzin
22307 Lannion Cedex, France
e-mail: lionel.delphinpoulat@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

10.21437/Eurospeech.1999-333x

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 an additive bias. To track the variations of this bias, we explicitly model the way in which the bias can vary by a state equation. We derive a frame-synchronous estimator of this bias based on Kalman recursions. We use this estimator to compensate for the mismatch in the recognition process. Finally, we report recognition experiments carried out over both public switched telephone network (PSTN) and cellular telephone network to show the efficiency of the method in a real context.

1. INTRODUCTION

In quiet conditions, nowadays speech recognizers can achieve acceptable performances. But the recognition rates rapidly decrease, even for small vocabularies, when the recognition has to be performed over very disturbed channels. To compensate for those disturbances, one has to choose a model for the speech signal and for the disturbances. To perform recognition, the speech signal is often modelled in the cepstral domain, since cepstrum coefficients efficiently extract information from speech. One of the important sources of disturbance is the distortion due to the transmission channel. This distortion can be modelled by a convolution in time domain, or equivalently by a bias in cepstral domain. Several techniques have been proposed to compensate for this bias: we can subtract the cepstral mean, apply a high-pass filter [2] or perform a blind equalization [3, 4]. However, those techniques rely on a rough speech signal model.

That is why it was proposed [5, 7, 8] to view the HMM used to perform recognition as a clean speech signal model. In [7], the HMM is reduced to a mixture of Gaussians. In [8], the reference model is an HMM and the disturbance is modelled by a general parametric functions and the function parameters are estimated in an off-line fashion thanks to the Expectation Maximization algorithm. On-line estimates can be obtained thanks to the multipath stochastic equalization framework (MUSE) [4]: an equalization function is associated to each state sequence (also called path) in the HMM. The parameters (for instance bias) are evaluated for each path according to a maximum likelihood criterion and then a path-dependent compensation is applied to the noisy features.

In all these models, the disturbance parameters are assumed to be constant over time. In practice, that might not be the case: the disturbance parameters may vary, even within a single utterance. To take this effect into account, it was suggested to introduce a forgetting factor [1] in the variables used to compute the bias in the MUSE technique. Another way to cope with the parameters variability is to assign a prior density to the parameters [9]. Given that density function it is possible to obtain the likelihood of the observed noisy data. This method is referred to as the predictive approach.

In this work, we propose to explicitly model the way in which the bias can vary with time by a state equation: this introduces an a priori knowledge on the bias. The reference model for speech is the HMM with cepstral observations used to perform recognition. The presented technique is also a predictive technique, since we are able to compute the actual distribution of the observed noisy data.

In the next section, we establish the theoretical framework and a link between the proposed estimator and the blind equalization technique. In section 3, we present experimental results. First, we validate our approach on artificially generated data, we show that the method can track the variation of a piecewise constant or piecewise linear bias. Then, we carry out recognition experiments over PSTN and cellular telephone network. Finally, in section 4, we draw conclusions and give some prospects.

2. THEORETICAL FRAMEWORK

2.1. Model

Let $\mathbf{Y}_t = \mathbf{y}_1, \ldots, \mathbf{y}_\tau, \ldots, \mathbf{y}_t$ be a sequence of noisy observations and $\mathbf{X}_t = \mathbf{x}_1, \ldots, \mathbf{x}_\tau, \ldots, \mathbf{x}_t$. We assume that:

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{b}_t \tag{1}$$

$\mathbf{X}_t$ is distributed according to an HMM with Gaussian-state dependent densities (with mean $\boldsymbol{\mu}_i$ and covariance

matrix $\mathbf{\Sigma}_i$ for state i). We denote $S_t = s_1, \ldots, s_t$ any possible state sequence in the HMM. For $\mathbf{b}_t$, we assume that we have the following state equation:

$$\mathbf{b}_t = \mathbf{F}\mathbf{b}_{t-1} + \boldsymbol{\epsilon}_t \tag{2}$$

$\boldsymbol{\epsilon}_t$ is a sequence of i.i.d. Gaussian random vectors with a zero-mean and a covariance matrix $\mathbf{\Gamma}_{\boldsymbol{\epsilon}}$. The sequence $\{\boldsymbol{\epsilon}_t\}$ is indepent of the sequence $\{\mathbf{X}_t\}$. Let λ denote all the model parameters (*i.e.* the parameters related to the HMM and the parameters related to the state equation). To perform recognition, we have to search the optimal state sequence:

$$\hat{S}_T = \underset{S_T}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{Y}_T, S_T|\lambda) \tag{3}$$

We clearly have:

$$\mathrm{p}(\mathbf{Y}_T, S_T|\lambda) = \mathrm{p}(\mathbf{Y}_T|S_T, \lambda)\,\Pr(S_T|\lambda)$$

First, we are going to demonstrate that it is possible to compute $\mathrm{p}(\mathbf{Y}_T|S_T, \lambda)$ in a recursive fashion, for any state sequence. Then we will briefly give a solution to find the best path in the HMM. Finally, we will establish the link between the proposed method and the channel blind equalization method.

2.2. Likelihood Computation

To perform the recognition, we have to compute:

$$\mathrm{p}(\mathbf{Y}_T|S_T, \lambda) = \prod_{t=1}^{T} \mathrm{p}(\mathbf{y}_t|\mathbf{Y}_{t-1}, S_T)$$

We can notice that, in Kalman Filtering terminology, equation 1 is the observation equation and equation 2 is the state equation. We have for $t \leq T$:

$$\mathrm{p}(\mathbf{b}_t|\mathbf{Y}_{t-1}, S_T) = \mathrm{p}(\mathbf{b}_t|\mathbf{Y}_{t-1}, S_{t-1})$$

If we apply Kalman filter theory (see e.g. [6]), we obtain that given $\mathbf{Y}_{t-1}$ and S_{t-1}, $\mathbf{b}_t$ follows a Gaussian law with mean $\hat{\mathbf{b}}_{t|t-1}(S_{t-1})$ and covariance matrix $\mathbf{\Gamma}_{t|t-1}(S_{t-1})$. These values can be computed recursively as follows; first, we perform the prediction step:

$$\hat{\mathbf{b}}_{t|t-1}(S_{t-1}) = \mathbf{F}\hat{\mathbf{b}}_{t-1|t-1}(S_{t-1}) \tag{4}$$

$$\mathbf{\Gamma}_{t|t-1}(S_{t-1}) = \mathbf{F}^{\mathrm{T}}\mathbf{\Gamma}_{t-1|t-1}(S_{t-1})\mathbf{F} + \Gamma_{\boldsymbol{\epsilon}} \tag{5}$$

Then, we perform the estimation step:

$$\mathbf{K}(S_t) = \mathbf{\Gamma}_{t|t-1}(S_{t-1})(\mathbf{\Gamma}_{t|t-1}(S_{t-1}) + \mathbf{\Sigma}_{s_t})^{-1} \tag{6}$$

$$\hat{\mathbf{b}}_{t|t}(S_t) = \hat{\mathbf{b}}_{t|t-1}(S_{t-1}) + \mathbf{K}(S_t)(\mathbf{y}_t - \boldsymbol{\mu}_{s_t} - \hat{\mathbf{b}}_{t|t-1}(S_{t-1})) \tag{7}$$

$$\mathbf{\Gamma}_{t|t}(S_t) = \mathbf{\Gamma}_{t|t-1}(S_{t-1}) - \mathbf{K}(S_t)\mathbf{\Gamma}_{t|t-1}(S_{t-1}) \tag{8}$$

$\mathbf{K}(S_t)$ is the Kalman gain factor. Furthermore :

$$\mathrm{p}(\mathbf{x}_t|\mathbf{Y}_t, S_t) = \mathrm{p}(\mathbf{x}_t|s_t)$$

Thus knowing $\mathbf{Y}_{t-1}$ and S_t, $\mathbf{y}_t$ follows a Gaussian law with mean $\boldsymbol{\mu}_{s_t} + \hat{\mathbf{b}}_{t|t-1}(S_{t-1})$ and covariance matrix $\mathbf{\Sigma}_{s_t} + \Gamma_{t|t-1}(S_{t-1})$. The path likelihood $\mathrm{p}(\mathbf{Y}_T|S_T, \lambda)$ can be computed recursively.

We can make approximations to compute the likelihood $\mathrm{p}(\mathbf{Y}_t|S_t, \lambda)$. We can suppose that all covariance matrices are diagonal, which will be the case in recognition experiments. Let us denote $\gamma^2_{t|t-1}(S_t)$ any diagonal element of $\Gamma_{t|t-1}(S_t)$ and σ_i^2 the correspong diagonal element of $\mathbf{\Sigma}_i$. If $\gamma^2_{t|t-1}(S_t) \ll \sigma_i^2$, then knowing $\mathbf{Y}_{t-1}$ and S_t, $\mathbf{y}_t$ follows a Gaussian law with mean $\boldsymbol{\mu}_{s_t} + \hat{\mathbf{b}}_{t|t-1}(S_{t-1})$ and covariance matrix $\mathbf{\Sigma}_{s_t}$. This approximation means that the conditionnal distribution of $\mathbf{b}_t$ given $\mathbf{Y}_{t-1}$ and S_{t-1} is very sharp compared to the distribution of $\mathbf{x}_t$ given s_t.

2.3. Optimal Path Search

The optimal solution would consist in computing:

$$\mathrm{p}(\mathbf{Y}_T|S_T, \lambda)\,\Pr(S_T)$$

for any possible state sequence in the HMM. This is obviously not feasible. Therefore, as it was suggested in [5], we perform conventional Viterbi pruning: this is a sub-optimal solution, since the Kalman filtering technique is nested with pruning. Alternative solutions, such as keeping the list of the K best paths have also been proposed.

2.4. Link with Blind Equalization

We can use the proposed method in a filtering scheme. In this case, speech is represented by a general model (like a phoneme loop). We perform the previous algorithm; at each time instant t, we suppose that $\hat{S}_t$ is determined as in equation 3:

$$\hat{S}_t = \underset{S_t}{\operatorname{argmax}}\, \mathrm{p}(\mathbf{Y}_t, S_t|\lambda)$$

and we obtain an estimate for $\mathbf{x}_t$:

$$\hat{\mathbf{x}}_t = \mathbf{y}_t - \hat{\mathbf{b}}_{t|t}(\hat{S}_t) \tag{9}$$

Let us assume that we have a rough model for speech: cepstrum coefficients are i.i.d Gaussian random vectors with mean $\boldsymbol{\mu}_1$ and covariance matrix $\mathbf{\Sigma}_1$. Moreover, we set $\mathbf{F}$ to the identity matrix. Then equations 4 and 7 can be combined:

$$\hat{\mathbf{b}}_{t|t} = \hat{\mathbf{b}}_{t-1|t-1} + \mathbf{K}_t(\mathbf{y}_\tau - \boldsymbol{\mu}_1 - \hat{\mathbf{b}}_{t-1|t-1}) \tag{10}$$

This equation is similar to the equation in the blind equalization framework [3]. In the present case, $\mathbf{K}_t$ is optimally computed for each frame, while in [3] it is an a priori chosen coefficient.

3. EXPERIMENTAL RESULTS

3.1. Convergence Measures

We 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. We generated 1000 scalar observations. Two kinds of time varying bias were experimented. First, we disturbed the generated data by adding a piecewise constant bias:

$$\mathbf{b} = \begin{cases} 2 & \text{if } 0 \leq t < 333 \\ 1 & \text{if } 333 \leq t < 667 \\ 3 & \text{if } 667 \leq t < 1000 \end{cases}$$

Secondly, we added a piecewise linear bias:

$$\mathbf{b} = \begin{cases} 1 + \frac{t}{100} & \text{if } 0 \leq t < 500 \\ 6 - \frac{t-500}{100} & \text{if } 500 \leq t < 1000 \end{cases}$$

We compared the values obtained by Kalman Filtering and the values obtained by MUSE. For the MUSE technique, we introduced a forgetting factor λ_{ff} [1] and along each path S_t we compute:

$$\begin{aligned} \mathbf{X}_1(S_t) &= \mathbf{\Sigma}_{s_\tau}^{-1}(\mathbf{y}_\tau - \boldsymbol{\mu}_{s_\tau})^{\mathrm{T}} + \lambda_{ff}\mathbf{X}_1(S_{t-1}) \\ \mathbf{X}_2(S_t) &= \mathbf{\Sigma}_{s_\tau}^{-1} + \lambda_{ff}\mathbf{X}_2(S_{t-1}) \\ \hat{\mathbf{b}}_t(S_t) &= (\mathbf{X}_2(S_t))^{-1}(\mathbf{X}_1(S_t)) \end{aligned}$$

λ_{ff} was set to 0.975. In the Kalman filtering scheme, the matrices $\mathbf{F}$ and $\mathbf{\Gamma}_\epsilon$ are reduced to scalar f and γ_ϵ^2, we set $f = 1$ and $\gamma_\epsilon^2 = 0.001$ In the presented experiments, we kept the 100 most likely paths. We plotted the optimal bias for the optimal path at time t versus time t. The results are reported on figures 1 and 2 We can see that both methods

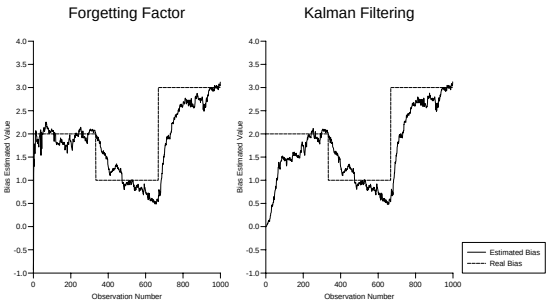

Figure 1: Tracking of a Piecewise Constant Bias

behave very similarly for a large number of frames. In the case of Kalman filtering, the convergence is slower than in the case of the MUSE estimation. This feature might be useful since it avoids strong overshoot at the beginning.

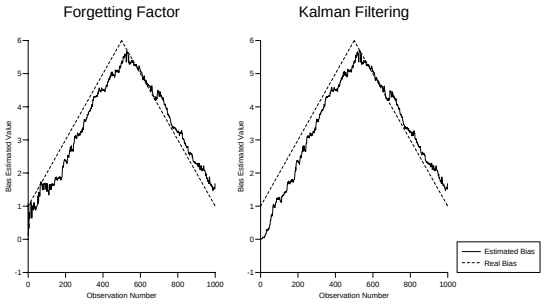

Figure 2: Tracking of a Piecewise Constant Bias

3.2. Models and Databases

The technique was evaluated on a digit database and on a 50-word-vocabulary database. Those databases were recorded on both PSTN and GSM network (European cellular telephone network). Both database contain hundreds of calls made by different speakers from different regions of France. They both contain 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 dimension of the feature vector separately. The system works in a speaker-independent mode.

3.3. Speech Recognition Results

In the results presented below, for both databases, 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. We plotted the recognition error rate versus the different testing conditions on figure 3 for the digit database and 4 for the 50-word vocabulary database. We denote by GSM the results obtained on the whole GSM recorded data.

We compared the results obtained thanks to the proposed method (referred to as KALMAN on the figures) to the baseline results (*i. e.* without adapting data). We also compared these results to those obtained thanks to MUSE with forgetting factor [1] (referred as MUSE). In the case of MUSE, the forgetting factor is set to 0.98. For the Kalman filtering technique, $\mathbf{F}$ is a diagonal matrix with all its diagonal elements set to 0.999 and $\mathbf{\Gamma}_\epsilon$ is also a diagonal matrix. The i^{th} diagonal element is approximately

set to $10^{-5}\sigma_g^2(i)$ ($\sigma_g(i)$ is the global standard deviation of the i^{th} feature). Those values were not tuned precisely. For both methods (MUSE and KALMAN), equalization was performed only on static coefficient, since we assume that the bias vary slowly enough not to disturb dynamic parameters. Furthermore, we made the approximation detailed in section 2.2 to compute the likelihood of the current frame given the path and the previous observations.

We can see that both methods lead to significant recog-

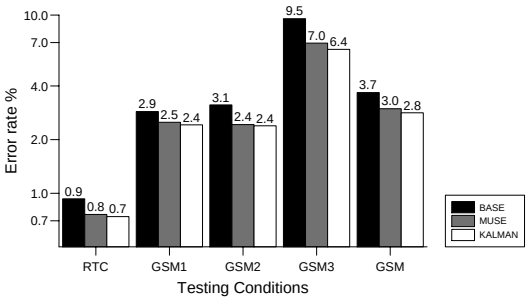

Figure 3: Error Rates on the Digit-Database

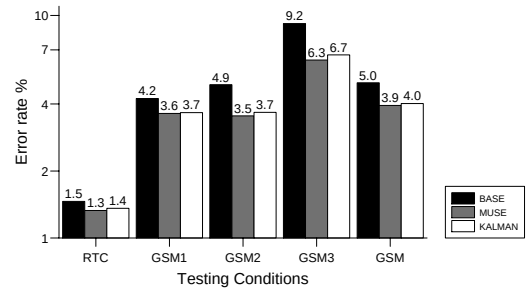

Figure 4: Error Rates on the 50-Word-Vocabulary Database

nition improvements. Those improvements are almost the same for both methods. Those results are consistant with the preliminary experiments 3.1.

4. CONCLUSION

In this paper, we gave a new view of bias compensation. This technique can be classified as a predictive approach, since it enables to compute the actual distribution of the data. The bias variations are explicitly model by a first order state equation. Thanks to Kalman recursions, it is possible to compute the likelihood of each path in order to perform the recognition process by a Viterbi search. We illustrated the properties of the technique on artificially generated data for two simple cases. Then, we showed that the method can efficiently compensate for the mismatch between PSTN recorded data and GSM-recorded data. In the future, it might also be interesting to study the effect of modelling the bias by a smoother function and to investigate the influence of the values given to the matrices $\mathbf{F}$ and $\mathbf{\Gamma}_\epsilon$. Furthermore, the traking possibilities of the method might be enlighten on longer recordings. It may also be possible to study the tracking of the parameters of a more complex transform. But, even for a simple case such as an affine transform, the estimation equations will become non-linear.

5. REFERENCES

[1] Delphin-Poulat L. and Mokbel C. (1997), Signal Bias Removal Using the Multi-path Stochastic Equalization Technique. Proceedings of Eurospeech, pp. 2575-2578, Rhodes, Greece.

[2] Hermansky H., Morgan N., Bayya A. and Kohn P. (1991), Compensation for the Effect of the Communication Channel in Auditory-Like Analysis of Speech. Proceedings of Eurospeech, pp. 1367-1370, Genova, Italy.

[3] Mauuary L. (1998), Blind Equalization in the Cepstral Domain for Robust Telephone Based Speech Recognition. Proceedings of Eusipco, pp. 359-362, Rhodes, Greece.

[4] Mokbel C., Jouvet D. and Monné J. (1995), Blind Equalization Using Adaptive Filtering for Improving Speech Recognition over Telephone. Proceedings of Eurospeech, pp. 1987-1990, Madrid, Spain.

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

[6] Picinbono B. (1995), Random Signal and Systems, Prentice Hall Signal Processing Series.

[7] Rahim M.G. and Juang B.-H. (1996), Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition, IEEE Trans. on Speech and Audio Processing, vol. 4, n. 1, pp. 19-30, Jan. 1996.

[8] Sankar A. and Lee C.-H. (1996), "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.

[9] Surendram A.C. and Lee C.-H. (1998), Predictive Adaptation and Compensation for Robust Speech Recognition. Proceedings of ICSLP, pp. 463-466, Sydney, Australia.