# TOWARDS INTRODUCING LONG-TERM STATISTICS IN MUSE FOR ROBUST SPEECH RECOGNITION

*Christopher Kermorvant and Chafic Mokbel*

IDIAP
P.O.Box 592, Martigny, Switzerland.
kermorva,mokbel@idiap.ch

## ABSTRACT

In this paper, we propose new developments of the MUlti-path Stochastic Equalization techniques (MUSE). The MUSE technique is based on an enriched model of speech, composed of both a classical model of clean speech with HMM and equalization functions. This technique is able to reduce the recognition error rate due to a mismatch between the training and testing conditions. In order to track long-term variation of this mismatch, the introduction of a priori statistics on the equalization function is studied. In the case of Bias Removal, this approach has been implemented in HTK and tested on the Numbers95 database. Experiments show that the convergence of the bias computation is fast enough and limits the effect of the a priori values. However, both the fast convergence property and the proposed framework open research directions towards more complex equalization functions.

## 1. INTRODUCTION

Recent advances made it possible to integrate speech recognition techniques in commercial products and applications. However, the performance of the recognisers still highly rely on the conditions in which they are used. If the recognition takes place in an environment which is close to the environment the recogniser has been designed for, high recognition scores can be achieved. But as soon as there is a mismatch between the training and testing environments, performance drops rapidly.

Many equalization scheme have been developed to reduce this mismatch both in the spectral [2][6] or in the cepstral [1][7] domain. However, the equalization process has always been separated from the recognition process.

The first attempt to combine an equalization scheme with HMM modeling during recognition was proposed by Ephraim [4]. Then, the Stochastic Matching technique [9] has been proposed : this technique uses a Maximum Likelihood approach to compute the parameters of a mapping function in order to reduce the mismatch between the observed utterance and the speech models during recognition. In this case, both the mapping function and the state sequence are optimized using the EM (Expectation-Maximisation) algorithm.

Recently, a Multi-path Stochastic Equalization (MUSE) technique has been developed [8]. MUSE provides an enriched model of speech signals. By combining usual HMM models and equalization functions, MUSE can model both the variations of the speech signals and the variations of the environment. In the case of bias removal, MUSE has already shown its ability to track local variation of the bias. In this paper, we introduce a method to learn and integrate long-term characteristics of the bias. The implementation of MUSE and of the proposed extension into a classical decoder, namely HTK[10], is also presented.

This article is organized as follow. The theoretical framework behind MUSE is recalled in Section 2. A particular application of MUSE to bias removal is developed in Section 3. In this section, the introduction of a priori statistics on the bias is also presented. Recognition experiments designed to assess this approach are presented and analyzed in Section 4. Finally, conclusions is given in Section 5.

## 2. THEORETICAL FRAMEWORK

The basic idea behind MUSE is to associate an equalization function to every possible state sequence hypothesized during the decoding. The parameters of the equalization function are computed using either a Maximum Likelihood or a Maximum A Posteriori criterion, as developed in [8]. In this section, we recall the the theoretical development of the technique using a Maximum A Posteriori (MAP) criterion.

We denote by $Y = \vec{Y_1}, \vec{Y_2}, \ldots, \vec{Y_T}$ a sequence of T speech frames observed at the output of a transmission channel and by $\lambda$ the parameter set of the HMM modeling clean speech. We suppose that we can obtain an estimate of the clean speech $X$ from the observed sequence $Y$ using an equalizing function $T_\theta(.)$ :

$$\widehat{\vec{X_t}} = T_\theta(\vec{Y_t}) \tag{1}$$

For a given path $s_t$ in the HMM model, we can derive the optimal values for the equalizing function parameter with a Maximum A Posteriori criterion :

$$\begin{aligned}\widehat{\theta}(s_t) &= arg\max_{\theta}[p_Y(\vec{Y_1},\dots,\vec{Y_t}|\theta(s_t),s_t,\lambda)\\ &\quad \cdot p(\theta(s_t)|s_t,\lambda)] \quad (2)\end{aligned}$$

$$\begin{aligned}\widehat{\theta}(s_t) &= arg\max_{\theta}\Bigg[\frac{p_X(T_{\theta(s_t)}(\vec{Y_1}),\dots,T_{\theta(s_t)}(\vec{Y_t})|s_t,\lambda)}{\prod_{\tau=t_0}^{t}\|J(\vec{Y_\tau})\|}\\ &\quad \cdot p(\theta(s_t)|s_t,\lambda)\Bigg] \quad (3)\end{aligned}$$

where $\|J(Y_\tau)\|$ is the absolute value of the Jacobian associated to the change of variable. With certain types of function $T_\theta(.)$, Equation 3 can be solved analytically to find the values of the equalization parameters $\theta$. If we assume, as it is generally the case in HMM, that the distribution of the state $i$ is Gaussian with a mean vector $\vec{\mu_i}$ and a covariance matrix $\Sigma_i$, then $\widehat{\theta}(s_t)$ is the solution of the following equation :

$$\begin{aligned}&\sum_{\tau=t_0}^{t}\Bigg\{\left[T_{\theta_t}(\vec{Y_\tau})-\vec{\mu}_{s_\tau}\right]^{\#}\cdot\Sigma_{s_\tau}^{-1}\cdot\left[\frac{\partial}{\partial\theta_t}T_{\theta_t}(\vec{Y_\tau})\right]\\ &+\left[\frac{\partial}{\partial\theta_t}\log\|J(\vec{Y_\tau})\|\right]\Bigg\}-\frac{\partial}{\partial\theta_t}\log p_{\Phi}(\theta_t|s_t,\lambda)=0_{1,q} \quad (4)\end{aligned}$$

Once $\theta$ is computed for each path, the most likely state sequence can be found by :

$$\begin{aligned}\widehat{s}_t &= arg\max_{s_t}[p(s_t|\vec{Y_1},\dots,\vec{Y_t},\theta(s_t),\lambda)]\\ &= arg\max_{s_t}\Bigg[\frac{p_X(T_{\theta(s_t)}(\vec{Y_1}),\dots,T_{\theta(s_t)}(\vec{Y_t})|s_t,\theta(s_t),\lambda)}{\prod_{\tau=t_0}^{t}\|J(\vec{Y_\tau})\|}\\ &\quad \cdot p(s_t|\lambda)\Bigg] \quad (5)\end{aligned}$$

## 3. INTRODUCTION OF LONG-TERM STATISTICS FOR BIAS REMOVAL

The approach presented in the previous section can be applied to a simple equalization function like a bias removal function in the form $T_\theta(\vec{Y_t})=\vec{Y_t}-\vec{b}$. In this case, in Equation 4, we have :

$$\frac{\partial}{\partial\theta_t}T_{\theta_t}(\vec{Y_\tau})=-I_p$$
$$\log\|J(\vec{Y_\tau})\|=0$$

where $I_p$ is the p-dimensional identity matrix. In order to introduce a priori statistics on bias, we suppose that having a bias $\vec{b_t}$ at time $t$ is equivalent to having $\vec{b}_1=\dots=\vec{b}_{t-1}=\vec{b}_t$. We also suppose that the a priori distribution of the bias is Gaussian with mean $\vec{\mu}_{apriori}$ and variance $\frac{\Sigma_{apriori}}{\alpha^t}$. The form of this distribution is based on the fact that the importance of the a priori distribution of the bias is high at the beginning of the decoding, and decreases with the time. The probability of a bias $\vec{b_t}$ at time $t$ given its a priori distribution is expressed as :

$$p(\vec{b_t}|s_t,\lambda)=\prod_{\tau=t_0}^{t}\mathcal{N}(\vec{b_t},\vec{\mu}_{apriori},\frac{\Sigma_{apriori}}{\alpha^t}) \quad (6)$$

Then, In Equation 4, we have

$$\frac{\partial}{\partial\theta_t}\log p_{\Phi}(\theta_t|s_t,\lambda)=\sum_{\tau=t_0}^{t}\alpha^t((\vec{b_t}-\vec{\mu}_{apriori})^{\#}\Sigma_{apriori}^{-1} \quad (7)$$

By solving Equation 4, we can derive an analytical formula to compute the bias :

$$\vec{b}=\frac{\left[\sum_{\tau=t_0}^{t}[\vec{Y_\tau}-\vec{\mu}_{s_\tau}]^{\#}\cdot\Sigma_{s_\tau}^{-1}+\alpha^t\vec{\mu}_{apriori}^{\#}\Sigma_{apriori}^{-1}\right]}{\left[\sum_{\tau=t_0}^{t}(\Sigma_{s_\tau}^{-1}+\alpha^t\Sigma_{apriori}^{-1})\right]} \quad (8)$$

With this bias value, the log likelihood of a sequence $Y$ of observation vectors $Y_\tau$ (of dimension $n$) given the states $S_t$ is now given by :

$$\begin{aligned}&log(p_X(T_{\theta_t}(Y)|S_t,\vec{b},\lambda))=-\frac{nt\log(2\pi)}{2}\\ &-\frac{1}{2}\sum_{\tau=t_0}^{t}\left\{log(||\Sigma_{s_\tau}||)+(Y_\tau-\vec{\mu}_{s_\tau})^{\#}\Sigma_{s_\tau}^{-1}(Y_\tau-\vec{\mu}_{s_\tau})\right\}\\ &+\left[\sum_{\tau=t_0}^{t}[\vec{Y_\tau}-\vec{\mu}_{s_\tau}]^{\#}\cdot\Sigma_{s_\tau}^{-1}\right]\vec{b}-\frac{1}{2}\vec{b}^{\#}\left[\sum_{\tau=t_0}^{t}\Sigma_{s_\tau}^{-1}\right]^{-1}\vec{b} \quad (9)\end{aligned}$$

In this equation, one can recognize the classical formula for the likelihood computation (first and second line) plus two terms associated with the bias (third line).

## 4. EXPERIMENTS

### 4.1. Experimental Setups

The speech database chosen for the experiments is the Numbers'95 database [3] from the Center for Spoken Language Understanding (CSLU). This database contains digits sequences continuously spoken over the telephone. We used the 3590 sentences of the training set for the training of our models and we tested them on the 1206 sentences of the development-test set.

We used the front-end developed at IDIAP to extract the feature vectors from the speech files. We computed 26 mel-scaled filter bank coefficients, over a 32 ms hamming window, with a 10 ms shift. Then 13 mel-cepstral coefficients were derived together with their first and second order derivatives (for a total of 39 coefficients).

The recognition system was based on Gaussian mixture HMMs. It was trained with HTK [10]. The system was composed of 81 triphones modeled by 3 states HMMs; each state had a 10 Gaussian mixture pdf and a diagonal covariance matrix. No language model was used. The recognition has been done with a modified version of HTK in which has been implemented the MUSE technique.

### 4.2. Implementation issues

#### *4.2.1. Bias computation with a priori statistics*

In order to compute the bias, we used two accumulators $\vec{A}_t$ and $\vec{B}_t$ defined as :

$$\begin{aligned}\vec{A}_t &= \vec{A}_{t-1} + (\vec{Y}_t - \vec{\mu}_{s_t})^{\#}\Sigma_{s_t}^{-1} + \alpha^t \vec{A}_{apriori} \cdot \vec{B}_{a_priori} \\ \vec{B}_t &= \vec{B}_{t-1} + \Sigma_{s_t}^{-1} + \alpha^t \vec{B}_{apriori}\end{aligned}$$

These two accumulators are updated with each new frame and with a priori values of the bias. We have computed $\vec{A}_{apriori}$ and $\vec{B}_{apriori}$, using the first and second order statistics of the bias,on the training database. The multiplicative factor $\alpha^t$ exponentially decreases with time ($0 < \alpha < 1$), and allows to give more importance to the a priori estimation of the bias at the beginning of the utterance. Several values have been tested for the parameter $\alpha$. Results are shown in the Section4.3. For each frame, once the two accumulators are updated, the bias is then computed as $\vec{b} = \vec{A}_t \cdot \vec{B}_t^{-1}$.

#### *4.2.2. On-line estimation of a priori statistics*

We have also introduced an on-line process to estimate the a priori estimation of the bias. At the end of each utterance, the a priori means and variances of the bias are updated recursively with the current value of the bias with parameter $\beta$ and $1 - \beta$. This allows to track the long-term variation of the bias, and therefore to adapt on-line the a priori statistics of the bias.

### 4.3. Results

We first present the results of an adaptation to a mismatch between training and testing conditions using MUSE. The best recognition results we obtained so far on the Numbers95 database was obtained by pre-processing the data with a Cepstral Mean Subtraction (CMS) scheme [5]. However, this preprocessing is not frame synchronous, and thus can not be used in real systems. Using classical MFCC feature vectors with the models trained on data pre-processed with CMS yields a decrease in recognition results. As shown on table 1, the recognition word error rate increase from 5.33% to 11.16% when using MFCC feature vectors with models trained with CMS. This bias can be significantly reduce by using MUSE. Indeed, using MUSE reduces the WER from 11.16% to 6.10% and almost recovers the WER obtained in matching conditions. It is also important to note that using MUSE in mismatch conditions (MFCC features with CMS based models) outperforms the baseline system, based on both MFCC features and models

| | baseline | CMS/CMS | MFCC/CMS | MUSE |
|---|---|---|---|---|
| WER | 7.07 | 5.33 | 11.16 | 6.10 |

Table 1: *Recognition error rate on the Numbers95 database in matching conditions (CMS/CMS), mismatch conditions (MFCC/CMS) and mismatch condition with MUSE (MUSE).*

The second set of results concerns the introduction of a priori statistics on the equalization function parameters in the MUSE technique. These a priori statistics have been computed on the train set and consist in the mean bias and its variance for each cepstral coefficient. The data are then used in the MUSE decoding process as described in the previous section.

Table 2 presents the variation of the recognition Word Error Rate according to the value of the factor $\alpha$. First, when $\alpha$ increases, the recognition score decreases. This is due to the fact that the a priori statistics should be used only at the beginning of the recognition. Therefore, $\alpha$ should be small enough not to hinder the convergence of the estimated bias. Second, even with a small value for $\alpha$, we get no improvement of the recognition results.

Two reason might explain why the introduction of a priori statistics on the bias does not improve the recognition scores. The first one is illustrated in Figure 1. This figure shows the convergence of the estimated bias compared to the real bias computed by a forced Viterbi alignment. We can see that the convergence of the estimated bias is quite fast. This means that, in the case of bias removal, the estimation of the parameters of the equalization function is simple enough not to need many frames to converge. The introduction of the a priori values for the bias is therefore of no help for the convergence. Secondly, when looking at the a priori values of the bias computed on the training set, we see that these values present a very large variance. The effect of these values in the computation of the bias is thus very limited.

| | 0.5 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|
| WER | 6.17 | 6.17 | 6.21 | 6.47 |

Table 2: *Recognition error rate on the Numbers95 database in mismatch condition with MUSE for different values of the factor $\alpha$.*

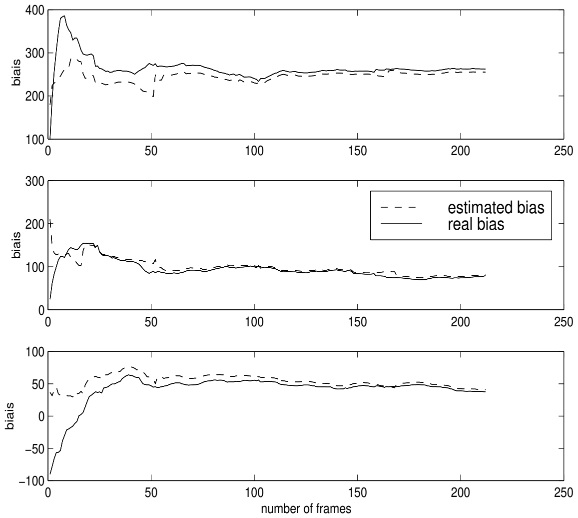


Figure 1: *Convergence of the estimated bias compared to the real bias for the three first cepstral coefficients*

However, these results give us some perspective for the future work. Firstly, the fast converge of the MUSE algorithm allows to consider more complex equalization functions. Secondly, the MAP approach to MUSE allows to introduce equalization functions which depend on phonemes. In these two cases, the introduction of a priori statistics on the parameter of the equalization functions , as proposed in this paper, will be necessary.

## 5. CONCLUSION

In this paper, we have proposed a development of the MUSE technique towards the introduction of a priori statistics on equalization functions, in the case of Bias Removal. The experiments on the Numbers95 database have shown that in the case of Bias Removal, the convergence of the MUSE algorithm is quite fast. Therefore, the introduction of the a priori statistics does not improve the recognition score. However, the proposed framework for the introduction of adaptative a priori statistics and the good convergence properties of MUSE open research directions towards more complex equalization functions.

## Acknowledgments:


This work is supported by the Swiss Federal Office for Education and Science (FOES) through the European TMR SPHEAR and RESPITE projects.


## 6. REFERENCES


[1] B. S. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *JASA*, 55(6):1304–1312, June 1974.

[2] Steven F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2), April 1979.

[3] R.A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at cslu. *Proc. European Conf. on Speech Communication and Technology*, 1:821–824, 1995.

[4] Yariv Ephraim. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80(10):1555, October 1992.

[5] C. Kermorvant. A comparison of noise reduction techniques for robust speech recognition. IDIAP-RR 10, 1999.

[6] L. Mauuary. Blind equalization for robust telephone based speech recognition. In *Proc. European Signal Processing Conference*, 1996.

[7] L. Mauuary. Blind equalization in the cepstral domain for robust telephone based speech recognition. In *Proc. European Signal Processing Conference*, 1998.

[8] C. Mokbel. Muse : Multipath stochastic equalization. a theoretical framework to combine equalization and stochastic modeling. In *ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channel*, Pont-à-Mousson, April 1997.

[9] A. Sankar and C.H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 4(3):190–202, May 1996.

[10] Steve Young. *The HTK Book*. Cambridge University, March 1997.