On-Line Frame-Synchronous Noise Compensation

Vincent Barreaud and **Irina Illina** and **Dominique Fohr**

LORIA/INRIA
57602 Villers-lès-Nancy FRANCE
{barreaud,illina,fohr}@loria.fr

ABSTRACT

We present a frame-synchronous noise compensation algorithm that uses Stochastic Matching approach to cope with time-varying unknown noise. This method proposes to estimate simple mapping function in parallel with Viterbi alignment. The technique is entirely general since no assumption is made on the nature, level and variation of noise. Our algorithm is evaluated on the VODIS database recorded in a moving car. For various tasks, our technique outperforms significantly classical methods. For instance, using a simple additive bias the proposed algorithm gives an error rate improvement of 13.3 % compared to Parallel Model Combination (PMC), 15.5 % on Spectral Subtraction (SS) and 27.8 % on frame-synchronous Cepstral Mean Subtraction (CMS) for the numbers recognition task in a moving car.

1 INTRODUCTION

An automatic speech recognition (ASR) system experiences a significant degradation of its performances when used in an environment that does not match its training environment. This mismatch is due mostly to additional noise sources and discrepancies in channels and speakers that transform a clean speech sequence X into a distorted speech sequence Y. Those mismatch sources may be non-stationary and little a priori information about them is available.

Several techniques have been proposed to enhance speech in a robust manner. Two possible approaches can be explored. First, the parameters of the acoustic models can be modified to make the transformed stochastic models better characterize the distorted features. This approach, called adaptation, brings together several techniques such as PMC [1], MAP [2] and MLLR [3]. Second, the corrupted features can be adjusted thanks to a transformation that is estimated from the noise characteristics. This set of methods, called compensation, gathers techniques such as CMS and Stochastic Matching [4]. The method developed here belongs to this category.

When acoustic environments are known to be non-stationary, three types of compensation methods can be used. First, noise and channel can be characterized by models trained on prior measurement of the environment [5]. Second, a bank of Kalman filters can be used to compensate the effect of time-varying noise [6]. Finally, sequential estimation algorithms can be used to track slowly time-varying environments. [7] uses this framework to estimate an additive bias and [8] uses this approach to perform frame-synchronous stochastic matching. The idea of stochastic matching is to reduce the mismatch between the distorted features of Y and the acoustic model Λ_X associated to the clean feature of X.

Frame synchronous sequential algorithms are naturally appealing to cope with non-stationary noise sources even if these algorithms often face convergence problems linked to the scarcity of data. One of the most popular frame synchronous technique is Cepstral Mean Subtraction (CMS): at each frame, the mean of the incoming sequence of cepstra is calculated. This mean is then subtracted to the observation.

We believe that this last method can be enhanced by taking into account statistics derived during the recognition process. Our aim is to merge the simplicity of CMS and the efficient and intuitive approach of stochastic matching. Our work is based on [8] where the mismatch between training and testing environment is estimated in order to maximize the Kullback-Leibler information. The basic idea of [8] was to compute the parameters of a linear transform according to the distance of the observations sequence to a partial sequence of models. Those derivations led to a recursively updated bias which expression was close to the one obtained in [4] with a Maximum-Likelihood approach.

The paper is organized as follow. Section 2 exposes the stochastic recognition process using Hiden Markov Models. Then section 3 presents the proposed compensation algorithm and section 4 gives some experimental results. Finally, conclusions are given in section 5.

ISBN 1-876346-48-5

2 RECOGNITION USING HMM

2.1 HIDDEN MARKOV MODELS

The main idea of ASR is pattern-recognition. Most of state-of-the-art recognition applications use a template based approach: observations are classified according to their distances to acoustic models. Hidden Markov Models (HMM) are used as acoustic models to characterize the spectral properties of speech frames. They are used to determine the probability of a word under the hypothesis of a sequence of observations. The basic units of speech (phones, for example) are described as a sequence of HMM states. The acoustic observations associated to each state are modeled through a probability density functions. Typically, a mixture of continuous observation density function is of the following formulation:

$$p(x_t|s_t = n) = b_n(x_t) = \sum_{k=1}^{K} c_{(n,k)} \mathcal{N}(x_t, \mu_{(n,k)}, \Sigma_{(n,k)})$$

where x_t and s_t are the observed vector and the state at time t, $\mathcal{N}$ is Gaussian with mean vector $\mu_{(n,k)}$ and covariance matrix $\Sigma_{(n,k)}$ for the k-th mixture component in state n and $c_{(n,k)}$ is the mixture coefficient for the k-th mixture in state n [9]. The order of the states is governed by the transition probabilities between the states. The following equation gives the transition probability between states i and j:

$$a_{i,j} = p(s_{t+1} = j|s_t = i) \text{ for j>=i}$$
$$a_{i,j} = 0 \text{ otherwise}$$

In the rest of this paper, let us consider a Hidden Markov Model recognition system of N-states models with diagonal covariance matrices. Each state n is characterized by a mixture of K gaussian probability functions of mean $\mu_{(n,k)}$ and variance $\sigma_{(n,k)}$. The one dimension case is treated, all the derivation being easily extended to multidimensional case.

2.2 THE VITERBI DECODING

During recognition, all models are concatenated in order to form a lattice of possible word sequences or paths. Then, Viterbi algorithm is used to find the best state sequence S^{opt} given the speech sequence. More precisely, this algorithm maximizes the likelihood of a state sequence $S_T = \{s_1, s_2, \ldots, s_T\}$ given a sequence of observations $X_T = \{x_1, x_2, \ldots, x_T\}$ and the acoustic models λ:

$$S^{opt} = \underset{S_T}{argmax}\ p(S_T|X_T, \lambda)$$
$$p(S_T|X_T, \lambda) \approx p(X_T|S_T, \lambda).p(S_T)$$
$$p(X_T|S_T, \lambda) = \prod_{t=1}^{T} p(x_t|s_t, \lambda)$$

The best state sequence is obtained according to:

$$\alpha_t(i) = \underset{s_1 s_2 \ldots s_{t-1}}{argmax}\ p[s_1 s_2 \ldots s_{t-1}, s_t = i, x_1 x_2 \ldots x_t|\lambda]$$

often called *forward probability*, and calculated as:

$$\alpha_t(i) = [\underset{j}{argmax}\ (\alpha_{t-1}(j)a_{ji})].b_i(x_t)$$

3 FRAME-SYNCHRONOUS COMPENSATION

3.1 CORRUPTED SIGNAL

We assume that the clean signal spectrum x_s is distorted by noise and gives y_s (the s subscript denotes the spectral domain). It can be modeled as follows:

$$y_s = h_s \otimes x_s + n_s \quad (1)$$

where $\otimes$ is the convolution operator, h_s is the channel noise, n_s is the additive noise. In the cepstral domain, (1) becomes:

$$y = x - d(x_s, n_s, h_s) \approx x - d(y)$$

where $d(x_s, n_s, h_s)$ is a non-linear function without any regular expression. Usually, the exact values of x_s, n_s and h_s are unknown. In practice, this function is approximated by $d(y)$. The goal of compensation is to find a transformation f such that $f(y)$ approaches x: $f(y) \approx x$.

3.2 PROPOSED COMPENSATION

The stochastic matching framework supposes the use of stochastic distance of the noisy observation to a state sequence. For the batch estimation of mismatch function, this distance is calculated using the optimum state sequence [4]. But for a frame synchronous method, only partial state sequences are avaliable. In this case, two solutions can be envisaged: statistics can be derived on short windows as in [8], or approximated by *forward probabilities* as proposed in this paper.

The basic idea of our method is as follows. First, the hypothesis is made that at each time instant during the Viterbi alignment, the states linked to the highest *forward* probabilities give a good modelisation of the speech observations. Then, the parameters of the mismatch function are estimated in order to maximize the likelihood of the observation given those states. Consequently, this on-line algorithm performs compensation in parallel with recognition and does not need any *a priori* information on the nature of the noise. Compensation transform is estimated frame per frame and confidence in its parameters is gained as forward probabilities computation goes along.

3.3 THEORETICAL FRAMEWORK

Consider θ as the set of parameters of a transformation $f_\theta(y)$ from the testing observation space into the training space. It has been shown in [8] that the set θ maximizing the Kullback-Leibler information

ISBN 1-876346-48-5

$J(\theta) = E\{log(p(Y_t|\theta)\}$ can be approximated by a sequence $\{\theta_i\}$ that maximizes the auxiliary function Q:

$$\theta_{t+1} = \underset{\theta}{argmax}\ Q_{t+1}(\Theta_t, \theta)$$

$$Q_{t+1}(\Theta_t, \theta) = \sum_{\tau=1}^{t+1} L_{\tau|t+1}(\Theta_{\tau-1})$$

with $\Theta_t = (\theta_0, \ldots, \theta_t)$. The auxiliary function is defined by the following expression of likelihood:

$$L_{\tau|t+1}(\Theta_{\tau-1}) = \log(|f'_\theta(y_\tau)|) - \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{\tau|t+1,\Theta_{\tau-1}}(n,k)\frac{(f_\theta(y_\tau)-\mu_{(n,k)})^2}{\sigma^2_{(n,k)}}$$

In which $f'_\theta(y_\tau)$ is the partial derivative of the compensation function with respect to the observation y_τ for the time frame τ and $\gamma_{\tau|t+1,\Theta_{\tau-1}}(n,k)$ is the probability that the τ-th emitting state s_τ being n and its principal Gaussian component g_τ being k knowing the sequence of observations $Y_{t+1} = \{y_1, \ldots, y_{t+1}\}$ and $\Theta_{\tau-1}$.

Let a simple transformation $f_B(y_{t+1}) = y_{t+1} + b_t$. Then the bias parameters $B_t = \{b_0, \ldots, b_t\}$ can be estimated over the optimum Viterbi path:

$$b_{t+1} = b_t - \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{t+1|t+1,B_t}(n,k)\frac{y_{t+1}+b_t-\mu_{(n,k)}}{\sigma^2_{(n,k)}}}{\sum_{\tau=1}^{t+1}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\gamma_{\tau|t+1,B_{\tau-1}}(n,k)}{\sigma^2_{(n,k)}}} \quad (2)$$

where

$$\gamma_{\tau|t+1,B_{\tau-1}}(n,k) = p(s_\tau = n, g_\tau = k|Y_{t+1}, B_{\tau-1})$$

Equation (2) converges toward an optimum bias that maximizes the likelihood of a state sequence.

The $\gamma_{\tau|t+1,B_{\tau-1}}(n,k)$ probability is unavailable during alignment. In our algorithm, we make the hypothesis that the *forward probability*

$$\alpha_{\tau|B_{\tau-1}}(n,k) = p(Y_\tau, s_\tau = n, g_\tau = k|B_{\tau-1})$$

could be used instead of γ in equation (2) and leads to the following expression:

$$b_{t+1} = b_t - \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\alpha_{t+1|B_t}(n,k)\frac{y_{t+1}+b_t-\mu_{(n,k)}}{\sigma^2_{(n,k)}}}{\sum_{\tau=1}^{t+1}\sum_{n=1}^{N}\sum_{k=1}^{K}\frac{\alpha_{\tau|B_{\tau-1}}(n,k)}{\sigma^2_{(n,k)}}} \quad (3)$$

Equation (3) can be simplified: we assume that the sums over all possible states and Gaussian components at time τ can be fairly approximated by the contribution of the pair (n,k) that maximizes $\alpha_{\tau|B_{\tau-1}}(n,k)$ alone. Let $(n,k)_\tau$ be that pair. Then (3) becomes:

$$b_{t+1} = b_t - \frac{\frac{y_{t+1}+b_t-\mu_{(n,k)_{t+1}}}{\sigma^2_{(n,k)_{t+1}}}}{\sum_{\tau=1}^{t+1}\frac{1}{\sigma^2_{(n,k)_\tau}}} \quad (4)$$

In the case of an affine transformation function $f_t(y_{t+1}) = A_t.y_{t+1} + b_t$, the parameters (A_t, b_t) can be developped in the same manner. This transform will be called *Affine* in the following.

3.4 IMPLEMENTATION OF THE ALGORITHM

In our method, computation of the bias at time t does not require backtracking along a path: at each frame t, the most probable state in the *forward probability* sense is used to re-estimate the transformation parameters. The compensation algorithm for the simple bias transform can then be described as follows:
1) initialization: $b_0 := 0$, $t := 0$;
2) at time t, compute, for each (n,k)
$\alpha_{t|B_{t-1}}(n,k) := p(y_t + b_{t-1}, s_t = n, g_t = k)$
used in the Viterbi alignment;
3) at time t, compute b_t according to equation (4);
4) $t := t + 1$;
5) if $t = T$ exit, else return to step 2.

4 EXPERIMENTAL FRAMEWORK

4.1 VODIS DATABASE

All the experiments have been conducted on the Voice-Operated Driver Information Systems (VODIS) Database. This corpus collects 200 french speakers. The speakers were divided into two sets: the training set (*Training*, 159 speakers) and the test set (*Test*, 41 speakers). Sentences were pronounced in french, in a moving car with various driving situations (opened window, traffic/highway, radio). Speakers were asked to utter phone numbers (*phone numbers* task, 95% confidence interval is ±1%) and numbers up to 12000 (*numbers* task, 95% confidence interval is ±1%). Notice that french phone numbers are composed of numbers ranging from 0 to 99. The speech sequences have been collected by two microphones, synchronously. The first microphone (*close talk*) was placed close to the mouth of the speaker and collected "clean" speech with an average Signal to Noise Ratio (SNR) of 20.7 dB. The second one was placed on the rear-view mirror and collected distorted speech with an average SNR of 10.8 dB (*far-talk*). The signal was sampled at 11025 Hz, and encoded in 36 dimensions cepstra sequence composed by 12 MFCC, 12 Δ and 12 ΔΔ. We used 3-states phoneme models, each state composed of a mixture of 8 Gaussian probability density functions. The models were trained on all the *close-talk* utterances of *Training* set whereas experiments were made on the *far-talk* utterances of *Test* set.

4.2 EXPERIMENTAL RESULTS

Table 1 represents the results of classical compensation methods (Baseline, CMS, SS, PMC) and transforms given by our algorithm (*Bias* and *Affine*). It

ISBN 1-876346-48-5

shows that our method outperforms significantly all classical methods for both tasks. For the *numbers* task, *Affine* method gives an error rate improvement of 13.3 % compared to PMC, 15.5 % compared to SS and 27.8 % on frame-synchronous CMS. For the *phone numbers* task, *Bias* method gives an error rate improvement of 10.5 % compared to PMC, 20.2 % compared to SS and 14.1 % on frame-synchronous CMS.

	Baseline	CMS	SS	PMC	Bias	Affine
numbers	63.5	67.3	72.1	72.8	72.9	**76.4**
phone numbers	78.6	80.8	79.3	81.6	**83.5**	**86.3**

Table 1: Word accuracy on far-talk test set(SNR:10.8dB).

4.3 STUDY OF BIAS INITIALISATION

As reported in section 3.4, the initial value of bias is set to 0 at the beginning of each new sentence. After a period of fluctuations, the bias converges to its final value. In a real life ASR application, the variability in environments and speakers between two successive utterances can be assumed low. Hence, at the beginning of a sentence, the initial value of the bias could be set to the final value of the bias computed during the previous sentence. In this way, bias convergence is obtained more rapidly. Figure 1 represents the evolution of the bias computed on the second dimension of the observation space (c_1), for one *phone number* sentence without any initialization (plain line) and when initialized with the bias obtained during the previous sentence (doted line). A direct implementation of this initialisation method on VODIS database did not give significant improvements compared to previous results.

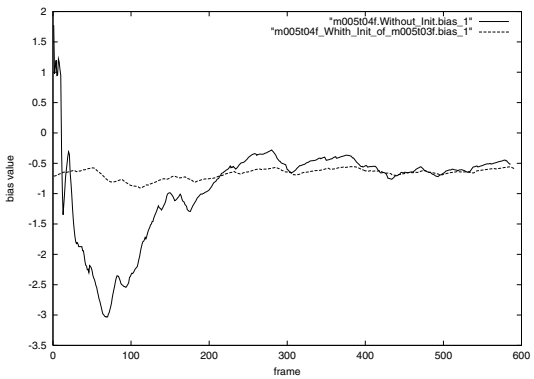

Figure 1: Influence of initialization on the evolution of the bias of the first cepstral dimension

5 CONCLUSION

This article presents an on-line frame-synchronous noise compensation algorithm using the theoretical framework of Stochastic Matching. This algorithm does not need any *a priori* information on the environment and compensates non-stationary noise. The basic idea is to use *forward probabilities* in the estimation of the simple affine transformation's parameters. To evaluate the algorithm we have chosen to recognize *numbers* and *phone numbers* pronounced in a moving car. For both these tasks, our method significantly outperforms the frame-synchronous CMS, spectral subtraction techniques and PMC. Moreover, our technique does not require any specific models training and, thus, can be used along with other compensation techniques. Future work will involve studies of class-specific transforms based on a tree structure.

REFERENCES

[1] M.J.F. Gales, *Model-Based Techniques for Noise Robust Speech Recognition*, Ph.D. thesis, Gonville and Caius College, September 1995.

[2] J.-L. Gauvain and C.-H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transaction on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[3] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[4] A. Sankar and C.H. Lee, "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Transaction on Speech and Audio Processing*, pp. 190–202, 1996.

[5] A. Varga and R.K. Moore, "Hidden Markov Model Decomposition of Speech and Noise," in *ICASSP*, 1990, pp. 845–848.

[6] N.S. Kim, "Time-Varying Noise Compensation Using Multiple Kalman Filters," in *ICASSP*, 1999, pp. 1540–1543.

[7] N.S. Kim, D.K. Kim, and S.R. Kim, "Application of Sequential Estimation to Time Varying Environment Compensation," in *IEEE Workshop on Speech Recognition and Understanding*, 1997, pp. 389–395.

[8] L. Delphin-Poulat, C. Mokbel, and J. Idier, "Frame Synchronous Stochastic Matching Based on the Kullback-Leibler Information," in *ICASSP*, 1998, pp. 89–92.

[9] L.R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

ISBN 1-876346-48-5