Structural State-Based Frame Synchronous Compensation

Vincent Barreaud, Irina Illina, Dominique Fohr, Filipp Korkmazsky

Speech Group, LORIA/INRIA
54602 Villers-Les-Nancy, France
{barreaud,illina,fohr,korkmazs}@loria.fr

Abstract

In this paper we present improvements of a frame-synchronous noise compensation algorithm that uses Stochastic Matching approach to cope with time-varying unknown noise. We propose to estimate a hierarchical mapping function in parallel with Viterbi alignment. The structure of the transformation tree is build from the states of acoustical models. The objective of this hierarchical transformation is to better compensate non-linear distortions of the feature space. 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, proposed technique significantly outperforms classical compensation/adaptation methods.

1. Introduction

An automatic speech recognition (ASR) system gives a significant degradation in performances when used in an environment that does not match its training environment. This mismatch is mostly due to additional noise sources and discrepancies in channels and speakers. 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. First, the parameters of the HMMs can be modified to better characterize the distorted features. This approach, called adaptation, regroups several techniques such as PMC [1], MAP [2] and MLLR [3]. Second, the corrupted features can be adjusted with a transformation that is estimated from the noise characteristics. This set of methods, called compensation, includes techniques such as Spectral Subtraction (SS), Cepstral Mean Normalization (CMN) [4] and Stochastic Matching [5]. The method developed here belongs to this set.

Frame synchronous algorithms are naturally appealing to cope with non-stationary noise sources even if they often face convergence problems linked to the scarcity of data. One of the most popular frame synchronous technique is Cepstral Mean Normalization: at each time frame, the mean of the incoming sequence of cepstra is estimated. This mean then subtracted from each observation. Our work is based on frame synchronous algorithm developed in [6], where an approximation of the mismatch function was performed in order to reduce the Kullback-Leibler information. Those derivations led to a recursively updated bias which expression was close to the one obtained in [5] with a Maximum-Likelihood approach. Compared to [5], where the batch estimation of mismatch function is derived, we use a frame per frame approach. This on-line algorithm performs compensation in parallel with recognition and does not need *a priori* information on the nature of noise.

To improve the results of this method, we propose a structural state-based transformation. This approach is motivated by several observations. First, it is often assumed that observations which are very similar will be affected in a similar manner by variations in the environment. Hence, a set of subspace-specific transformations should give better results.

Second, as explained in [7], subspace-specific transformations face a data scarcity problem that can be overcome by the use of hierarchical transformation: a tree of transformations. For each node of this tree, a transformation function is estimated according to the observations of the current sentence. If the transformation associated with a node is poorly estimated, its parent will be used. Similar structural approach is used for the Stochastic Matching approach proposed in [8]. Compared to the approach of [8], where each node regroup Gaussian components of the acoustic models, in our approach each node regroup the states of the models.

In section 2, a brief reminder of the theoretical framework, developed in [6], is given. After, the construction and use of the tree structure in the hierarchical version of our algorithm is proposed. In section 4, experimental results that compare the structural and non structural version of our algorithm are presented. In the same section, the mismatch function initialization and the continuity of the transform are studied. Finally, in section 5, we draw conclusion and describe future work.

2. Theoretical Framework

The following is a brief summary of the non-structural version of our frame synchronous compensation algorithm, developed in [6]. 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 means $\mu_{(n,k)}$ and variances $\sigma_{(n,k)}$ and weights $w_{(n,k)}$.

Consider θ as the set of parameters of a transformation $f_\theta(y)$ from the testing observation space to the training space. It has been shown in [9] that the set θ maximizing the Kullback-Leibler information $J(\theta) = E\{log(p(Y_t|\theta)\}$ can be approximated by a sequence $\{\theta_i\}$ maximizing 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:

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

where

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

Equation (1) 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 (1) and leads to the following expression:

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

Equation (2) 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.

3. Structural State-Based Compensation

3.1. Motivation

It is often assumed that observations which are very similar will be affected in a similar manner by variations in the environment. Consequently, a global compensation transform as in equation (2) is inevitably a suboptimal solution for any subspace of the test space. A set of subspace-specific transformations should probably give better results than a global transformation [1].

To perform the partition of the test space in the subspaces, the states of acoustic model HMMs can be used. Indeed, each state models some subspace of feature space. During the recognition, observation y_t is associated to a particular state according to Viterbi alignement. Consequently, y_t is associated to a particular subspace corresponding to this state and will be compensated using state specific transformation.

Thus a possible evolution of the algorithm presented in section 2 is to associate a specific transformation $f_{\theta_t^n}$ to each state n or to cluster of states. In this case, the compensation algorithm will be as follow:

1) initialization: $t := 0$;
2) at time t, compute, for each $n \in 1, \ldots, N$ $\alpha_{t|B_{t-1}}(n,k) := p(f_{\theta^n_{t-1}}(y_t), s_t = n, g_t = k|B_{t-1})$ used in the Viterbi alignment;
3) at time t, compute $\theta_t^n = b_t^n$;
4) $t := t + 1$;
5) if $t = T$ exit, else return to step 2.

This solution can face a data scarcity problem. Indeed, θ_t^n need several time frames to achieve convergence [6]. Thus, during the compensation of one sentence, θ_t^n of rarely used states may be badly estimated and will be unusable. Consequently, a hierarchical partition of the state space have been considered: each node of the tree structure contains collections of states. For each node of this tree, a transformation function is estimated. During compensation, if the transformation associated with a node is not properly estimated, then transformation associated with its father node may be used.

3.2. Tree Construction

The tree construction is done prior to recognition. To built the tree, a bottom-up approach is used. We built a binary tree from the acoustic models. Each node contains a collection of states (Gaussian mixtures) and not Gaussian components of Gaussian mixtures, contrary to trees used in adaptation techniques.

First, each state of the acoustic models are associated with a node. After, each node is merged with the closest node to form a higher-level node according to the following distance measure:

$$D(i,l) = \sum_{k_i=1}^{K}\sum_{k_j=1}^{K} w_{(i,k_i)}w_{(l,k_l)}KL((i,k_i),(l,k_l)) \quad (3)$$

where $D(i,l)$ is the distance between states i and l and $KL((i,k_i),(l,k_l))$ is the Kullback-Leibler distance of k_i-th Gaussian of state i to the k_l-th Gaussian of state l. The distance between two nodes is the sum of the distances between the states of the first node and those of the second node.

This operation is repeated until all the nodes are regrouped into a single node: *root* node. Figure 1 represents an example of the tree, build from 4 states of HMM: s_1, s_2, s_3, s_4.

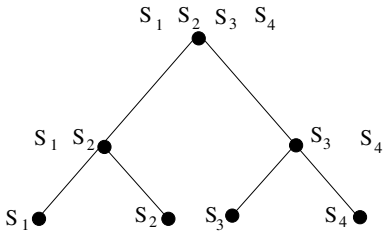

Figure 1: *Example of tree for* 4 *states.*

For every tree node (d,j), where d is a depth and j is a depth specific node number, a transformation $\theta_{(d,j,t)}$ is estimated at each time instant t as described in the next section.

3.3. Using the Tree During the Compensation

At time t, for a given state i, all the nodes, containing this state, update their estimation of $\theta_{(d,j,t)}$ as follow:

$$b_{(d,j,t+1)} = b_{(d,j,t)} - \frac{\frac{y_{t+1}+b_{(d,j,t)}-\mu_{(i,k)_{t+1}}}{\sigma^2_{(i,k)_{t+1}}}}{den_{(d,j,t+1)}}$$

$$den_{(d,j,t+1)} = den_{(d,j,t)} + \frac{1}{\sigma^2_{(j,k)}} \quad (4)$$

[1] This transformation uses a single set of state-independent parameters θ_t for the entire test space.

Thus, a node (d, j) of the tree contains the set of states, specific to this node, the frame specific transformation $b_{(d,j,t)}$ and cumulation of covariances $den_{(d,j,t)}$.

Consequently, the final compensation algorithm is:

1) initialization: $t := 0$, $b_{(d,j,0)} := 0$, $den_{(d,j,0)} = 0$;
2) at time t, compute, for each HMM state n
   $\alpha_{t|B_{t-1}}(n) := p(f_{\theta_n(t-1)}(y_t), s_t = n)$
   used in the Viterbi alignment
3) at time t, for each node j of each depth d
   which contain HMM state n, compute $b_{(d,j,t)}$
   according to equation (4);
4) at time t, choose for the compensation the deepest
   transformation $b_{(d,j,t)}$. If the number of
   observations used to estimate $b_{(d,j,t)}$ is too
   small (compared to some threshold T) then
   instead of $b_{(d,j,t)}$ its parent is used.
5) $t := t + 1$;
6) 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 $\pm 1\%$) and numbers up to 12000 (*Numbers* task, 95% confidence interval is $\pm 1\%$). 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 $\Delta\Delta$. 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.

4.2. Results

Experiments were conducted on the *far-talk* part of the test set. In table 1, the results of the non-structural version of our algorithm (*Bias*) are compared with those of classical compensation and adaptation techniques: *Cepstral Mean Normalization* (CMN), *Spectral Subtraction* (SS) and *Parallel Model Compensation* (PMC). We can notice that all methods are frame synchronous methods. The notation *Baseline* means recognition results without adaptation/compensation. The table shown that the best results are obtained using our frame synchronous method.

Method	*Baseline*	*CMN*	*SS*	*PMC*	*Bias*
Numbers	63.5	67.3	72.1	72.8	73.2
Phone Numbers	78.6	80.8	79.3	81.6	83.4

Table 1: *Word Accuracy (%) for Numbers and Phone Numbers tasks.*

Table 2 represents results using the Structural State-Based version of our algorithm proposed in this paper. Those results are presented in function of the tree depth. The depth of 0 corresponds to no tree (only root node). For all the experiments, the minimum number of frame necessary to start using a node-specific bias was set to 10 frames. Moreover, the transformation was not used on the first 10 frames.

Tree Depth	0	1	2	3
Numbers	73.2	73.7	71.8	71.6
Phone Numbers	83.4	85.2	84.3	83.5

Table 2: *Word Accuracy (%) for Phone Numbers and Numbers tasks using proposed Structural-State-Based algorithm.*

As can be observed in table 2, the best performance is obtained for the tree of depth 1. The performances decrease for values of depth superior to 1. Two hypothesis have been proposed to explain this behaviour. First, lower nodes may suffer from a bad initialization of the bias for the first frames. Indeed, for each sentence the bias is estimated independently of previous sentence (for $t = 0$, $b_{(d,j,0)} = 0$). Second, the discontinuity in the transformation may disturb the convergence of state-specific biases. In the following section we study these problems.

4.3. Bias Initialization and Smoothing

As we said, better *bias initialization* can improve the recognition results. Indeed, a properly initialized bias converge more quickly and be rapidly available for compensation. We propose to initialise all the biases $\{b_{(d,j,0)}, \forall d, \forall j\}$ with the final set of biases obtained during the recognition of the sentence, uttered in a similar environment. As similarity measure, the SNR level is used. The test set has been segmented into classes of similar SNRs:

- The *Numbers* test set was segmented in 4 clusters (SNR$<$ 3 dB, 3 dB$\leqslant$SNR$<$ 9 dB, 9 dB$\leqslant$SNR$<$ 14 dB and SNR$\geqslant$ 14 dB).
- The *Phone Numbers* task set was segmented in 3 clusters (SNR$<$ 7 dB, 7 dB$\leqslant$SNR$<$ 14 dB and SNR$\geqslant$ 14 dB).

In order to reduce discontinuities in the transformation, a *smoothing* procedure has been used. As the root node contains all HMM states, its transformation is updated at each time frame and represent a no state-specific transformation. One solution to introduce continuity in our algorithm is to create a transformation that has a state-depend component and a continuous component. For example, for the state n, $\theta_n(t) = (\theta_{(j,l,t)} + \theta_{(0,0,t)})/2$. This simple smoothing procedure is used in our algorithm, jointly with bias initialization.

Figures 2 and 3 represent the word accuracy on *Numbers* and *Phone Numbers* tasks, without initialization (*noInit*), with initialization (*Init*) and with smoothing and initialization (*Smooth_Init*).

For the *Numbers* task, the initialized bias outperforms those of non-initialized bias. For the *Phone Numbers* task, the difference between the initialized bias and non-initialized bias is small. For all tasks it is better to use smoothing that to not use them. More elaborated smoothing procedure will give probably higher improvement and will be studied in the future. Again, for *Numbers* and *Phone Numbers* tasks, the best performance is obtained for the tree depth of 1.

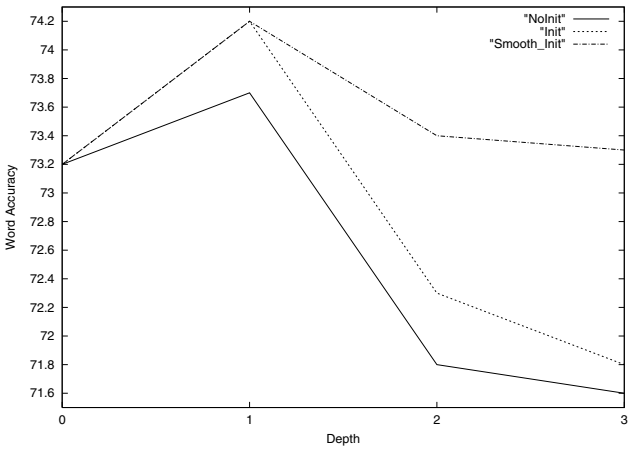

Figure 2: *Word Accuracy on Numbers task with noInit, Init and Smooth.*

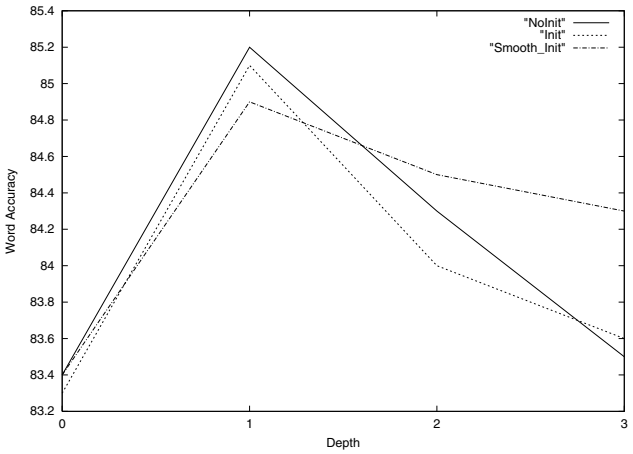

Figure 3: *Word Accuracy on Phone Numbers task with noInit, Init and Smooth.*

Compared to Table 2, bias initialisation and smoothing improve the results only for *Numbers* task (from 73.7 to 74.2). But, for *Phone Numbers* task initialisation and smoothing give no improvement. This can be explained as follow. Good initialization of the bias accelerate its convergence to a good estimate. This is important for short sentences such as in *Numbers* task (mean length of 100 frames). For the long sentences as in *Phone Numbers* (mean length of 600 frames), influence of initialization process on compensation is less noticeable, since convergence of bias is achieved before the end of the sentence wheter initialization of bias is performed or not.

Compared to CMN, SS and PMC approaches, structural state-based frame synchronous compensation, proposed in this paper, improves significantly the results for all used recognition tasks. Compared to non-structural approach developed in [6], the structural approach significantly improves the results for *Phone Numbers* task. This can be explained by the more statistics avaliable for the node-specific transformations estimation.

5. Conclusions

In this paper, we have presented a structural state-based frame-synchronous compensation algorithm aimed at dealing with time-varying unknown noise. The objective of the proposed hierarchical transformation is to better compensate for non-linear distortions of the feature space. The transformation tree is built on the states of acoustical models. During recognition, the node specific transformation is applied to the noise corrupted observation to compensate it. This approach outperform classical compensation methods on *Numbers* and *Phone Numbers* recognition in a moving car. The improvement is limited by a scarcity of data problem that occurs when depth of the tree is important. In this work, this problem was solved by the bias initialization and smoothing procedures which resulted in recognition accuracy improvement. Future work will involve more accurate smoothing procedure, a structural *MAP* approach and experimental validation on Aurora database.

6. References

[1] M.J.F. Gales. *Model-Based Techniques for Noise Robust Speech Recognition*. PhD 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*, 2(2):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*, 9:171–185, 1995.

[4] C. Mokbel, P. Paches-Leal, D. Jouvet, and J. Monn. Compensation of Telephone Line Effects for Robust Speech Recogni tion. In *Proceedings of International Conference on Spoken Language Processing*, April 1994.

[5] A. Sankar and C.H. Lee. A Maximum Likelihood Approach to Stochastic Matching for Robus t Speech Recognition. *IEEE Transaction on Speech and Audio Processing*, pages 190–202, 1996.

[6] V. Barreaud, I. Illina, and D. Fohr. On-Line Frame-Synchronous Compensation of Non-Stationary noise. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, April 2003.

[7] K. Shinoda and C.-H. Lee. A Structural Approach to Speaker Adaptation. *IEEE Transaction on Speech and Audio Processing*, 9(3):276–286, march 2001.

[8] Hui Jiang, Frank Soong, and Chin-Hui Lee. Hierarchical stochastic feature matching for robust speech reco gnition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2001.

[9] L. Delphin-Poulat, C. Mokbel, and J. Idier. Frame Synchronous Stochastic Matching Based on the Kullback-Le ibler Information. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 89–92, 1998.