Issues in acoustic modeling of speech for automatic speech recognition

Yifan Gong, Jean-Paul Haton, Jean-François Mari

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Issues in acoustic modeling of speech for automatic speech recognition

Yifan Gong, Jean-Paul Haton,

Jean-Francois Mari

N° 2368

Octobre 1994

PROGRAMME 3

ISSN 0249-6399

Issues in acoustic modeling of speech for automatic speech recognition

Yifan Gong, Jean-Paul Haton,
Jean-Francois Mari

Programme 3 — Intelligence artificielle, systèmes cognitifs et interaction homme-machine
Projet RF-IA

Rapport de recherche n° 2368 — Octobre 1994 — 15 pages

Abstract: Stochastic modeling is a flexible method for handling the large variability in speech for recognition applications. In contrast to dynamic time warping where heuristic training methods for estimating word templates are used, stochastic modeling allows a probabilistic and automatic training for estimating models. This paper deals with the improvement of stochastic techniques, especially for a better representation of time varying phenomena.

Key-words: Speech recognition, HMM, stochastic trajectory modeling

(Résumé : tsvp)

chapter in the book "Progress and Prospects of Speech Research and Technology", H. Nieman, R. De Mori and G. Hanrieder, editors, INFIX, Sankt Augustin, 1994

Unité de recherche INRIA Lorraine
Technôpole de Nancy-Brabois, Campus scientifique,
615 rue de Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY (France)
Téléphone : (33) 83 59 30 30 – Télécopie : (33) 83 27 83 19
Antenne de Metz, technopôle de Metz 2000, 4 rue Marconi, 55070 METZ
Téléphone : (33) 87 20 35 00 – Télécopie : (33) 87 76 39 77

Contribution à la modélisation acoustique en reconnaissance automatique de la parole

Résumé : La modélisation stochastique est une méthode souple pour tenir compte de la grande variabilité de la parole. Contrairement à la programmation dynamique qui utilise des méthodes heuristiques pour construire des formes de référence robustes, les modèles stochastiques permettent un apprentissage rigoureux reposant sur la théorie des probabilités. Ce rapport décrit des techniques stochastiques adaptées aux phénomènes transitoires propres à la parole. Il présente deux apports de l'équipe RF-IA au problème : les modèles de Markov du second-ordre et le modèle stochastique de trajectoire.

Mots-clé : Reconnaissance de la parole, Modèles de Markov, Modèle stochastique de trajectoire

1 Introduction

The design of an adequate modeling of speech patterns has been a constant concern since the beginning of automatic speech recognition research. The first techniques proposed relied on the use of acoustic, "vocoder-like" reference patterns together with dynamic time warping comparison [Sakoe 78]. In such template methods, the acoustic variability modeling of a vocabulary consisted in storing several references for the same lexical unit, or in deriving typical sequences of acoustic frames resorting to some kind of averaging method. These solutions were rather inefficient and expensive, even though they can provide a viable solution for a variety of applications.
The idea of the statistical modeling of spectral properties of speech gave a new dimension to the problem. The underlying assumption for all statistical methods is that speech can be adequately characterized as a random process whose parameters can be estimated in a proper way. The most widely used statistical method is the hidden Markov model (HMM) approach, first implemented for speech recognition during the seventies [Baker 75], [Jelinek 76]. The basic HMM model has led to very good performances in various domains of speech recognition. However, the intrinsic limitations of this model were progressively pointed out, as well as the necessity of incorporating into the model some knowledge about the speech communication process. Some solutions were proposed to overcome these limitations, especially in terms of frame correlation and trajectory modeling.
The use of artificial intelligence knowledge-based techniques was also proposed. Despite some success in phonetic decoding, these techniques suffer from several drawbacks, especially with respect to the lack of global criteria for parameter optimization and the severe difficulty of acoustic-phonetic knowledge elicitation. They could again be used in the future maybe in conjunction with other techniques once solutions are found for the preceding problems.
A great amount of effort has also been devoted to the development and the improvement of the HMM at several levels. These include enhancing the models themselves, and search techniques and finding methods for speaker representations [Levinson 86], [Schwartz 91], [Normandin 94], [Bahl 93].
This paper presents two models developed by our group in acoustic modeling for speech recognition, ie second-order HMM and stochastic trajectory models (STM). This paper is organized as follows. In section 2, we present a second-order Markov model and compare its performance with classical first-order models. We then propose in section 3 a new model referred to in the sequel as Stochastic Trajectory Model (STM) and highlight its interest for recognition. We conclude with a comparative study of different models.

2 Higher-order hidden Markov models

2.1 Increasing HMM order

HMM based speech modeling assumes that the input signal can be split into segments modeled as states of an underlying Markov chain and that the waveform of each segment is

a stationary random process. In a first-order hidden Markov Model (HMM1), the sequence of states is assumed to be a first-order Markov chain. This assumption is mainly motivated by the existence of efficient and tractable algorithms for model estimation and speech recognition. HMM1 does however suffer from several drawbacks. For instance, HMM1 assume segment frames to be independent, and does not include trajectory modeling (i.e. frame correlation) in the frame space. By incorporating short term dynamic features to model spectrum shape, HMM1 can be made to overcome this drawback. Modeling segment duration, which as a function of time, follows a geometric law, remains another major drawback for HMM1. In a second-order Markov model (HMM2), the underlying state sequence is a second-order Markov chain. The state duration in this model is governed by two parameters : the probability of entering a state only once, and the probability of visiting a state at least twice, with the latter modeled as a geometric decay.

Thus, HMM2 can explicitly model the event that a state can be visited just one time, and eliminate singular alignments given by the Viterbi algorithm in the recognition process when a state captures just one frame whereas all other speech frames fall into the neighboring states.

2.2 Second-order HMM

Unlike the first-order Markov chain where the stochastic process is specified by a 2-dimensional matrix of a priori transition probabilities a_{ij} between states s_i and s_j, the second-order Markov chain is specified by a 3 dimensional matrix a_{ijk}. Thus, in a second-order Markov chain, we have :

$$\begin{aligned} Prob(q_t = s_k / q_{t-1} = s_j, q_{t-2} = s_i, q_{t-3} = ...) = \\ Prob(q_t = s_k / q_{t-1} = s_j, q_{t-2} = s_i) = a_{ijk} \end{aligned} \tag{1}$$

with the constraints :

$$\sum_{k=1}^{N} a_{ijk} = 1 \quad with\ 1 \leq i \leq N\ ,\ 1 \leq j \leq N$$

The probability of the state sequence $Q \stackrel{\triangle}{=} q_1, q_2, ..., q_T$ is defined as :

$$Prob(Q) = \Pi_{q_1} a_{q_1 q_2} \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t}$$

where Π_i is the probability of state s_i at time t = 1 and a_{ij} is the probability of the transition $s_i \rightarrow s_j$ at time t = 2. Each state is associated with a mixture of Gaussian distributions :

$$b_i(O_t) \stackrel{\triangle}{=} \sum_{m=1}^{M} c_{im} \mathcal{N}(O_t; \mu_{im}, \Sigma_{im}), \quad with \quad \sum_{m=1}^{M} c_{im} = 1 \tag{2}$$

where O_t is the input vector (the frame) at time t. Given a sequence of observed vectors $O \stackrel{\triangle}{=} O_1, O_2, ..., O_T$ the joint state-output probability $Prob(Q, O/\lambda)$, is defined as :

$$Prob(Q, O/\lambda) =$$

$$\Pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \prod_{t=3}^{T} a_{q_{t-2} q_{t-1} q_t} b_{q_t}(O_t) \quad (3)$$

Basically, HMM2 suffers from two major theoretical drawbacks : i) it is computationnally less efficient than HMM1 since most of iterations run on the 2-fold product space **S x S**, and ii) each second-order Markov model has an equivalent first-order model on the 2-fold product space **S x S**. The first drawback is overcome by considering only the couples of states that are transitions in the model rather than considering the entire set **S x S**. In the model depicted in figure 1 this number is 2 times the number of states of the model. Point ii) is true when we have infinite data to train a model, since going back to first-order increases dramatically the number of states in the model. Figures 2 shows the equivalent model associated with the model depicted in figure 1. Moreover, there is a strong similarity between this process and the process of expanding a state of a conventional HMM and considering it as a *sub-HMM* , as in [Levinson 86], [Russell 87]. In the model depicted in figure 2, the duration in state j may be defined as :

$$\begin{aligned} d_j(0) &= 0 \\ d_j(1) &= a_{ijk}, \quad i \neq j \neq k \\ d_j(n) &= (1 - a_{ijk}) \cdot a_{jjj}^{n-2} \cdot (1 - a_{jjj}), \quad n \geq 2 \end{aligned}$$

It is interesting to note that HMM2 converges naturally to Ferguson-like models, [Mari 94] hence improving the capability of state duration modeling.

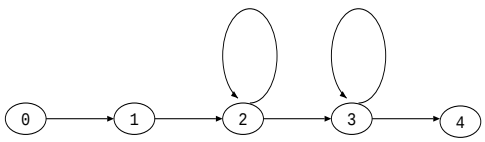

Figure 1: original second-order model

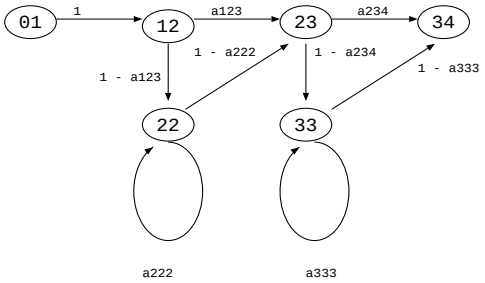

Figure 2: first-order equivalent model

2.3 Duration model

Even if the duration of a segment is better modeled by two parameters in a HMM2, thus avoiding singular state assignment as mentioned in 2.1, it is necessary to implement duration

constraints based on the relative duration of the segments corresponding to successive states as in [Suaudeau 93]. The reason is that most of the errors of our HMM2-based word recognition system come from singular alignments given by the Viterbi algorithm. We conducted an experiment on the training corpus to get statistics on the state duration of each HMM2. We observed that state durations were strongly correlated for states in a model. In order to take this correlation into account, we have specified a set of classes of correct alignments on one class per model basis. Given an utterance, an alignment between a model and the utterance is defined by a vector of relative duration of the states of the model. This alignment is found using the Viterbi algorithm. We denote by:

- $\mathbf{w}$, a d-frame long word which has been aligned with HMM λ. Each state i among the N states of λ captures d_i frames. If all states must be visited, we have: $d = d_1 + d_2 + ... + d_N$
- $$x \stackrel{\triangle}{=} (d, \frac{d}{d_2}, \frac{d}{d_3}, ..., \frac{d}{d_N}) \tag{4}$$
- g_λ, the mean vector associated with the class of λ, and V_λ the covariance matrix
- $detV_\lambda{}^{1/q}$, a normalizing factor that ensures that all matrices have a determinant equal to one.

Given a word model and the class of correct alignments of this model, we measure the distance between alignments using the Mahalanobis distance :

$$d^2(x, g_\lambda) \stackrel{\triangle}{=} detV_\lambda{}^{1/q}(\mathbf{x} - \mathbf{g}_\lambda)^t \mathbf{V}_\lambda{}^{-1}(\mathbf{x} - \mathbf{g}_\lambda) \tag{5}$$

This distance weights the probability of the Viterbi's alignment during a post-processing step where the N best [1] answers given by the recognition algorithm [Schwartz 91] are rescored.

$$FinalScore = A \cdot d^2(x, g_w) + B \cdot log(P(O/\lambda_w)) \tag{6}$$

A and B are normalizing constants determined empirically on the training set.

2.4 Test Protocol

First-order HMM and second-order HMM have been comparatively assessed using the same database of digits, i.e. the adult part of the TI-NIST database [Leonard 84]. This database contains connected digits strings from 225 adult speakers divided into a group of 112 to be used for training only and a group of 113 to be used for testing. Note that the sequence *"zero oh"* which causes most insertions errors is not present in this corpus. However our system accepts this sequence.
The vocabulary is made up of 23 models, one per digit and gender, and one for the background noise. The state output densities are mixtures of 9 Gaussian estimates with full

[1] N does not refer to the number of states of a model but rather to the number of alignments

covariance matrices. For the comparison, we have used models with the same topology and same number of pdfs. In particular, digit models have 6 states with 5 self loops and no skip transition, whereas the background noise model has only 2 states and one self loop. For computational convenience 2 extra states were added in HMM2 but no pdf was asssociated with them.

2.5 Parameterization

The speech signal in the TI-NIST database was recorded in a quiet laboratory environment and sampled at 20 kHz. Using a frame shift of 12 ms and a 25 ms window, we computed 12 cepstral coefficients corresponding to an approximate Mel-frequency warped spectrum. The first coefficient, called loudness, was removed. In some experiments we stack dynamic coefficients (usually called Δ, ΔE, $\Delta\Delta$ and $\Delta\Delta E$) over the 11 higher order static coefficients. Thus, each frame captures events in an overall window of 102 ms duration. Two analysis feature vectors incorporating dynamic features, have been specified in order to explore the capability of HMM2 to capture frame correlations :

- 24 coefficients : 11 static, 12 dynamic first-order coefficients plus the second-order energy coefficient $\Delta\Delta E$.
- 35 coefficients : 11 static, 12 dynamic first-order coefficients plus 12 dynamic second-order coefficients.

2.6 HMM1/HMM2 comparison

Tables 1 and 2 summarize the recognition results. In these tables, we give the string error rates and the 95% confidence intervals. Table 3 gives the results at the word level. In the different experiments, we used the 8700 strings from the test part of the TI-NIST database containing 28383 digits.

Three major conclusions can be drawn from these results:

1. HMM2 outperforms HMM1 in the absence of post-processing, and HMM2 without post-processing is almost equivalent in performances to HMM1 with post-processing (see tables 1 and 2).
2. Acceleration coefficients do not significantly improve performance, especially with HMM2 (see table 1).
3. The offset in performances is greatly reduced when a post-processor is used to take into account the duration constraints.

Point 1 can be explained by the capability of HMM2 to model the probability that the hidden Markov process stays only one time in specific states. Thus, the trajectory of speech, in terms of state sequence, is better modeled by HMM2.
Since the beginning of this study in 1990, several sytems have produced better performances

Parameterization	Male + Female	
	HMM1	HMM2
$11MFCC + 11\Delta +$ $\Delta E + \Delta\Delta$ E	4.5% (4.1 5.0)	2.4% (2.1 2.7)
$11MFCC + 11\Delta +$ $\Delta E + \Delta\Delta E + 11\Delta\Delta$	3.7% (3.3 4.1)	2.4% (2.1 2.7)

Table 1: String error rates (without post-processing)

Parameterization	Male + Female	
	HMM1	HMM2
$11MFCC + 11\Delta +$ $\Delta E + \Delta\Delta$ E	2.8% (2.5 3.2)	2.2% (1.9 2.5)
$11MFCC + 11\Delta +$ $\Delta E + \Delta\Delta E + 11\Delta\Delta$	2.3% (2.0 2.6)	2.1% (1.8 2.4)

Table 2: String error rates (with post-processing)

	HMM1	HMM2
Insertions	174	159
Deletions	14	20
Substitutions	31	34
String error rate	2.3 %	2.4 %
% correct	99.8	99.8
Accuracy	99.2	99.2

Table 3: Comparison between HMM1 (with post-processing) and HMM2 (without post-processing)

on the TI-NIST corpus [Haeb-Umbach 93], [Cardin 93]. These systems involve sophisticated parameterization and training techniques. Our word recognition system, based on HMM1 models, which serves as the reference system to which the HMM2-based system was compared, gives results similar to the system described by Wilpon in 1993 [Wilpon 93], ie 2.4 % string error rate with a 10 state model with 9 Gaussian pdf per mixture and telephone bandwith speech. In our system, we have 6 states per model (no frame is consumed in extra states) but 2 models per digit. This keeps the number of parameters slightly constant.
Point 2 has already been mentioned in relation to clean speech and HMM1 models [Hanson 90]. The analysis of errors in HHM1 and HMM2 show that most are insertion errors. 60 % insertion errors are due to the insertion of *"oh"* after a *"zero"* and 20 % are insertion of *"eight"* between a word and a silence. We guess that such errors could be avoided with an appro-

priate energy modeling at the state level. Almost all the deletion errors involve *"oh"*. The small number of confusion errors does not allow any conclusion regarding the discriminative power of both methods.

3 Stochastic trajectory modeling

3.1 Motivation

In a parametric space (e.g. cepstral space), a speech signal can be represented as a point which moves as articulatory configuration changes during continuous speech production. The sequence of moving points is called the trajectory of speech. Since a given point can belong to different trajectories, models for speech recognition should rely more on the trajectory of speech rather than on the geometrical position of observations in the parameter space. As already mentioned in section-2, the inherent state independency assumption in basic hidden Markov models cannot preserve trajectory information. Particularly, the pdf of different groups of trajectories are mixed up and clusters of trajectories cannot be well represented, because the information on the continuity of each individual trajectory is lost. Trajectories are *folded*, leading to a poor discriminability in complex phonetic contexts (cf. Figure-3).

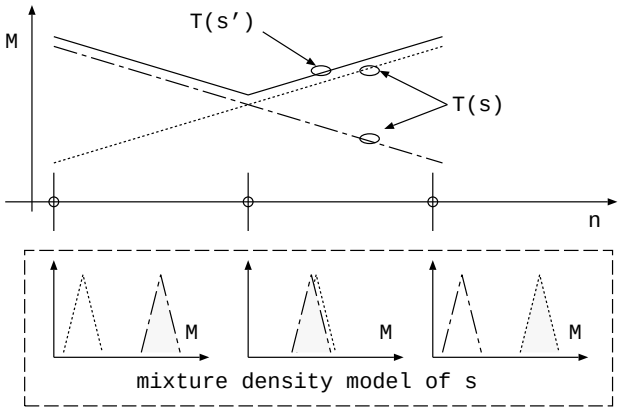

Figure 3: Illustration of trajectory folding with a three state HMM phoneme model. The trajectories of s, $T(s)$, in different phonetic contexts are modeled by mixture probability densities. For the given trajectory of $s' \neq s$, $T(s')$, never appeared in training data, $p(T(s')|s)$ will be as high as $p(T(s)|s)$.

Explicit statistical dependency between the current observation vector and the last observed vector can be modeled by Gaussian estimates as in [Wellekens 87]. The correlation between observation vectors can also be modeled by a bigram constraint [Paliwal 93]. The state observation probability can be conditioned on the previous observation vectors, as well

as on the current vector [Takahashi 93]. The sequential constraints can also be captured by recurrent neural networks [Robinson 92]. The previously proposed stochastic segment models handle segmental information by modeling the pdf of a grand vector, made up of the concatenation of resampled frames of the observation vector sequence, by a multi-variate Gaussian distribution [Ostendorf 89]. However, no mixture probability density notion is used, that implies the impossibility of modeling trajectory clusters.

We consider trajectories as observations of a random variable and propose a stochastic mixture trajectory model (STM) [Gong 94] for its recognition. We model phoneme-based speech units as clusters of trajectories in their parameter space. The trajectories are modeled by mixture of state sequences of multi-variate Gaussian density functions, optimized at the state sequence level. Duration of trajectories are integrated in the model.

3.2 Principle

3.2.1 Phoneme probability

Let us consider a phone segment parameterized as a sequence of N vectors : $\mathbf{o}_0, \mathbf{o}_1, ...\mathbf{o}_n, ..., \mathbf{o}_N$. Each point $\mathbf{o}_n \in I\!R^D$ is a D-dimensional vector in some parameter space. Let $\mathbf{X}_n$ be a sequence of Q vectors centered at time slot n. The Q vectors are linearly mapped from the $\mathbf{o}$ sequence.

$$\mathbf{X}_n \stackrel{\triangle}{=} \mathbf{x}_{n-\frac{Q}{2}}, \mathbf{x}_{n-\frac{Q}{2}+1}, ...\mathbf{x}_n, ...\mathbf{x}_{n-\frac{Q}{2}+Q-1} \tag{7}$$

It has been observed that non-linear mapping [Afify 94] results in a slight recognition improvement, but it introduces additional computational cost. Let $I\!P \stackrel{\triangle}{=} \{s_1, s_2, ..., s_H\}$ be a set of H symbols representing phonemes. In our formulation, it is assumed that each phoneme symbol is associated with K stochastic trajectory generators, $T_1, T_2, ..., T_K$. Let $p(\mathbf{X}_n, T_k, d, s)$ be the joint probability density function (pdf) of vector sequence $\mathbf{X}_n$, component trajectory source T_k, duration d and phoneme symbol $s \in I\!P$:

$$p(\mathbf{X}_n, T_k, d, s) = p(\mathbf{X}_n|T_k, d, s)Pr(T_k|d, s)Pr(d|s)Pr(s) \tag{8}$$

where $p(\mathbf{X}_n|T_k, d, s)$ is the pdf of $\mathbf{X}_n$ given T_k, d and s, $Pr(T_k|d, s)$ the probability of T_k given d, and s, $Pr(d|s)$ the probability of d given s, and $Pr(s)$ the *a priori* probability of s.

We use p for continuous probability density functions and Pr for discrete probabilities. The marginal pdf of $p(\mathbf{X}_n, s)$ can be obtained by summing up $p(\mathbf{X}_n, T_k, d, s)$ over all trajectories T_k and all durations d:

$$p(\mathbf{X}_n, s) = Pr(s) \sum_d \sum_{k=1}^{K} p(\mathbf{X}_n|T_k, d, s)Pr(T_k|d, s)Pr(d|s) \tag{9}$$

The probability of phoneme s given the observation $\mathbf{X}_n$ is therefore:

$$Pr(s|\mathbf{X}_n) = \frac{p(\mathbf{X}_n, s)}{p(\mathbf{X}_n)} = \frac{Pr(s)}{p(\mathbf{X}_n)} \sum_d p(\mathbf{X}_n|d, s)Pr(d|s) \tag{10}$$

where $p(\mathbf{X}_n|d,s)$ is the pdf of $\mathbf{X}_n$ given d and s:

$$p(\mathbf{X}_n|d,s) \stackrel{\triangle}{=} \sum_{k=1}^{K} p(\mathbf{X}_n|T_k,d,s)Pr(T_k|d,s) = \sum_{k=1}^{K} p(\mathbf{X}_n|T_k,d,s)Pr(T_k|s) \tag{11}$$

where we assume that the probability of a trajectory T_k does not depend on durations, i.e. for a given symbol different durations are tied together:

$$Pr(T_k|d,s) = Pr(T_k|s) \tag{12}$$

The duration probability of each phoneme symbol $Pr(d|s)$ is modeled by Γ-distributions.

3.2.2 Component trajectory

The critical part in our formulation is the modeling of $p(\mathbf{X}_n|T_k,d,s)$ introduced in Eq-11. T_k is a component trajectory in the mixture of pdfs of K trajectory generators.

Assuming that each of the Q points of the component trajectory T_k is produced by an independent distribution, the pdf of $\mathbf{X}_n$, given T_k, d and s is modeled as:

$$p(\mathbf{X}_n|T_k,d,s) \stackrel{\triangle}{=} \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{x}_{n-\frac{d}{2}+i\frac{d}{q}};\mathbf{m}^s_{k,i},\mathbf{\Sigma}^s_{k,i})^{w_i} \tag{13}$$

where $\mathcal{N}(\mathbf{x};\mathbf{m},\mathbf{\Sigma})$ is a Gaussian distribution with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$. w_i weights the pdf of each state on the trajectory to obtain larger contribution from the center part.

3.2.3 Sentence recognition

Sentence recognition consists in evaluating a cumulated log-probability measure over all possible sequences of phonemes and finding the most plausible sequences.

Let $I\!F$ be the set of all grammatical sentences. A particular sentence $\omega \in I\!F$ is made up of $L(\omega)$ symbols:

$$\omega \stackrel{\triangle}{=} a_0, a_1, ...a_h, ...a_{L(\omega)-1}, \ \forall h,\ a_h \in I\!P.$$

From Eq-10. the log-probability of symbol a_h at time slot i is available:

$$\mu_{n,a_h} \stackrel{\triangle}{=} \log Pr(a_h|\mathbf{X}_n), \ \ 0 \leq n < N, 0 \leq h < L(\omega).$$

The duration of a symbol is considered as a random variable τ, and the probability of symbol a_h with duration $\tau = d$ is $Pr(d|a_h)$ introduced in section-3.2.1. Let t_h be a time slot index of the vector sequence of a_h. We introduce the cumulated log-probability for a_h, which is the sum of plausibilities cumulated for the symbol a_h from t_h to $t_{h+1}-1$, weighted by the γ^{th} power of the corresponding duration probability:

$$q(h) = Pr(t_{h+1}-t_h|a_h)^{\gamma} \sum_{t_h \leq n < t_{h+1}} \mu_{n,a_h} \tag{14}$$

where $\gamma \geq 0$ is constant during recognition. The log-probability of a sentence is defined as the normalized non-overlapping sum of cumulated log-probability of its composing symbols:

$$\theta(\omega|t_0, t_1, ...t_{L(\omega)}) = \frac{1}{N} \sum_{0 \leq h < L(\omega)} q(h) \tag{15}$$

where $t_0 = 0, t_h < t_{h+1}$, and $t_{L(\omega)} = N - 1$. This log-probability is therefore a function of $t_h, \forall h \in [0, L(\omega)]$. We optimize t_h's so that $\theta(\omega)$ is maximized:

$$\Theta(\omega) = \max_{t_0, t_1, ...t_{L(\omega)}} \theta(\omega|t_0, t_1, ...t_{L(\omega)}) \tag{16}$$

The $t_h \; \forall h$ which maximize Eq-16 are the starting time slots of the symbol $a_h's$. A backtracking can be applied to obtain $t_h \; \forall h$ if necessary.

Sentence recognition consists in evaluating $\Theta(\omega)$ for all possible sentences, and in assigning the most probable sentence as the recognized sentence $\hat{\omega}$:

$$\hat{\omega} = \underset{\omega \in I\!F}{\operatorname{argmax}} \, \Theta(\omega) \tag{17}$$

To evaluate $\Theta(\omega)$, we introduce the following auxiliary function of l (frame slot) and j (phoneme order) [Gong 94]:

$$\Pi(l, j) = \max_{0 \leq k < l} \{\Pi(k, j-1) + Pr(l-k|a_j)^\gamma \sum_{k \leq n < l} \mu_{n,a_j} \;\; 0 \leq l \leq N, 0 \leq j < L\} \tag{18}$$

We have

$$\Theta(\omega) = \frac{1}{N} \Pi(N, L(\omega) - 1)).$$

3.3 Experimental Results

We have tested the STM model for French using a 1010 word vocabulary grammar with a word-pair perplexity of 26. The same grammar was used for two tasks : Newspaper real-estate ads dictation (*"real-estate"*) and working report dictation for nuclear power plant inspection (*"working report"*). 33 context-independent phoneme models were used for all tests. 13 mel-cepstral coefficients were computed. Previous experiments showed that time derivatives did not improve recognition performance. Each model has 5 states with up to 8 components in a mixture. Table-refTR summerizes the results for two male speakers.

3.4 Comparison with HMM

The basic idea of HMM consists in modeling speech variabilities. HMMs use a sequence of states to capture speech variability. Typically, each state is associated with a mixture of Gaussian distributions. On the other hand, STM is designed to avoid the trajectory folding phenomenon, and thereby to improve the ability to deal with complex phonetic contexts.

	real estate	working report
speaker	lar	yfg
training style	vocabulary-dependent	vocabulary-independent
training speech (sentences/minutes)	140/4	80/3
number of pdfs	700	655
number of training phone tokens	2773	2353
microphone	sun-desk	shure M10
signal to noise ratio	15dB	40dB
test speech (sentences/words)	241/1482	161/685
word recognition rate	98%	96%

Table 4: Recognition accuracy on two tasks by stochastic trajectory models

STM uses a mixture of sequences of states. Each state has one Gaussian distribution. While the number of parameters in the two schemes are basically identical, there is a fundamental difference between the two: in STM, the mixture of densities is defined on the state sequence whereas in HMM it is defined on individual states. In addition, STM exploits an accurate explicit phone duration probability modeling in phoneme recognition. It takes into account the fact that the center of a segment has a smaller variance than its extremities by weigthing state observation probabilities.

Based on context-independent phoneme models, STM gives equivalent recognition accuracy than context-dependent HMM on similar tasks (i.e.: ARPA RM task), with much less training data.

4 Conclusion

We have presented in this paper two contributions of our group to the difficult problem of acoustic modeling for automatic speech recognition.
These two contributions address the problem differently. The first one consists in modeling speech by second-order Markov models instead of the usual first-order model. Experiments with the TI-NIST database show that HMM2 outperform basic HMM1, and give comparable results to HMM1 plus post-processing for duration.
The second approach deals with the explicit modeling of speech trajectories in some parametric space. We propose a Stochastic Trajectory Model based on a stochastic mixture representation. Results obtained in continuous speech recogniton for French with 3-4 minutes of training speech prove the relevance of STM for modeling contextual variations of phones.

References

[Afify 94] M. Afify, Y. Gong and J.-P. Haton. Non-linear time alignment in stochastic trajectory models for speech recognition. In *Proc. of Int. Conf. on Spoken Language Processing 1994*, Yokohama, Japan, September 1994.

[Bahl 93] L. R. Bahl, J. R. Bellarda, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo and M. A. Picheny. Multonic Markov Word Models for Large Vocabulary Continuous Speech Recognition . *IEEE Transactions on Speech and Audio Processing*, 1(3):334–344, July 1993.

[Baker 75] J.K. Baker. The Dragon system- An overview. *IEEE Trans. on ASSP*, 23(11):24 – 29, 1975.

[Cardin 93] R. Cardin, Y. Normandin and E. Millien. Inter-Word Coarticulation Modeling and MMIE Training for Improved Connected Digit Recognition. In *Proc. ICASSP*, volume 2, pages 243 – 246, 1993.

[Gong 94] Y. Gong and J.-P. Haton. Stochastic trajectory modeling for speech recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume I, pages 57–60, Adelaide, Australia, April 1994.

[Haeb-Umbach 93] R. Haeb-Umbach, D. Geller and H. Ney. Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities. In *Proc. ICASSP*, pages 239 – 242, 1993.

[Hanson 90] B. A. Hanson and T. Applebaum. Robust Speaker-Independent Word Recognition Using Static, Dynamic, and Acceleration Features: Experiments with Lombard and Noisy Speech. In *Proc. ICASSP*, pages 857 – 860, 1990.

[Jelinek 76] F. Jelinek. Continuous Speech Recognition by Statistical Methods. *IEEE Trans. on ASSP*, 64(4):532 – 556, April 1976.

[Leonard 84] R. G. Leonard. A Database for Speaker Independent Digit Recognition. In *Proc. ICASSP*, pages 42.11.1 – 42.11.4, San Diego, CA, March 1984.

[Levinson 86] S. E. Levinson. Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition. *Computer Speech and Language*, 1:29 – 45, 1986.

[Mari 94] J.-F. Mari and J.-P. Haton. Automatic Word Recognition Based On Second-Order Hidden Markov Models . In *Proc. ICSLP*, page S07.17, Yokohama, September 1994.

[Normandin 94] Y. Normandin, R. Cardinand and R. De Mori. High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation . *IEEE Transactions on Speech and Audio Processing*, 2(2):299–311, April 1994.

[Ostendorf 89] M. Ostendorf and S. Roucos. A stochastic segment model for phoneme-based continuous speech recognition. *ASSP*, 37(12):1857–1869, 1989.

[Paliwal 93] K. K. Paliwal. Use of temporal correlation between successive frames in a hidden Markov model based speech recognizer. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume II, pages 215–218, 1993.

[Robinson 92] T. Robinson. A real-time recurrent error propagation network word recognition system. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume I, pages 617–620, 1992.

[Russell 87] M. J. Russell and A. Cook. Experimental Evaluation of Duration Modelling Techniques For Automatic Speech Recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2376–2379, Dallas, 1887.

[Sakoe 78] H. Sakoe and S. Chiba. Dynamic Programming Optimization for Spoken Word Recognition, . *IEEE Trans. on ASSP*, 26(11):43 – 49, 1978.

[Schwartz 91] R. Schwartz and S. Austin. A Comparison of Several Approximate Algorithms for Finding Multiple (N-BEST) Sentence Hypotheses. In *Proc. ICASSP*, pages 701 – 704, 1991.

[Suaudeau 93] N. Suaudeau and R. André-Obrecht. Sound Duration Modelling and time variable Speaking rate in a Speech Recognition System. In *Eurospeech*, pages 307 – 310, 1993.

[Takahashi 93] S. Takahashi, T. Matsuoka, Y. Minami and K. Shikano. Phoneme HMMS constrained by frame correlations. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume II, pages 219–222, 1993.

[Wellekens 87] Wellekens. Explicite time correlation in hidden Markov models for speech recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 384–386, Dallas, 1987.

[Wilpon 93] J. G. Wilpon, C.-H. Lee and L. R. Rabiner. Connected Digit Recognition Based on Improved Acoustic Resolution. *Computer Speech and Language*, 7:15–26, 1993.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
ISSN 0249-6399