VITERBI ALGORITHM FOR ACOUSTIC VECTORS GENERATED BY A LINEAR STOCHASTIC DIFFERENTIAL EQUATION ON EACH STATE

Marco SAERENS
IRIDIA Laboratory, Université Libre de Bruxelles, cp. 194/6,
50 av. F. Roosevelt, 1050 Bruxelles, BELGIUM
Email: saerens@ulb.ac.be

ABSTRACT

When using hidden Markov models for speech recognition, it is usually assumed that the probability that a particular acoustic vector is emitted at a given time only depends on the current state and the current acoustic vector observed. In this paper, we introduce another idea, i.e. we assume that, in a given state, the acoustic vectors are generated by an linear stochastic differential equation. This extends our previous model, in which we assumed that the acoustic vectors are generated by a continuous Markov process. This work is motivated by the fact that the time evolution of the acoustic vector is inherently dynamic and continuous, so that the modelling could be performed in the continuous-time domain instead of the discrete-time domain. By the way, the links between the discrete-time model obtained after sampling, and the original continuous-time signal are not so trivial. In particular, the relationship between the coefficients of a continuous-time linear process and the coefficients of the discrete-time linear process obtained after sampling is nonlinear. We assign a probability density to the continuous-time trajectory of the acoustic vector inside the state, reflecting the probability that this particular path has been generated by the stochastic differential equation associated with this state. This allows us to compute the likelihood of the uttered word. Reestimation formulae for the parameters of the process, based on the maximization of the likelihood, can be derived for the Viterbi algorithm [34]. As usual, the segmentation can be obtained by sampling the continuous process, and by applying dynamic programming to find the best path over all the possible sequences of states.

1. INTRODUCTION

Hidden Markov models (HMM) are widely used for speech recognition ([20], [26], [32], [16]). However, strong assumptions have to be made to render the model computationally tractable (see, for instance, [2]). One of these assumptions is the observation independence of the acoustic vectors. Indeed, it is usually assumed that the probability that a particular acoustic vector is emitted at a given time only depends on the current state and the current acoustic vector observed. This does not take account of the time dynamic behaviour of the acoustic vector inside a state. This problem of time dynamic modelization has become an important research topic in speech recognition. For instance, Furui ([11], [12]) and Gurgen, Sagayama & Furui [15] introduce features including the time-derivative of the acoustic vectors. Deng ([5], [6]) modelizes the temporal evolution of the acoustic feature inside a state by a given function of time, i.e. a polynomial trend function of time t spend in the state. Wellekens [41] assumes explicit dependence between the current vector and the last observed vector. He shows that, in the case of a correlated Gaussian probability distribution function, the emission probabilities depend on the prediction error of a first order linear predictor. On the other hand, Poritz [31], Juang [21], Juang & Rabiner [22] (see also [23], [39], [43]) use Gaussian autoregressive densities per state, assuming that the acoustic vectors are generated by linear autoregressive processes, corrupted by Gaussian additive noise. Once more, the emission probabilities depend on the prediction error of a linear predictor (given by the autoregressive model). An extension of the Gaussian autoregressive densities model, in which we allow the autoregressive coefficients to be stochastic variables, is presented in [35].

More recently, some authors have considered the possibility of using non-linear prediction models (mostly multi-layer neural networks) for speech recognition with hidden Markov models ([40], [24], [25], [37], [38], [30], [17], [18], [4]). In this case, the acoustic vectors are assumed to be generated at each frame by a discrete nonlinear process, different for every state, corrupted by an additive uncorrelated Gaussian noise. It generalizes the work mentioned above, where linear prediction models were considered. Another interesting work, relying on a dynamical system approach with parameter training based on the EM algorithm, can be found in ([7], [8]).

Our work can be considered as a continuous-time version of the linear autoregressive modelling just mentioned. It tries to address the following question: Isn't it possible to build a continuous-time formulation of hidden Markov modelling for speech recognition, following the fact that speech is continuous and dynamic by nature. In other words, what is the continuous-time counterpart of the linear discrete-time modelling. In most domains dealing with continuous-time physical systems, such as control systems theory, the analysis and the modelling can be performed in the continuous-time domain, so that sampling only occurs afterwards for the purpose of digital computation. We have to keep in mind that the relationship between a continuous-time linear system and the discrete-time linear system obtained after sampling of the continuous-time signal is non so trivial.

More precisely, in this paper, we assume that, in a given state, the acoustic vectors are generated by an ordinary first-order linear stochastic differential equation. This extends previous work, in which we assumed that the acoustic vectors are generated by a continuous Markov process [34]. In other words, we assign a probability density to the continuous path of the acoustic vector inside the state, reflecting the probability that this particular path has been generated by the continuous stochastic process associated with the state. This computation relies on the concept of path integral – also known as Wiener integral [42] – widely used in theoretical physics (see, for instance, [9], [10], [14], [36]).

As usual, the sequence of states is assumed to follow a first order discrete Markov process. However, in our model, the state transitions do not occur at regular time intervals, so that it should be more appropriate to speak about semi-Markov process [3]. The probability of a succession of states and the observed time evolution of the acoustic vector can be computed as the product of transition probabilities between the states and path probabilities inside the states. This leads to the computation of the likelihood of the uttered word.

This approach leads to the introduction of a time-derivative which is to be added to the acoustic vector, and is therefore related to Furui ([11], [12]) and Gurgen, Sagayama & Furui's work [15], which also consider the time-derivative of the cepstral vector as a feature. Once the segmentation is fixed, reestimation formulae for the parameters of the process (based on the maximization of the likelihood) can be derived for the Viterbi algorithm in the same way as in [34]. The segmentation can be obtained by sampling the continuous process, and by applying dynamic programming to find the best path over all the possible sequences of states. This computation follows the same line as in [34].

2. MOTIVATIONS: WHY CONTINUOUS-TIME MODELLING ?

Continuous-time modelling could appear mere as an *exercice de style*, so that we will give some motivations to follow the developments.

(i) First of all, while in speech processing sampling is often assumed a piori, the speech production system is continuous by nature. It implies that the time evolution of the acoustic vector is inherently dynamic and continuous, so that the analysis and the modelling could be performed in the continuous-time domain.

Now, let us examine what happens when the dynamic of the speech process is modelled after sampling by a discrete-time linear model (see, for instance, [1]). In this case, the speech samples are supposed to follow an autoregressive process corrupted by additive Gaussian white noise in each state of the Markov model. In state-space configuration, this means that the state vector $\mathbf{x}_k$ at time step k is following

$$\mathbf{x}_{k+1} = \mathbf{B}\,\mathbf{x}_k \tag{1}$$

However, let us assume that the time evolution of the state vector is in fact following a continuous-time linear differential equation:

$$\frac{d\mathbf{x}(t)}{dt} = \dot{\mathbf{x}} = \mathbf{A}\,\mathbf{x} \tag{2}$$

The solution of this differential equation is

$$\mathbf{x}(t) = \exp[\mathbf{A}\,(t-t_0)]\,\mathbf{x}(t_0) \tag{3}$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

where $\mathbf{x}(t_0)$ is the initial value at time t_0, and exp[A t] is defined as

$$\exp[\mathbf{A}\, t] = \sum_{n=0}^{\infty} \mathbf{A}^n \frac{t^n}{n!} \tag{4}$$

Now, by sampling this continuous-time process with a sampling period Δt, we obtain the discrete-time process:

$$\mathbf{x}_{k+1} = \exp[\mathbf{A}\, \Delta t]\, \mathbf{x}_k \tag{5}$$

where we used the following notation: $\mathbf{x}_k = \mathbf{x}(k\, \Delta t)$. The matrix **B** defining the autoregressive process (1), obtained after sampling, is therefore given by

$$\mathbf{B} = \exp[\mathbf{A}\, \Delta t] \tag{6}$$

This shows that there is no trivial correspondance between the coefficients of the continuous-time differential equation and the coefficients of the difference equation obtained after sampling. In addition, the correspondance is nonlinear. While certain dimensions can be uncorrelated in the original process, the sampling action can introduce correlations. Despite the fact that both the discrete and the continuous model are linear, and therefore do essentially the same think, the model could (or not; only experiments could decide) be simpler and more easily identified in the continuous-time domain.

Moreover, if we change the sampling period from Δt to $\Delta t'$, the resulting discrete-time process will be:

$$\mathbf{x}_{k+1} = \exp[\mathbf{A}\, \Delta t']\, \mathbf{x}_k \tag{7}$$

and, once more, there is no trivial correspondance between the coefficients of the two discrete-time processes (5) and (7), having a different sampling frequency. This means that the modelling by a discrete-time process at a given sampling frequency is only valid at that frequency. This fact makes the continuous-time formulation more attractive.

In addition, the entire left half plane of the p-plane (Laplace transform) is mapped into the unit circle of the z-plane (z transform). The correspondence between the zeroes of the continuous-time polynomial p_i and the zeroes of the discrete-time process q_i obtained after sampling is given by $q_i = \exp[p_i\, \Delta t]$. As $\Delta t \rightarrow 0$, the zeroes in the p-plane are mapped into a close cluster near z=1 in the z-plane. This leads to numerical ill-conditioning: for instance, we have $dq_i/dp_i = \Delta t \exp[p_i\, \Delta t] \rightarrow 0$ with $\Delta t \rightarrow 0$, so that the q_i become increasingly insensitive to p_i. We therefore need higher precision for the coefficients of the discrete-time polynomial when $\Delta t \rightarrow 0$ (this has been observed in control theory; see [27], [33]).

Now, let us briefly resume the other reasons that lead us to the introduction of the continuous-time model that will be described in this paper, while these reasons will only become clear later. Some of these reasons are not specific to continuous-time models.

(ii) The model takes account of the dynamic of the acoustic vectors since the mean value of the acoustic vector is supposed to follow a temporal trajectory. Indeed, we will show later that the acoustic vector is following the solution of a p-order vectorial linear differential equation in average. The dynamics of the vector is therefore represented by a vectorial linear differential equation subject to random fluctuations, for each state.

(iii) It models the uncertainty related to the observed process. For times near the initial condition, the acoustic vector is expected to be near the observed value, while, as time goes on, the position is becoming more and more fuzzy. For instance, this gives us the possibility to consider processes for which the observations are generated at a nonconstant time period (while, of course, this kind of process is extremely rare).

(iv) The model is just an extension of the one-Gaussian-per-state continuous hidden Markov models, by allowing the mean value and the variance-covariance matrix of the Gaussian densities to evolve over time.

3. General overview of the model

The model is formulated in the framework of word recognition with Viterbi algorithm. In the future, both for parameter estimation and for segmentation, we will be interested in the evaluation of the total probability (the likelihood) of the observations P(X, S), where S is a sequence of states (s_0, s_1, ..., s_Q) defining a word together with its segmentation, and X is the time evolution of the acoustic vector $\mathbf{x}(t)$, observed over the whole word. The estimation of the parameters will be based on the maximization of this likelihood, while the state segmentation will be chosen so as to maximize the a posteriori probability P(S | X). We have

$$\mathrm{P(X, S)} = \mathrm{P(X \mid S)\ P(S)} \tag{8}$$

Now, the sequence of states is assumed to be modelled by a discrete Markov process, and the time evolutions of the observations arising from any state are assumed to be independent, so that we can write

$$\mathrm{P(S)} = [\prod_{k=1}^{Q} \pi(s_k \mid s_{k-1})]\, \pi_0 \tag{9}$$

$$\mathrm{P(X \mid S)} = \prod_{k=0}^{Q} \mathscr{P}_{s_k}[\mathbf{x}(t)] \tag{10}$$

where $\pi(s_k \mid s_{k-1})$ is the transition probability of the discrete Markov model of states, and $\mathscr{P}_s[\mathbf{x}(t)]$ is the probability density of the observed continuous acoustic vector trajectory $\mathbf{x}(t)$ on state s.

The problem is, of course, to compute the probability density of a path $\mathscr{P}_s[\mathbf{x}(t)]$ on a state s. In the following, we will suppose that the acoustic vectors $\mathbf{x}(t)$ are generated by a stochastic differential equation in each state, and, from this assumption, we will be able to compute the probability density of the observed path $\mathbf{x}(t)$.

4. Markov process generated by an ordinary stochastic differential equation

Most of the material in this paragraph is taken from Gardiner [13]; therefore, see this monograph for more details. Let us assume that, in each state $s \in \{0, 1, ..., q\}$, the acoustic vector $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_d(t)]^{\mathrm{t}}$ at time t is generated by a stochastic differential equation:

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \dot{\mathbf{x}} = \mathbf{A}_s\,(\mathbf{x} - \mathbf{x}_s) + \mathbf{B}_s\, \boldsymbol{\xi}(t) \tag{11}$$

where $\mathbf{A}_s$ and $\mathbf{B}_s$ are d-dimensional square matrices of real values, and the vector $\boldsymbol{\xi}(t)$ is a rapidly fluctuating random term, simulating a noise source. An idealised mathematical formulation of the concept of a "rapidly fluctuating random term" is that for $t \neq t'$, $\boldsymbol{\xi}(t)$ and $\boldsymbol{\xi}(t')$ are statistically independent. We also require that the mean value is zero, i.e. $<\boldsymbol{\xi}(t)> = 0$. We thus have for the covariances $<[\boldsymbol{\xi}]_i(t)\ [\boldsymbol{\xi}]_j(t')> = \delta_{ij}\, \delta(t-t')$ where $[\boldsymbol{\xi}]_i$ is coordinate i of vector $\boldsymbol{\xi}$ (this correponds to white noise).

Now, it is known from the theory of stochastic processes that if the variable defined as

$$\mathbf{u}_s(t) = \int_{t_0}^{t} \boldsymbol{\xi}(t')\ \mathrm{d}t' \tag{12}$$

is a continuous function of t, the process described by (11) is the so-called multivariate Ornstein-Uhlenbeck process. As seen from (11) and (12), it corresponds to a linear process undergoing uncorrelated fluctuations so that the overall observed trajectory is continuous. The "noise" $\boldsymbol{\xi}$ is responsible for the stochastic nature of the process, so that the position of $\mathbf{x}$ is not deterministic but is characterised by a probability density function, reflecting the probability of observing this position. Ornstein-Uhlenbeck process occurs, for instance, for the velocity of a particle in Brownian motion.

Expressions (11) and (12) define a continuous Markov process in the sense that the conditional probability density of finding a particular value of the acoustic vector $\mathbf{x}$ at time t is determined entirely by the knowledge of the most recent condition, that is, by the most recent observation. This means that if $p(\mathbf{x}, t \mid \mathbf{x}_0, t_0, \mathbf{x}_1, t_1)$ with $t_0 < t_1 < t$ represents the conditional probability of observing $\mathbf{x}$ at time t, given that we observed $\mathbf{x}_0$ at t_0 and $\mathbf{x}_1$ at t_1, $p(\mathbf{x}, t \mid \mathbf{x}_0, t_0, \mathbf{x}_1, t_1)$ is simply equal to $p(\mathbf{x}, t \mid \mathbf{x}_1, t_1)$.

For a deterministic initial condition $p_s(\mathbf{x}, t_0) = \delta(\mathbf{x} - \mathbf{x}_0)$, the corresponding conditional probability density $p_s(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$ is Gaussian with mean

$$<\mathbf{x}(t)> = \mathbf{m}_s(t) = \exp[\mathbf{A}_s\,(t - t_0)]\,(\mathbf{x}(t_0) - \mathbf{x}_s) + \mathbf{x}_s \tag{13a}$$

and variance

$$\begin{aligned} <[\mathbf{x}(t) - <\mathbf{x}(t)>]\ [\mathbf{x}(t) - <\mathbf{x}(t)>]^{\mathrm{t}}> &= \Lambda_s(t) \\ &= \int_{t_0}^{t} \exp[\mathbf{A}_s\,(t - t')]\ \mathbf{B}_s\,(\mathbf{B}_s)^{\mathrm{t}} \exp[(\mathbf{A}_s)^{\mathrm{t}}\,(t - t')]\ \mathrm{d}t' \end{aligned} \tag{13b}$$

where t denotes the transpose of the matrix, and exp[$\mathbf{A}_s$ t] is defined by (4).

Therefore, we nave:

$$p_s(\mathbf{x}, t \mid \mathbf{x}_0, t_0) = \frac{1}{\sqrt{(2\pi)^d\ |\Lambda_s(t)|}}\ .$$

$$\exp\{-\frac{[\mathbf{x}(t)-\mathbf{m}_s(t)]^{\mathrm{t}}\,(\Lambda_s(t))^{-1}\,[\mathbf{x}(t)-\mathbf{m}_s(t)]}{2}\} \qquad (14)$$

At any fixed time, it corresponds to a Gaussian distribution, but it has a time-varying mean and variance. The mean is following a exponential curve with parameters $\mathbf{x}_s$, $\mathbf{A}_s$, describing the dynamical behaviour of the acoustic vector inside state s. The probability density (14) becomes a Dirac distribution centered on $\mathbf{x}_0$ for $t \to t_0$, reflecting the fact that we indeed observed the point at position $\mathbf{x} = \mathbf{x}_0$. The variance defined by (13b) is the integral of a positive definite matrix; therefore, the variance matrix is continuously increasing, that is, the uncertainty about the position is growing. This provides the justifications for the properties (ii), (iii), and (iv) mentioned in section 2.

A remarquable think is that the Gaussian nature of Ornstein-Uhlenbeck process follows from the continuity of the observations (12), and must not be assumed a priori. Roughly speaking, this results from the central limit theorem: In a small time interval Δt, the process undergoes a large number of infinitesimal uncorrelated fluctuations that results in a Gaussian behaviour.

5. Computation of the path probability density for each state

Let us define for each state $s \in \{0, 1, ..., q\}$ the conditional probability density of observing the acoustic vector $\mathbf{x}(t) = [x_0(t), x_1(t), ..., x_d(t)]^{\mathrm{t}}$ at time t, given that the acoustic vector was $\mathbf{x}_0$ at time t_0, as $p_s(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$. Each conditional probability density $p_s(\mathbf{x}, t \mid \mathbf{x}_0, t_0)$ is characterized by some parameters that will be labelled by s, and is supposed to be independent of the other states. We assume that the process is a continuous time – continuous state Markov process, such as the one described by (14) ([13], [29]). As pointed out in preceding paragraph, this means that the conditional probability density of finding a particular value of acoustic vector $\mathbf{x}$ at time t is determined entirely by the knowledge of the most recent observation.

Let us consider that state s is segmented from t_0 to t_f. By dividing the interval $[t_0, t_f]$ into N equal parts of length $\Delta t = (t_f - t_0)/N$, the path probability density $\mathscr{P}_s[\mathbf{x}(t)]$ can be computed as follows (see Appendix):

$$\mathscr{P}_s[\mathbf{x}(t)] = \lim_{N\to+\infty} [p_s(\mathbf{x}_1, t_1 \mid \mathbf{x}_0, t_0) \ldots p_s(\mathbf{x}_N, t_N \mid \mathbf{x}_{N-1}, t_{N-1})]$$

$$= \lim_{N\to\infty} \frac{1}{\sqrt{((2\pi\,\Delta t)^d\ |\Sigma_s|)^N}} \cdot$$

$$\exp\{-\frac{\sum_{i=1}^{N} [\dot{\mathbf{x}}_{i-1} - \mathbf{A}_s\,(\mathbf{x}_{i-1} - \mathbf{x}_s)]^{\mathrm{t}}\,(\Sigma_s)^{-1}\,[\dot{\mathbf{x}}_{i-1} - \mathbf{A}_s\,(\mathbf{x}_{i-1} - \mathbf{x}_s)]\,\Delta t}{2}\} \qquad (15)$$

where we have posed $\mathbf{x}_i = \mathbf{x}(t_i)$, $\Sigma_s = \mathbf{B}_s\,(\mathbf{B}_s)^{\mathrm{t}}$, and $\dot{\mathbf{x}}_i = \frac{(\mathbf{x}_{i+1} - \mathbf{x}_i)}{\Delta t}$.

This result can be extended to higher order stochastic differential equations. Indeed, since any system of n linear m-order differential equations can be put in the form of a system of (n x m) first order linear differential equations, the vector $\mathbf{x}$ can also be interpreted as a vector containing the acoustic vector, the first derivative, etc (state-space configuration). If, for instance, we decide to modelize the time evolution of the acoustic vector by a p-order linear differential equation, we should construct an augmented vector $\mathbf{x}$ including the acoustic vector and its (p–1) derivatives, as well as the corresponding matrix $\mathbf{A}_s$. Of course, the weakness of the method is that we need the derivatives of the acoustic vector, but since the parameters are supposed to be fixed during the estimation stage, these derivatives can be evaluated in a reliable way by using a Kalman-Bucy filter for each state (see, for instance, [19]).

We observe that assuming that the acoustic vectors are generated by a linear first-order stochastic differential equation leads to the addition of a "time-derivative" to these vectors in the computation of the distance. This kind of feature transformation has been studied by Furui [12] and Gurgen, Sagayama & Furui [15] for LPC analysis. They call it "combination of instantaneous and transitional LPC frequencies in the parameter domain". They compared this feature combination with the more usual "combination in the distance domain", which means that an augmented vector $(\mathbf{x}, \dot{\mathbf{x}})$ is constructed. Experimental results [15] show that the first method has a slightly better recognition performance than the second. Combination in the parameter domain is also advantageous in terms of computation; that is combination can be obtained during speech signal analysis and thus it does not result in extra computation at the recognition level [15].

6. Segmentation and parameters estimation

Segmentation and parameters estimation (by Viterbi algorithm) is based on the maximisation of the likelihood:

$$\mathscr{L} = [\prod_{k=1}^{Q} \pi(s_k \mid s_{k-1})] \cdot [\prod_{k=0}^{Q} \mathscr{P}_{s_k}[\mathbf{x}(t)]]$$

A detailed derivation can be found in [34].

7. Conclusion

In this work, we have considered that, on each state, the acoustic vectors are generated by a linear first-order stochastic differential equation. This is to be opposed to the current assumption, i.e. that acoustic vectors are emitted according to a probability distribution that only depends on the current acoustic vector observed. This allows us to consider the time trajectory of the acoustic vector as a continuous dynamic path, and to derive the probability distribution of observing this trajectory, given the state. It measures the "distance" of the observed trajectory to an ideal trajectory (not corrupted by noise), which is supposed to be modelled by a linear differential equation. Once the segmentation is fixed, reestimation formulae for the parameters of the continuous Markov process can be derived for the Viterbi algorithm [34]. The segmentation can be obtained by sampling the continuous process, and by applying dynamic programming to find the best path over all the possible segmentations of states [34]. This provides some enlightenments to related work of Poritz [31], Juang [21], Juang & Rabiner [22], Furui ([11], [12]) and Gurgen, Sagayama & Furui [15]. In another paper [34], where we considered a less general continuous Markov process, we have shown that duration models are easily introduced, but dynamic programming must then be performed in three dimensions to find the best path through all the possible successions of states and all the possible durations [3]. We also sketched a possible generalization to path mixtures, for which different trajectories are available in each state.

References

[1] Aström K. & Wittenmark B. (1990). *Computer-controlled systems, 2nd ed.* Prentice-Hall.

[2] Bourlard H. & Morgan N. (1994). *Connectionist speech recognition. A hybrid approach.* Kluwer Academic Press.

[3] Bourlard H. & Wellekens C. (1986). Connected speech recognition by phonemic semi-Markov chains for state occupancy modelling. *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Young I.T. et al. (editors), pp. 511-514.

[4] Deng L., Hassanein K. & Elmasry M. (1991). Neural-network architecture for linear and nonlinear predictive hidden Markov models: Application to speech recognition. In Juang H., Kung S. & Kamm C. (editors), *Proceedings of the 1991 IEEE Workshop on Neural Networks for Signal Processing*, Princeton New Jersey, pp. 411-421.

[5] Deng L. (1992a). A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal. *Signal Processing*, **27** (1), pp. 65-78.

[6] Deng L. (1992b). A class of nonstationary-state hidden Markov models with applications to speech modeling and recognition. *Proceedings of the International Conference on Signal Processing Applications and Technology*, Boston, pp. 1025-1034.

[7] Digalakis V., Rohlicek J.R. & Ostendorf M. (1991). A dynamical system approach to continuous speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, pp. 289-292.

[8] Digalakis V., Rohlicek J.R. & Ostendorf M. (1993). ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. *IEEE Transactions on Speech and Audio Processing*, **1** (4), pp. 431-442.

[9] Feynman R. (1948). Space-time approach to nonrelativistic quantum mechanics. *Reviews of Modern Physics*, **20** (2), pp. 367-387.

[10] Feynman R. & Hibbs A. (1965). *Quantum mechanics and path integrals.* Mc Graw-Hill.

[11] Furui S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **ASSP-34** (1), pp. 52-59.

[12] Furui S. (1991). Recent advances in speech recognition. *Proceedings of EUROSPEECH*, Genova, pp. 3-10.

[13] Gardiner C.W. (1985). *Handbook of stochastic methods.* Springer-Verlag.

[14] Gel'fand I.M. & Yaglom A.M. (1960). Integration in functional spaces and its applications in quantum physics. *Journal of Mathematical Physics*, **1** (1), pp. 48-69.

[15] Gurgen F., Sagayama S. & Furui S. (1990). Line spectrum pair frequency-based distance measures for speech recognition. *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, Japan, pp. 13.1.1-13.1.4.

[16] Huang X.D., Ariki Y. & Jack M.A. (1990). *Hidden Markov models for speech recognition.* Edinburgh University Press.

[17] Iso K.-I. & Watanabe T. (1990). "Speaker-independent word recognition using a neural prediction model". *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Albuquerque, pp. 441-444.

[18] Iso K.-I. & Watanabe T. (1991). Large vocabulary speech recognition using neural prediction model. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Toronto, pp. 57-60.

[19] Jazwinski A. (1970). *Stochastic processes and filtering theory.* Academic Press.

[20] Jelinek F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE,* **64**, pp. 532-556.

[21] Juang B.-H. (1984). On the hidden Markov model and dynamic time warping for speech recognition – A unified view. *AT&T Laboratories Technical Journal,* **63** (7), pp. 1213-1243.

[22] Juang B.-H. & Rabiner L. (1985). Mixture autoregressive hidden Markov models for speech signal. *IEEE Transactions on Acoustics, Speech and Signal Processing,* **ASSP-33** (6), pp. 1404-1413.

[23] Kenny P., Lennig M. & Mermelstein P. (1990). A linear predictive HMM for vector-valued observations with applications to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing,* **ASSP-38** (2), pp. 220-225.

[24] Levin E. (1990). Word recognition using hidden control neural architectures. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Albuquerque, pp. 433-436.

[25] Levin E. (1991). Modeling time varying systems using hidden control neural architecture. Proceedings of *Advances in Neural Information Processing Systems (NIPS),* 3, pp. 147-154.

[26] Levinson S.E., Rabiner L.R. & Sondhi M.M. (1983). An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal,* **62** (4), pp. 1035-1074.

[27] Middleton R. & Goodwin G. (1986). Improved finite word length characteristics in digital control using delta operators. *IEEE Transactions on Automatic Control,* **AC-31 (11)**, pp. 1015-1021.

[28] Montroll E. (1952). Markoff chains, Wiener integral and quantum theory. *Communications on Pure and Applied Mathematics,* **5**, pp. 415-453.

[29] Papoulis A. (1991). *Probability, random variables, and stochastic processes, 3th ed.* McGraw-Hill.

[30] Petek B., Waibel A. & Tebelskis M. (1992). Integrated phoneme and function word architecture of hidden control neural networks for continuous speech recognition. *Speech Communication,* **11**, pp. 273-282.

[31] Poritz A. (1982). Linear predictive hidden Markov models and the speech signal. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Paris, pp. 1291-1294.

[32] Rabiner L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* **77** (2), pp. 257-286.

[33] Rao G. P. & Sinha N. K. (1991). Continuous-time models and approaches. In *Identification of continuous-time systems,* Sinha N. K. & Rao G. P. (editors), Kluwer Academic Publishers.

[34] Saerens M. (1993). A continuous-time dynamic formulation of Viterbi algorithm for one-Gaussian-per-state hidden Markov models. *Speech Communication,* **12** (4), pp. 321-333.

[35] Saerens M. & Bourlard H. (1993). Linear and nonlinear prediction for speech recognition with hidden Markov models. *Proceedings of the EUROSPEECH Conference,* Berlin, pp. 807-810.

[36] Schulman L. (1981). *Techniques and applications of path integration.* John Wiley & Sons.

[37] Tebelskis J. & Waibel A. (1990). Large vocabulary recognition using linked predictive neural networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Albuquerque, pp. 437-440.

[38] Tebelskis J., Waibel A., Petek B. & Schmidbauer O. (1991). Continuous speech recognition using linked predictive neural networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Toronto, pp. 61-64.

[39] Tishby N. (1991). On the application of mixture AR hidden Markov models to text independent speaker recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing,* **ASSP-39** (3), pp. 563-570.

[40] Tsuboka E., Takada Y. & Wakita H. (1990). Neural predictive hidden Markov model. *Proceedings of the 1990 International Conference on Spoken Language Processing,* Kobe, Japan, pp. 31.2.1-31.2.4.

[41] Wellekens C. (1987). Explicit time correlation in hidden Markov models for speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Dallas, pp. 384-386.

[42] Wiener N. (1930). Generalized harmonic analysis. *Acta Mathematica,* **55**, pp. 117-258.

[43] Woodland P. (1992). Hidden Markov models using vector linear prediction and discriminative output distributions. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* San Francisco, pp. 509-512.

APPENDIX: COMPUTATION OF THE PATH PROBABILITY DENSITY

We have to compute the probability density $\mathscr{P}_s[\mathbf{x}(t)]$ of an observed time trajectory of the acoustic vector $\mathbf{x}(t)$ on state s. This can be done by using the concept of path integral, widely used in theoretical physics (see, for instance, [9], [10], [14], [36]). A similar calculus has already been carried out for a particular continuous-time Markov process [34]; see this paper for more details.

Now, the probability that an acoustic vector starting at $(\mathbf{x}_0, t_0)$ is inside domain Ω_1 at time t_1, is inside domain Ω_2 at time t_2, ..., and is inside domain Ω_N at $t_N = t_f$ $(t_i < t_{i+1})$ is given by:

$$\int_{\Omega_1} d\mathbf{x}_1 \ldots \int_{\Omega_N} d\mathbf{x}_N \;\; p_s(\mathbf{x}_N, t_N \mid \mathbf{x}_{N-1}, t_{N-1}) \ldots p_s(\mathbf{x}_2, t_2 \mid \mathbf{x}_1, t_1) \;\; p_s(\mathbf{x}_1, t_1 \mid \mathbf{x}_0, t_0) \qquad \text{(A1)}$$

with $\mathbf{x}_i = \mathbf{x}(t_i)$. For $N \to +\infty$, and $\Delta t = (t_i - t_{i-1}) = (t_f - t_0)/N \to 0$, we can write (A1) symbolically as

$$\int_{\Omega(t)} \mathscr{P}_s[\mathbf{x}(t)] \; \mathscr{D}\mathbf{x}(t) \;, \qquad \text{(A2)}$$

which means that we are integrating over all the continuous paths $\mathbf{x}(t)$ lying in domain $\Omega(t)$, with initial condition $\mathbf{x}(t_0) = \mathbf{x}_0$. It has been shown ([42], [14]) that we obtain a measure on the space of continuous paths $\mathbf{x}(t)$ with $\mathbf{x}$ equal to $\mathbf{x}_0$ at $t = t_0$. We have now to compute

$$\mathscr{P}_s[\mathbf{x}(t)] = \lim_{N \to +\infty} \; [p_s(\mathbf{x}_1, t_1 \mid \mathbf{x}_0, t_0) \;\; p_s(\mathbf{x}_2, t_2 \mid \mathbf{x}_1, t_1) \ldots p_s(\mathbf{x}_N, t_N \mid \mathbf{x}_{N-1}, t_{N-1})]$$

for the conditional probability density defined by (14).

Since $\Delta t \to 0$, we can expand the mean and the variance to the first order at the initial value (for small values Δt in comparison with the characteristic response times of the process defined by (11), but large compared to the time interval between the small random fluctuations). For instance, in the case of $p_s(\mathbf{x}_1, t_1 \mid \mathbf{x}_0, t_0)$, we obtain at the first order:

$$\begin{aligned} \mathbf{x}(t_1) - \mathbf{m}_s(t_1) &\equiv \mathbf{x}(t_1) - [1 + \mathbf{A}_s \Delta t] \, (\mathbf{x}(t_0) - \mathbf{x}_s) - \mathbf{x}_s \\ &= [\, \dot{\mathbf{x}}_0 - \mathbf{A}_s \, (\mathbf{x}_0 - \mathbf{x}_s)] \, \Delta t \end{aligned} \qquad \text{(A3)}$$

where we have posed $\dot{\mathbf{x}}_i = \dfrac{(\mathbf{x}_{i+1} - \mathbf{x}_i)}{\Delta t}$. For the variance, we obtain:

$$\Lambda_s(t_1) \equiv \mathbf{B}_s \, (\mathbf{B}_s)^{\mathrm{t}} \, \Delta t \qquad \text{(A4)}$$

Let us now define

$$\Sigma_s = \mathbf{B}_s \, (\mathbf{B}_s)^{\mathrm{t}} \qquad \text{(A5)}$$

$$\Lambda_s(t_1) \equiv \Sigma_s \, \Delta t \qquad \text{(A6)}$$

Now, we can evaluate (for more details, see [34])

$$\begin{aligned} \mathscr{P}_s[\mathbf{x}(t)] &= \lim_{N \to +\infty} \; [p_s(\mathbf{x}_1, t_1 \mid \mathbf{x}_0, t_0) \;\; p_s(\mathbf{x}_2, t_2 \mid \mathbf{x}_1, t_1) \ldots p_s(\mathbf{x}_N, t_N \mid \mathbf{x}_{N-1}, t_{N-1})] \\ &= \lim_{N \to \infty} \frac{1}{\sqrt{((2\pi \, \Delta t)^d \; |\Sigma_s|)^N}} \; . \\ &\quad \exp\{- \frac{\sum_{i=1}^{N} \; [\dot{\mathbf{x}}_{i-1} - \mathbf{A}_s \, (\mathbf{x}_{i-1} - \mathbf{x}_s)]^{\mathrm{t}} \, (\Sigma_s)^{-1} \; [\dot{\mathbf{x}}_{i-1} - \mathbf{A}_s \, (\mathbf{x}_{i-1} - \mathbf{x}_s)] \, \Delta t}{2} \} \end{aligned} \qquad \text{(A7)}$$

which can also be written symbolically as

$$\mathscr{P}_s[\mathbf{x}(t)] = \lim_{N \to +\infty} \frac{1}{\sqrt{((2\pi \, \Delta t)^d \; |\Sigma_s|)^N}} \; . \; \exp\{- \frac{\int_{t_0}^{t_f} [\dot{\mathbf{x}} - \mathbf{A}_s \, (\mathbf{x} - \mathbf{x}_s)]^{\mathrm{t}} \, (\Sigma_s)^{-1} \; [\dot{\mathbf{x}} - \mathbf{A}_s \, (\mathbf{x} - \mathbf{x}_s)] \; dt}{2} \} \qquad \text{(A8)}$$