A continuous-time dynamic formulation of Viterbi algorithm for one-Gaussian-per-state hidden Markov models

Marco Saerens

IRIDIA Laboratory – cp. 194 /6, Université Libre de Bruxelles, 50, av. F. Roosevelt, B-1050 Bruxelles, Belgium

Received 29 September 1992
Revised 1 March 1993

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 a continuous Markov process. Indeed, the time evolution of the acoustic vector is inherently dynamic and continuous, and sampling only occurs for the purpose of computation. This allows us to assign a probability density to the time trajectory of the acoustic vector inside the state, reflecting the probability that this particular path has been generated by the continuous Markov process associated with this state. Roughly speaking, it measures the “adequacy” of the observed trajectory with respect to an ideal trajectory, which is modelled by a vectorial linear differential equation. This model is introduced in order to describe the dynamic behaviour of the acoustic vector inside a state. Once the segmentation is fixed, reestimation formulæ for the parameters of the continuous Markov process are derived for the Viterbi algorithm. 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 and all the possible durations. Finally, we sketch a possible generalization to path mixtures, for which different trajectories are available in each state. However, we have to stress that no experimental results are available at present. Indeed, we did not have the opportunity to test the algorithm on real speech. We are aware of the fact that the assumptions we did may not be appropriate for the modelling of speech.


Résumé. Lorsqu'on utilise des modèles de Markov cachés pour la reconnaissance de la parole, on fait habituellement l'hypothèse que la probabilité d'émission d'un vecteur acoustique ne dépend que de l'état courant et du vecteur acoustique actuellement observé. Dans cet article, nous envisageons une hypothèse moins restrictive: nous considérons qu'au sein d'un même état de la chaîne de Markov, les vecteurs acoustiques sont générés par un processus markovien continu. En effet, l'évolution temporelle du signal de parole est intrinsèquement continue, et l'échantillonnage n'est réalisé que pour les besoins de calculs numériques. Nous assignons une densité de probabilité à la trajectoire temporelle observée du vecteur.

* Part of this research was carried out when the author was working with Lernout & Hauspie Speech Products, Belgium.
acoustique, reflétant la probabilité que cette trajectoire particulière a été générée par le processus markovien continu associé à l'état. Elle mesure "l'adéquation" de cette trajectoire observée par rapport à une trajectoire idéale, supposée générée par une équation différentielle vectorielle linéaire. Une fois la segmentation fixée, les formules de réestimation pour les paramètres du modèle markovien continu peuvent être calculées dans le cadre de l'algorithme de Viterbi. La segmentation la plus probable (au sens du maximum de vraisemblance) s'obtient en échantillonnant le processus continu, et en calculant le meilleur chemin à travers toutes les segmentations possibles et toutes les durées possibles par programmation dynamique. Ensuite, nous mentionnons une extension possible à des mélanges de trajectoires, pour laquelle plusieurs trajectoires sont accessibles dans un même état. Cependant, aucun test expérimental n'a encore été effectué pour le moment, si bien que les hypothèses que nous avons faites ne pourraient pas s'appliquer au signal de parole.

Keywords. Speech recognition; Hidden Markov models; Viterbi algorithm

1. Introduction

Hidden Markov models are widely used for speech recognition (Jelinek, 1976; Levinson et al., 1983; Rabiner, 1989; Huang et al., 1990). However, strong assumptions have to be made to render the model computationally tractable (see, for example, (Bourlard, 1992)). 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.

In this paper, we introduce another idea, i.e. we assume that, in a given state, the acoustic vectors are generated by a continuous Markov process. Indeed, the time evolution of the acoustic vector is inherently dynamic and continuous, and sampling only occurs for the purpose of computation. In other words, we assign a probability density to the time trajectory of the acoustic vector inside the state, reflecting the probability that this particular path has been generated by the continuous Markov process associated with the state. This computation relies on the concept of path integral – also known as Wiener integral (Wiener, 1930) – widely used in theoretical physics (see, for instance, (Feynman, 1948; Feynman and Hibbs, 1965; Gel'fand and Yaglom, 1960; Schulman, 1981)).

As usual, the sequence of states is assumed to be 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 a semi-Markov process. 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. In this way, we hope to capture something about the dynamics of the speech vectors. Many authors have tried to take account of the dynamics of the speech. For instance, Furui (1986, 1991) and Gurgen et al. (1990) introduce features including the time-derivative of the acoustic vectors. Deng (1992a, 1992b) models the temporal evolution of the acoustic vector inside a state by a given function of time, i.e. a polynomial trend function of time $t$ spent in the state. Wellekens (1987) assumes explicit dependence between the current vector and the last observed vector. He shows that, in the case of a correlated Gaussian probability density function, the emission probabilities depend on the prediction error of a first order linear predictor. On the other hand, Poritz (1982), Juang (1984), Juang and Rabiner (1985) (see also (Kenny et al., 1990; Tishby, 1991; Woodland, 1992)) use Gaussian autoregressive densities per state; the Baum–Welch algorithm is applied for the reestimation of the parameters. Once more, the emission probabilities depend on the prediction error of a linear predictor. This special case is closely related to our work, while we are dealing with continuous processes. An extension of the Gaussian autoregressive densities model, in which we allow the autoregressive coefficients to be stochastic variables, is presented by Saerens and Bourlard (1993).

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.
(Tsuboka et al., 1990; Levin, 1990, 1991; Tebelskis and Waibel, 1990; Tebelskis et al., 1991; Petek et al., 1992; Iso and Watanabe, 1990, 1991; Deng et al., 1991). In this case, the acoustic vectors are assumed to be generated at each frame by a discrete non-linear process, different for each state, corrupted by an additive uncorrelated Gaussian noise. It generalizes the work mentioned above, where linear prediction (auto-regressive) models were considered. Another interesting work, relying on a dynamical system approach with parameter training based on the EM algorithm, can be found in (Digalakis et al., 1991).

Our approach leads to the introduction of a time-derivative which is to be added to the acoustic vector, and is therefore also related to Furui (1986, 1991) and Gurgan et al.'s work (1990), 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 continuous Markov process are derived for the Viterbi algorithm. 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 and all the possible durations (Bourdard and Wellekens, 1986). Finally, we sketch a possible generalization to path mixtures, for which different trajectories are available in each state.

2. Path probability density and continuous Markov processes

We will assume that, in each state $s$, the acoustic vector $x(t) = (x_0(t), x_1(t), \ldots, x_d(t))^T$ are generated by a specific continuous time-continuous state Markov process (Gardiner, 1985; Papoulis, 1991). This continuous Markov process is defined by its conditional probability density $p_c(x, t | x_0, t_0)$, that is, for each state $s \in \{0, 1, \ldots, q\}$, the conditional probability density of observing the acoustic vector $x(t) = (x_0(t), x_1(t), \ldots, x_d(t))^T$ at time $t$, given that the acoustic vector was $x_0$ at time $t_0$. Each probability density $p_c(x, t | x_0, t_0)$ is characterized by some parameters that will be labelled by $s$, and is independent of the other states. Assuming that the process is a continuous time-continuous state Markov process means that the conditional probability density of finding a particular value of the acoustic vector $x$ at time $t$ is determined entirely by the knowledge of the most recent condition.

In the case of a continuous time-continuous state Markov process, the conditional probability density satisfies the Chapman–Kolmogorov equation (Gardiner, 1985; Papoulis, 1991):

$$p_c(x, t | x_0, t_0) = \int_{-\infty}^{+\infty} p_c(x, t | x', t') p_c(x', t' | x_0, t_0) \, dx',$$

(1)

and also

$$\lim_{t \to t_0} p_c(x, t | x_0, t_0) = \delta(x - x_0),$$

(2)

where $\delta(x - x_0)$ is the Dirac distribution.

The model will be formulated in the framework of word recognition with the Viterbi algorithm. In the future, 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 $(s_0, s_1, \ldots, s_Q)$ defining a word, together with its segmentation, and $X$ is the time evolution of the acoustic vector $x(t)$, observed over the whole word. We have

$$P(X, S) = P(X \mid S) P(S).$$

Since the sequence of states is assumed to be modelled by a discrete Markov process, and since the time evolution of the observations arising from any state is independent, we can write

$$P(S) = \left[ \prod_{k=1}^{Q} \pi(s_k | s_{k-1}) \right] \pi_0,$$

$$P(X \mid S) = \prod_{k=0}^{Q} \mathcal{P}_s[x(t)] ,$$

where $\pi(s_k | s_{k-1})$ is the transition probability of the discrete Markov model of states, and $\mathcal{P}_s[x(t)]$ is the probability density of the observed acoustic vector trajectory $x(t)$ on state $s$. The problem is to compute the probability density of a path $\mathcal{P}_s[x(t)]$ on a state $s$ in terms of the conditional
probability density \( p_s(x, t \mid x_0, t_0) \) defining the continuous Markov process.

We will assume that the segmentation procedure (Section 5) indicates that state \( s \) is (most likely) segmented from \( t_0 \) to \( t_f \); that is, \( t \in [t_0, t_f] \). Let us first consider the one-dimensional case (\( x \) is a scalar, i.e. \( x = x \)). If \( x = x_0 \) at \( t = t_0 \), then at time \( t \), the probability that value \( x \) is between \( a \) and \( b \) is

\[
\int_a^b p_s(x', t \mid x_0, t_0) \, dx'.
\]  

Similarly, the probability that an acoustic vector starting at \((x_0, t_0)\) is between \(a_1\) and \(b_1\) at \(t_1\), is between \(a_2\) and \(b_2\) at \(t_2\), and so on, then at time \(t_N\) is given by

\[
\int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N \, p_s(x_N, t_N \mid x_{N-1}, t_{N-1}) \cdots p_s(x_2, t_2 \mid x_1, t_1) p_s(x_1, t_1 \mid x_0, t_0),
\]  

with \(x_i = x(t_i)\).

Now, let us divide the interval \([t_0, t_f]\) into \(N\) equal parts of length \(\Delta t = (t_f - t_0) / N\) (Figure 1). For the simple diffusion problem, it was demonstrated by Wiener (1930) that, when we take the limit \(N \to +\infty\) of expression (4), we obtain a measure on the space of continuous paths \(x(t)\) with \(x\) equal to \(x_0\) at \(t = t_0\); this is the Wiener measure (for more details, see Gel'fand and Yaglom, 1960; Schulman, 1981). In the case of an arbitrary continuous Markov process, we still obtain a measure (Gel'fand and Yaglom, 1960), and a general method for computing the analytic form of (4), based on the work of Kas, is reviewed in Gel'fand and Yaglom (1960). We will not enter into details, since the discrete form given by (4) will suffice.

Symbolically, expression (4), for \(N \rightarrow +\infty\), can be written in the form

\[
\int_{a(t)}^{b(t)} \mathcal{P}_x[x(t)] \, dx(t),
\]  

which means that we are integrating over all the continuous paths \(x(t)\) lying between the functions \(a(t)\) and \(b(t)\). \(\mathcal{P}_x[x(t)]\) is often referred to as the path probability density (Feynman and Hibbs, 1965). Notice that, from Chapman–Kolmogorov equation (1), if we consider all the possible paths, i.e. if \(a_t = a_2 = \cdots = a_N = -\infty\) and \(b_1 = b_2 = \cdots = b_N = +\infty\), the result of (4) is one.

The extension to a multidimensional space is straightforward. If we are dealing with \(d\)-dimensional vectors, we have to define a domain \(\Omega_i\) for each time \(t_i\). In this case, the probability that an acoustic vector starting at \((x_0, t_0)\) is inside domain \(\Omega_1\) at time \(t_1\), is inside domain \(\Omega_2\) at time \(t_2\), and so on, and is inside domain \(\Omega_N\) at \(t_N = t_f\) \(\forall t_i < t_{i+1}\), is given by

\[
\int_{\Omega_1} \cdots \int_{\Omega_N} \mathcal{P}_x[x_N, t_N \mid x_{N-1}, t_{N-1}] \cdots \mathcal{P}_x[x_2, t_2 \mid x_1, t_1] \mathcal{P}_x(x_1, t_1 \mid x_0, t_0),
\]  

with \(x_i = x(t_i)\). As before, for \(N \rightarrow +\infty\), we can write (6) symbolically as

\[
\int_{\Omega(t)} \mathcal{P}_x[x(t)] \, dx(t).
\]  

3. A particular continuous Markov process

For the sake of simplicity, we will first compute the path probability density, as defined by (6), for a scalar, i.e. \(x(t) = x(t)\). Let us consider a particular form of \(p_s(x, t \mid x_0, t_0)\):

\[
p_s(x, t \mid x_0, t_0) = \frac{\beta}{\sqrt{2\pi\sigma_s(1 - \exp[-\beta(t-t_0)])}} \times \exp\left(-\beta\left[(x-x_0) - \exp[\alpha_s(t-t_0)]\right]ight) \times (x_0-x_0)^2 \times [2\sigma_s(1 - \exp[-\beta(t-t_0)])^{-1}],
\]
with $\beta > 0$. This continuous Markov process is known as the Ornstein–Uhlenbeck process for $\alpha_x = -\beta/2$ (Gardiner, 1985; Cox and Miller, 1965; Papoulis, 1991), and occurs, for instance, for the velocity of a particle in Brownian motion. This distribution appears as the probability density of the solutions of a first-order ordinary linear differential equation in which a rapidly and irregularly fluctuating random function of time occurs (Gardiner, 1985). At any fixed time, it corresponds to a Gaussian density, but it has a time-varying mean and variance. The variance tends to a constant value for $t \to +\infty$, while at $t \to t_0$, the probability density (8) becomes a Dirac distribution centered on $x_0$, reflecting the fact that we indeed observed the point at position $x = x_0$. As time goes on, the position is becoming more fuzzy. The mean is following an exponential curve with parameters $x_0$, $\alpha_x$, where $x_0$ is the asymptotic value and $\alpha_x$ is the relaxation time. A remarkable think is that the Gaussian nature of the Ornstein–Uhlenbeck process follows from the assumed continuity of the observations, and must not be assumed a priori (see, for example, (Gardiner, 1985)).

Clearly, (2) is fulfilled. Moreover, if $\alpha_x < 0$, we have

$$
\lim_{t \to +\infty} p_x(x, t \mid x_0, t_0) = \frac{1}{\sqrt{2\pi \sigma_x^2}} \exp \left[ -\frac{(x-x_0)^2}{2\sigma_x^2} \right],
$$

with $\sigma_x = \sigma_x/\beta$, and we recover a standard Gaussian distribution centred on $x_0$. This provides a clear interpretation of the process defined by (8): We do not expect the value of $x$ to reach $x_0$ immediately after $t_0$, the moment where we enter the state $x$. Instead, we expect the value $x$ to be near $x_0$ at the beginning and to move gradually towards $x$. This is what (8) indicates. On the contrary, if $\alpha_x > 0$, the point moves away from the asymptotic value.

Let us resume the main reasons that lead us to the introduction of the process defined by (8) for the purpose of speech modelling. (i) It takes account of the dynamics of the acoustic vectors since the mean value is supposed to follow a temporal trajectory. In (8), the trajectory is a simple exponential curve but, in fact, the model is more general. Indeed, we will show later that the acoustic vector is supposed to follow the solution of a $p$-order vectorial linear differential equation in average. The dynamics of the vector is therefore represented by a vectorial differential equation subject to random fluctuations, for each state. (ii) 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 fuzzy. (iii) The process is continuous. Indeed, as already mentioned in the introduction, the time evolution of the acoustic vector is inherently dynamic and continuous, and sampling only occurs for the purpose of computation. This makes the modelling independent of the sampling frequency, which is not the case, e.g., for a discrete autoregressive model. (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.

Now, we have to compute

$$
\mathcal{F}[x(t)] = \lim_{N \to +\infty} \prod_{i=1}^N \left[ \frac{1}{\sqrt{2\pi \sigma_x^2}} \exp \left[ -\frac{(x-x_i)^2}{2\sigma_x^2} \right] \right]
$$

with $\sigma_x = \sigma_x/\beta$, and we recover a standard Gaussian distribution centred on $x_0$. This provides a clear interpretation of the process defined by (8): We do not expect the value of $x$ to reach $x_0$ immediately after $t_0$, the moment where we enter the state $x$. Instead, we expect the value $x$ to be near $x_0$ at the beginning and to move gradually towards $x$. This is what (8) indicates. On the contrary, if $\alpha_x > 0$, the point moves away from the asymptotic value.

Let us resume the main reasons that lead us to the introduction of the process defined by (8) for the purpose of speech modelling. (i) It takes account of the dynamics of the acoustic vectors since the mean value is supposed to follow a temporal trajectory. In (8), the trajectory is a simple exponential curve but, in fact, the model is more general. Indeed, we will show later that the acoustic vector is supposed to follow the solution of a $p$-order vectorial linear differential equation in average. The dynamics of the vector is therefore represented by a vectorial differential equation subject to random fluctuations, for each state. (ii) 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 fuzzy. (iii) The process is continuous. Indeed, as already mentioned in the introduction, the time evolution of the acoustic vector is inherently dynamic and continuous, and sampling only occurs for the purpose of computation. This makes the modelling independent of the sampling frequency, which is not the case, e.g., for a discrete autoregressive model. (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.

Now, we have to compute

$$
\mathcal{F}[x(t)] = \lim_{N \to +\infty} \prod_{i=1}^N \left[ \frac{1}{\sqrt{2\pi \sigma_x^2}} \exp \left[ -\frac{(x-x_i)^2}{2\sigma_x^2} \right] \right]
$$

with $\sigma_x = \sigma_x/\beta$, and we recover a standard Gaussian distribution centred on $x_0$. This provides a clear interpretation of the process defined by (8): We do not expect the value of $x$ to reach $x_0$ immediately after $t_0$, the moment where we enter the state $x$. Instead, we expect the value $x$ to be near $x_0$ at the beginning and to move gradually towards $x$. This is what (8) indicates. On the contrary, if $\alpha_x > 0$, the point moves away from the asymptotic value.

Let us resume the main reasons that lead us to the introduction of the process defined by (8) for the purpose of speech modelling. (i) It takes account of the dynamics of the acoustic vectors since the mean value is supposed to follow a temporal trajectory. In (8), the trajectory is a simple exponential curve but, in fact, the model is more general. Indeed, we will show later that the acoustic vector is supposed to follow the solution of a $p$-order vectorial linear differential equation in average. The dynamics of the vector is therefore represented by a vectorial differential equation subject to random fluctuations, for each state. (ii) 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 fuzzy. (iii) The process is continuous. Indeed, as already mentioned in the introduction, the time evolution of the acoustic vector is inherently dynamic and continuous, and sampling only occurs for the purpose of computation. This makes the modelling independent of the sampling frequency, which is not the case, e.g., for a discrete autoregressive model. (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.

Now, we have to compute

$$
\mathcal{F}[x(t)] = \lim_{N \to +\infty} \prod_{i=1}^N \left[ \frac{1}{\sqrt{2\pi \sigma_x^2}} \exp \left[ -\frac{(x-x_i)^2}{2\sigma_x^2} \right] \right]
$$

with $\sigma_x = \sigma_x/\beta$, and we recover a standard Gaussian distribution centred on $x_0$. This provides a clear interpretation of the process defined by (8): We do not expect the value of $x$ to reach $x_0$ immediately after $t_0$, the moment where we enter the state $x$. Instead, we expect the value $x$ to be near $x_0$ at the beginning and to move gradually towards $x$. This is what (8) indicates. On the contrary, if $\alpha_x > 0$, the point moves away from the asymptotic value.
By computing a first order expansion of \( \exp(-\beta \Delta t) \) and \( \exp(\alpha \Delta t) \), we obtain
\[
\begin{align*}
\lim_{N \to +\infty} p_t(x_1, t_1, x_0, t_0) p_t(x_2, t_2 | x_1, t_1) \\
\cdots p_t(x_N, t_N | x_{N-1}, t_{N-1})
&= \lim_{N \to +\infty} \frac{1}{\sqrt{(2\pi\sigma_s)^N(\Delta t)^N}} \\
&\quad \times \exp \left[ -\frac{\sum_{i=1}^{N} (\dot{x}_{i-1} - \alpha_i(x_{i-1} - x_s))^2 \Delta t}{2\sigma_s} \right].
\end{align*}
\]
(9a)

where we posed \( \dot{x}_i = (x_{i+1} - x_i)/\Delta t \).

This expression can also be written symbolically as
\[
\begin{align*}
\mathcal{R}_t[x(t)] &= \lim_{N \to +\infty} \frac{1}{\sqrt{(2\pi\sigma_s)^N(\Delta t)^N}} \\
&\quad \times \exp \left[ -\frac{\int_{t_0}^{t} (\dot{x} - \alpha_i(x - x_s))^2 \, dt}{2\sigma_s} \right].
\end{align*}
\]
(9b)

Of course, this value is maximum for
\[
\dot{x} - \alpha_i(x - x_s) = 0,
\]
(10)
so that the best path is
\[
x - x_s = \exp[\alpha_i(t - t_0)](x_0 - x_s).
\]
(11)

Roughly speaking, (9a, 9b) measures the "adequacy" of the observed trajectory with respect to the ideal trajectory given by (11).

Let us now consider the multidimensional case. It is well known that the solution of a system of \( d \) first order linear differential equations,
\[
\frac{dx}{dt} = A_s(x - x_s),
\]
(12)
where \( A_s \) is the \( d \times d \) real square matrix, can be written in the form (see, for example, (Friedman, 1956; Golub and Van Loan, 1989))
\[
x(t) - x_s = \exp[A_s(t - t_0)](x_0 - x_s),
\]
(13)
where \( \exp[A_s t] \) is defined as
\[
\exp[A_s t] = \sum_{n=0}^{\infty} \frac{[A_s]^n}{n!} t^n.
\]
(14)

Therefore, we consider a process of the kind
\[
\begin{align*}
&= \sqrt{\frac{\beta^d}{(2\pi)^d|\Sigma_i| (1 - \exp[-\beta(t - t_0)])^d}} \\
&\quad \times \exp \left[ -\beta [(x - x_s) - \exp[A_s(t - t_0)]] \\
&\quad \times (x_0 - x_s)^T (\Sigma_s)^{-1} [(x - x_s) \\
&\quad - \exp[A_s(t - t_0)](x_0 - x_s)] \\
&\quad / (2 \times \exp[-\beta(t - t_0)]),
\end{align*}
\]
(15)
where \( T \) denotes the transpose of the matrix and \( |\Sigma_s| \) is the determinant of the symmetrical positive definite matrix \( \Sigma_s \). The mean of the acoustic vector is varying in time and is following the solution of a system of first-order linear differential equations. The variance-covariance is growing towards the asymptotic value \( \Sigma_s/\beta \).

In this case, we have
\[
\begin{align*}
\mathcal{R}_t[x(t)] &= \lim_{N \to +\infty} \frac{1}{\sqrt{(2\pi\Delta t)^d|\Sigma_s|)}^N} \\
&\quad \times \exp \left[ -\frac{1}{2} \sum_{i=1}^{N} (\dot{x}_{i-1} - A_s(x_{i-1} - x_s))^T \\
&\quad \times (\Sigma_s)^{-1} (\dot{x}_{i-1} - A_s(x_{i-1} - x_s)) \Delta t \right].
\end{align*}
\]
(16)

The process defined by (15) is no more an Ornstein-Uhlenbeck process, while it is closely related. The multivariate Ornstein-Uhlenbeck process has a somewhat more complicated form (Gardiner, 1985), and generalization to this case, assuming acoustic vectors generated by a linear stochastic differential equation in each state, is developed elsewhere (Saerens, 1993).

Now, since any system of \( n \) linear \( m \)-order differential equations can be put in the form of a system of \( n \times m \) first-order linear differential
equations, the vector $x$ can also be interpreted as a vector containing the acoustic vector, the first derivative, etc. (state-space configuration). If we think that the ideal time evolution of the acoustic vector is correctly modelled by a $p$-order linear differential equation, we should construct an augmented vector $\tilde{x}$ including the acoustic vector and its $p-1$ derivatives, as well as the corresponding matrix $A_t$. As a special example, the case where the acoustic signal is supposed to follow in average the solution of a $p$-order linear differential equation will be considered in Section 7.

So, assuming that the acoustic vectors are generated by a continuous Markov model as defined by (15) 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 (1991) and Gureg et al. (1990) for LPC analysis. They call it a "combination of instantaneous and transitional LPC frequencies in the parameter domain". They compare this to a "combination in the distance domain", which means that an augmented vector $(x, \dot{x})$ is constructed. Experimental results (Gureg et al., 1990) 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, a combination can be obtained during speech signal analysis and thus it does not result in extra computation at the recognition level (Furui et al., 1990).

4. Parameters estimation

Once the segmentation is fixed, we must find the parameter values that maximize the total probability density (the likelihood) of the observation $P(X, S)$, where $S$ is the sequence of states $(s_0, s_1, \ldots, s_Q)$ together with its segmentation, and $X$ is the time evolution of vector $x$, observed over the whole word. The parameters estimation procedure will be described in the case of training of a single word. Of course, for a whole database, the statistics must be performed on all the data. We have

\[ P(X, S) = P(X|S)P(S). \] 

Since the sequence of states is assumed to be modelled by a discrete Markov process, and since the time evolution of the observations arising from any state in independent, we can write

\[ P(S) = \prod_{k=1}^{Q} \pi(s_k|s_{k-1}) \pi_0, \] 

\[ P(X|S) = \prod_{k=0}^{Q} P_{s_k}[x(t)] \mathcal{S}x(t), \]

where $\pi(s_k|s_{k-1})$ is the transition probability of the discrete Markov model of states, and $P_{s_k}[x(t)]$ is the probability density of the observed acoustic vector path $x(t)$ on state $s$. Maximizing (17) is equivalent to maximizing

\[ \mathcal{L} = \prod_{k=1}^{Q} \pi(s_k|s_{k-1}) \prod_{k=0}^{Q} P_{s_k}[x(t)]. \]

The reestimation formulae for the transition probabilities can be obtained in the same way as for standard one-Gaussian-per-state hidden Markov models.

If we collect the product on all the acoustic vectors labeled as state $s$ in (20) (each state appears a given number of times in the whole word), we can approximate (16) as

\[ P_{s_i}[x(t)] = \frac{1}{\sqrt{(2\pi\Delta t)^d |\Sigma_s|}} e^{-\frac{1}{2} \sum_{i \in s} [\hat{x}_i - A_s x_t - \mu_s]^T} \times (\Sigma_s)^{-1} [\hat{x}_i - A_s x_t - \mu_s] \Delta t, \]

where the sum is taken over all the $N_t$ acoustic vectors $x_t$ that have been labeled as state $s$ during the segmentation procedure, and we have posed $\mu_s = -A_s x_t$. This last expression (21) is similar to the one obtained by Vellekens (1987). In his work, Vellekens assumes explicit dependence between the current vector and the last observed vector. He shows that, in the case of a correlated Gaussian probability density function, the emission probabilities depend on the prediction error of a first order autoregressive model. However, for this model, the generalization to higher order predictors is not obvious.
Cancelling the derivatives of the total probability versus the parameters of the model $\mu_s$ and $A_s$ yields

$$N_s \mu_s + A_s \sum_{i \in s} x_i = \sum_{i \in s} \dot{x}_i,$$

(22a)

$$\mu_s \sum_{i \in s} (x_i)^T + A_s \sum_{i \in s} x_i(x_i)^T = \sum_{i \in s} \dot{x}_i(x_i)^T.$$  

(22b)

Let us introduce the intermediate vectors $\nu_s$ and $\dot{\nu}_s$:

$$\nu_s = \frac{1}{N_s} \sum_{i \in s} x_i,$$  

(23)

$$\dot{\nu}_s = \frac{1}{N_s} \sum_{i \in s} \dot{x}_i.$$  

(24)

If we proceed like in (Wellekens, 1987), we obtain

$$A_s = \left[ \frac{1}{N_s} \sum_{i \in s} (x_i - \nu_s)(x_i - \nu_s)^T \right]$$

$$\times \left[ \frac{1}{N_s} \sum_{i \in s} (x_i - \nu_s)(x_i - \nu_s)^T \right]^{-1},$$  

(25)

$$\mu_s = \dot{\nu}_s - A_s \nu_s,$$  

(26)

and we used the property

$$\frac{1}{N_s} \sum_{i \in s} (\dot{x}_i - \dot{\nu}_s)(x_i - \nu_s)^T$$

$$= \frac{1}{N_s} \sum_{i \in s} (x_i - \nu_s)^T - \dot{\nu}_s(\nu_s)^T.$$

For $\Sigma_s$, we obtain

$$\Sigma_s = \frac{1}{N_s} \sum_{i \in s} \left[ (\dot{x}_i - \dot{\nu}_s) - A_s(x_i - \nu_s) \right]$$

$$\times \left[ (\dot{x}_i - \dot{\nu}_s) - A_s(x_i - \nu_s) \right]^T \Delta t.$$  

(27)

There is something strange with this estimator of the variance-covariance matrix. Indeed, if we look at (27), the estimated values seem to be of order of magnitude $\Delta t$. In fact, this is not surprising, since for a diffusion process such as the one described by (8) or (15), the point undergoes displacements $\Delta x$ which are of the order of magnitude $\sqrt{\Delta t}$ in a small time interval $\Delta t$ (but large values $\Delta t$ compared to the time interval between the small random fluctuations; see for example (Cox and Miller, 1965; Gardiner, 1985)). In other words, we have $(\Delta x)^2 - \Delta t$, and the proportionality constant is called the diffusion coefficient. Now, the term $[(x_i - \nu_s) - A_s(x_i - \nu_s)]$ is simply the difference between the predicted value and the observed value of the acoustic vector, divided by $\Delta t$. It is therefore of order of magnitude $\Delta x/\Delta t + 1/\sqrt{\Delta t}$, so that $\Sigma_s$ is finite.

5. Segmentation procedure (dynamic programming)

As usual, to find the best path over all the possible segmentations of states, we use dynamic programming after sampling of the process (Bellman and Dreyfus, 1962). The best state segmentation, given the observations, is the one that maximizes

$$P(S \mid X) = \frac{P(X \mid S)P(S)}{P(X)},$$

where $S$ is the sequence of states $(s_0, s_1, \ldots, s_Q)$, modelling the uttered word, together with its segmentation. Maximizing this last expression is equivalent to maximizing (20):

$$\mathcal{L}(s_0, s_1, \ldots, s_Q; P_{s_0}, P_{s_1}, \ldots, P_{s_Q})$$

$$= \max \left\{ \prod_{k=1}^{Q} \pi(s_k \mid s_{k-1}) \left( \prod_{k=0}^{Q} P_{s_k}[x(t)] \right) \right\}.$$  

(28)

At a particular time $t = (l \Delta t)$, the process can either come from the same state or transit from another state at $t = ((l-1) \Delta t)$. This leads to the following recurrence:

$$\mathcal{L}(s_k; l \Delta t)$$

$$= \max \left\{ \mathcal{L}(s_k; (l-1) \Delta t) P_{s_k}[x((l-1) \Delta t); \Delta t], \right\}$$

$$\max_{s_{k-1}} \left\{ \mathcal{L}(s_{k-1}; (l-1) \Delta t) \pi(s_k \mid s_{k-1}) \right\} \times P_{s_k}[x((l-1) \Delta t); \Delta t]\}$$  

for $s_k = (s_0, s_1, \ldots, s_Q)$,

(29)

where $\mathcal{L}(s_k; l \Delta t)$ denotes the cumulative probability (28) at time $(l \Delta t)$, while being on state $s_k$. 
\( P_s[x(l\Delta t); \Delta t] \) is the probability density of observing the path \( x(t), t \in [(l-1)\Delta t, l\Delta t] \) on state \( s \) within a constant factor:

\[
P_s[x(l\Delta t); \Delta t] = \frac{1}{\sqrt{\mid \Sigma_x \mid}} \exp\left[-\frac{1}{2} [\dot{x}_{l-1} - A_s x_{l-1} - \mu_s]^T \Sigma_x^{-1} [\dot{x}_{l-1} - A_s x_{l-1} - \mu_s] \Delta t \right]. \tag{30}
\]

In (29), the maximum value is taken on all the possible predecessor states \( s_i \) of state \( s_k \).

During recognition, the likelihood (28) is computed for all the candidate words thanks to (29), and the most likely word is chosen as the best candidate.

6. Duration modeling

At this stage, we are able to include a duration model for each state (Russell and Moore, 1985; Rabiner et al., 1985). Let \( P_s(t_s) \) be the probability density of spending a duration of \( t_s \) on state \( s \). We rewrite (28) as

\[
\mathcal{L}(s_0, s_1, \ldots, s_Q; l_s, t_s, \ldots, t_Q; P_s, \mathcal{P}_s, \mathcal{P}_y, \mathcal{P}_p) = \frac{\prod_{k=1}^Q \pi(s_k | s_{k-1})}{\prod_{k=0}^Q P_s[x(t); \mathcal{P}_s(t_s)]}.
\]

Now, we have to find the best path through all the possible sequences of states \( s_k \) and all the possible durations \((l_s, \Delta t)\). Following Bourlard and Wellekens (1986), we find the following recurrences:

\[
\mathcal{L}(s_k; \Delta t; l_s) = \max_{s_{l-1}} \max_{t_h} \left[ \mathcal{L}(s_{l-1}; l_h, \Delta t; (l-1), \Delta t) \pi(s_{l-1} | s_l) \right] \mathcal{P}_s[x(l\Delta t); \Delta t] \tag{32a}
\]

for \( s_k = (s_0, s_1, \ldots, s_Q) \),

if we enter a state \( s_k \), and

\[
\mathcal{L}(s_k; l_s, \Delta t; l_s) = \mathcal{L}(s_k; (l_s - 1) \Delta t; (l - 1) \Delta t) \times \frac{\mathcal{P}_s[l_s \Delta t]}{\mathcal{P}_s[(l_s - 1) \Delta t]} \times \frac{\mathcal{P}_s[x(l\Delta t); l_s \Delta t]}{\mathcal{P}_s[x((l-1)\Delta t); (l_s - 1) \Delta t]} \tag{32b}
\]

for \( l \geq l_s \geq 2; \ s_k = (s_0, s_1, \ldots, s_Q) \).

if we remain in the same state \( s_k \). \( P_s[x(l\Delta t); l_s \Delta t] \) denotes the probability density of observing the path \( x(t), t \in [(l-1)\Delta t, l\Delta t] \), while being on state \( s \). \( \mathcal{L}(s; l_s \Delta t; l_s \Delta t) \) is the cumulative probability (31) at time \((l_s \Delta t)\), while having spent the last \((l_s \Delta t)\) on state \( s \).

Inverse Gaussian distributions (Sunadeau et al., 1992), Poisson distributions (Bourd and Wellekens, 1986), discrete distributions (Vaseghi and Conner, 1992) as well as gamma distributions (Levinson, 1986) have been used in the literature. Of course, it is interesting to introduce bounded state durations in order to reduce the computation time (Gu et al., 1991).

7. Two examples

Example 1

Let us consider the particular case of a first order totally decoupled process. We have

\[
[\Sigma_s]_{i,j} = [\sigma_s]_i \delta(i-j), \tag{33a}
\]

\[
[A_s]_{i,j} = [\alpha_s]_i \delta(i-j). \tag{33b}
\]

\( \delta(i-j) \) is equal to one if and only if \( i-j = 0 \), and is zero otherwise. \( [\alpha_s]_i \) is component \( i \) of vector \( \alpha_s \). We find for the estimators

\[
[\mu_s]_i = [\nu_s]_i - [\alpha_s]_i [\nu_s]_i, \tag{33c}
\]
\[
\begin{align*}
[\alpha_s]_i &= \frac{1}{N_s} \sum_{k \in s} \left( [\dot{x}_k]_i - [\alpha_s]_i [x_k]_i - [\mu_s]_i \right)^2 \Delta t, \\
[\alpha_s]_i &= \frac{1}{\sum_{k \in s} ( [x_k]_i - [\nu_s]_i )^2 } \sum_{k \in s} ( [x_k]_i - [\nu_s]_i ) \tag{33d}
\end{align*}
\]

Example 2

Let us now assume that, in each state, the acoustic signal \( x(t) \) follows a \( p \)-order linear differential equation in average:

\[
\frac{d^p x(t)}{dt^p} - \sum_{i=0}^{p-1} [\alpha_s]_i \frac{d^i x(t)}{dt^i} = 0. \tag{34}
\]

The differential equation (34) is supposed to be excited by a stochastic source in such a way that the probability density of the corresponding Markov process can be written as in (15).

In this case, we define the vector

\[
\begin{align*}
[x]_0 &= x(t), \\
[x]_1 &= \frac{d}{dt} [x]_0, \\
[x]_2 &= \frac{d}{dt} [x]_1, \\
&\vdots \\
[x]_{p-1} &= \frac{d}{dt} [x]_{p-2}.
\end{align*}
\]

Instead of (34), these \( p \) equations are satisfied exactly. Now, we define the matrix \( A_s \),

\[
[A_s]_{ij} = \delta((i + 1) - j)
\]

for \( i \in [0, p - 2]; \ j \in [0, p - 1] \), \( \tag{35a} \)

\[
[A_s]_{(p-1)j} = [\alpha_s]_j \quad \text{for } i = p - 1, \ j \in [0, p - 1] \), \( \tag{35b} \)

and from (34) we find

\[
\frac{d}{dt} [x]_{p-1} = \sum_{i=0}^{p-1} [A_s]_{(p-1)i} [x]_i \tag{36a}
\]

or

\[
\frac{d}{dt} [x]_{p-1} = (\alpha_s)^T x, \tag{36b}
\]

these last equations being subject to random fluctuations.

By examining the general form (16), we find

\[
\begin{align*}
[\mu_s]_i &= 0 \quad \text{for } i \in [0, p - 2], \tag{37a} \\
[\mu_s]_{p-1} &= [\dot{\nu}_s]_{p-1} - (\alpha_s)^T \nu_s, \tag{37b} \\
[\Sigma_s]_{ij} &= 0 \quad \text{for } i, j \in [0, p - 2], \tag{37c} \\
[\Sigma_s]_{(p-1)(p-1)} &= \frac{1}{N_s} \sum_{k \in s} \left( [x_k]_{p-1} - [\dot{\nu}_s]_{p-1} \right)^2 \Delta t, \tag{37d} \\
(\alpha_s)^T &= \left[ \sum_{k \in s} \left( [x_k]_{p-1} - [\dot{\nu}_s]_{p-1} \right) (x_k - \nu_s)^T \right] \times \left[ \sum_{k \in s} (x_k - \nu_s) (x_k - \nu_s)^T \right]^{-1}. \tag{37e}
\end{align*}
\]

This kind of modelling is similar to those developed by Poritiz (1982), Juang (1984), Juang and Rabíner (1985), Kenny et al. (1990) and Woodland (1992) by using discrete autoregressive Gaussian densities. The model we introduced in this paper can be considered as a continuous-time version of their work.

8. Mixture of paths

We can further assume that each state is described by a mixture of path probability densities, that is,

\[
\mathcal{P}[x(t)] = \sum_j p_j(\omega_j) \mathcal{P}_j[x(t) | \omega_j], \tag{38a}
\]

with

\[
\sum_j p_j(\omega_j) = 1, \tag{38b}
\]

where the conditional density \( \mathcal{P}_j[x(t) | \omega_j] \) is the \( j \)th component density of the mixture. This means that the acoustic vector can choose between sev-
eral trajectories while entering in a state, but cannot skip from one trajectory to the other while being in a particular state. The mixture coefficients \( p_s(\omega_j) \) are related to the a priori probabilities of observing the corresponding path. The problem is now to find the estimators that maximize the cumulative log-probability under constraint (38b). For this purpose, we introduce the Lagrange multipliers \( \lambda_i \):

\[
\mathcal{L} = \sum_{k=1}^{Q} \log[\pi(s_k | s_{k-1})] + \sum_{k=0}^{Q} \log\left[ \sum_j p_s(\omega_j) \mathcal{P}_s[x(t) | \omega_j] \right] + \sum_{k=0}^{Q} \log[\mathcal{P}_s(t_{s_k})] + \sum_{s=0}^{Q} \lambda_i \left( \sum_j p_s(\omega_j) - 1 \right). \tag{39}
\]

Let us define

\[
\mathcal{P}_s(\omega_j | x(t)) = \frac{\mathcal{P}_s[x(t) | \omega_j] p_s(\omega_j)}{\mathcal{P}_s[x(t)]}.
\]

By cancelling the derivative of (39) versus the parameters, we find

\[
p_s(\omega_j) = \frac{1}{n_s} \sum_{\sigma = s} \mathcal{P}_s(\omega_j | x(t)), \tag{40a}
\]

\[
\nu_{s;i,j} = \frac{\sum_{\sigma = s} [\mathcal{P}_s(\omega_j | x(t)) \sum_{i \in \sigma} x_i]}{\sum_{\sigma = s} \mathcal{P}_s(\omega_j | x(t))}, \tag{40b}
\]

\[
\dot{\nu}_{s;i,j} = \frac{\sum_{\sigma = s} [\mathcal{P}_s(\omega_j | x(t)) \sum_{i \in \sigma} \dot{x}_i]}{\sum_{\sigma = s} \mathcal{P}_s(\omega_j | x(t))}, \tag{40c}
\]

\[
A_{s;i,j} = \left[ \sum_{\sigma = s} \mathcal{P}_s(\omega_j | x(t)) \sum_{i \in \sigma} (\dot{x}_i - \nu_i)(x_i - \nu_i)^T \right]^{\frac{1}{2}} \times \left[ \sum_{\sigma = s} \mathcal{P}_s(\omega_j | x(t)) \sum_{i \in \sigma} (x_i - \nu_i)(x_i - \nu_i)^T \right]^{\frac{1}{2}} \tag{40d}
\]

\[
\sum_{s = 1} \frac{\mathcal{P}_s(\omega_j | x(t))}{\sum_{\sigma = s} \mathcal{P}_s(\omega_j | x(t))} \times \left[ \sum_{i \in \sigma} (x_i - \nu_i)(x_i - \nu_i)^T \right]^{-1}, \tag{40e}
\]

The summation on \( \sigma = s \) must be understood as the sum over all the occurrences \( \sigma \) of state \( s \) in the word, while the summation on \( i \in \sigma \) is taken on all the acoustic vectors that are segmented in \( \sigma \). \( n_s \) is the number of occurrences of state \( s \) in the word and \( N_\sigma \) is the number of acoustic vectors segmented in \( \sigma \):

\[
n_s = \sum_{\sigma = s} 1, \quad N_\sigma = \sum_{i \in \sigma} 1.
\]

The segmentation is found by applying a dynamic programming procedure analogous to (32a, 32b), since the length of the path must be known in order to compute the probability densities \( \mathcal{P}_s[x(t)] \) thanks to (38a). Moreover, since the equations are coupled and non-linear, they have to be solved iteratively (Duda and Hart, 1973); the EM algorithm is generally used.

9. Conclusion

In this work, we have considered that, on each state, the acoustic vectors are generated by a continuous Markov model. This is to be opposed to the current assumption, i.e. that acoustic vectors are emitted according to a probability density 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. By considering a particular continuous Markov process, we have derived the probability density of observing this trajectory, given the state. Roughly speaking, it measures the "adequacy" of the observed trajectory with respect to an ideal trajectory, which is generated by a vectorial linear differential equation. Once the segmentation is fixed, reestimation formulae for the parameters of the continuous Markov process are derived for the Viterbi algorithm. 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. Moreover, we have shown that duration models could be 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 (Bourlard and Wellekens, 1986). Thereafter, we consider two special cases, i.e., a totally uncorrelated process and a modelling of the speech signal by a $p$-order differential equation. This provides some enlightenments to related work of Poritz (1982), Juang (1984), Juang and Rabiner (1985) (see also Kenny et al., 1990; Tishby, 1991; Woodland, 1992), Furui (1986, 1991) and Gurger et al. (1990). Finally, we sketch a possible generalization to path mixtures, for which different trajectories are available in each state.

Further work will be devoted to the generalization to other continuous Markov processes. In particular, we computed the path probability density in the case of a vectorial linear stochastic differential equation (Saerens, 1993). It appears that the solutions of such an equation follow a multivariate Ornstein–Uhlenbeck Markov process, and are distributed according to a Gaussian density (Gardiner, 1985).

Finally, we have to stress that no experimental results are available at present. Indeed, we did not have the opportunity to test the algorithm on real speech. We are aware of the fact that the assumptions we did may not be appropriate for the modelling of speech. An experimental validation of the model is now necessary.

Acknowledgments

I thank Professor Christian Wellekens and Alain Soquet for their comments, and for the many useful discussions we had. I am also grateful to the two anonymous referees for their comments on a earlier version of the manuscript.

References


M. Saerens (1993), "Hidden Markov models assuming acoustic vectors generated by a linear differential equation on each state", Submitted for publication.


