Stochastic Calculus, Non-Linear Filtering, and the Internal Model Principle: Implications for Articulatory Speech Recognition

Gordon Ramsay

Institut de la Communication Parlée
UPRESA CNRS 5009 - INPG
46 avenue Félix Viallet
38031 Grenoble CEDEX 01
France.

ABSTRACT

A stochastic approach to modelling speech production and perception is discussed, based on Itô calculus. Speech is modelled by a system of non-linear stochastic differential equations evolving on a finite-dimensional state space, representing a partially-observed Markov process. The optimal non-linear filtering equations for the model are stated, and shown to exhibit a predictor-corrector structure, which mimics the structure of the original system. This is used to suggest a possible justification for the hypothesis that speakers and listeners make use of an "internal model" in producing and perceiving speech, and leads to a useful statistical framework for articulatory speech recognition.

1. INTRODUCTION

Traditional models for speech recognition are based on representing phonological sequences and acoustic feature trajectories by simple forms of hidden Markov process, with various patterns of statistical dependency introduced between a hidden discrete state space and an observed continuous measurement space.

A common criticism of such models is that there is little resemblance between the underlying model structure and the physical processes involved in producing real speech. Consequently, it has often been proposed that the introduction of true physical models of speech production might provide a useful means of constraining speech recognition, a more robust parameterisation of speech, and better possibilities for interpreting recognition results.

Conversely, traditional models of speech production typically attempt to describe articulatory and acoustic data by proposing deterministic relationships between trajectories of abstract control variables and trajectories of state variables representing the physical state of the vocal tract and auditory system.

These models often succeed in generating realistic state trajectories from a direct statement of basic physical principles, but cannot reproduce the statistical patterns of variability observed in real measurement data. Previous papers have suggested that this deficiency could be addressed if speech production models were represented in a statistical framework [1] [2].

Common to both speech production and speech recognition is the problem of *describing* how speakers systematically vary their control strategies in different contexts, and *explaining* how listeners succeed in recovering linguistic information from speech, where the consequences of this variability are apparent.

Many theories of speech production and perception claim that speakers and listeners possess considerable knowledge of the way their vocal tracts behave, and actively employ an "internal model" of articulatory behaviour to regulate and interpret articulatory-acoustic variability in different environments. Despite the appearance of numerous proposals of this type, the arguments justifying the hypothesis remain largely heuristic, and no quantitative or falsifiable mathematical model has ever been formulated or tested.

The aim of this paper is to demonstrate that speech production and perception can be modelled within a very general statistical framework, using stochastic calculus, and to show that procedures for speech synthesis and recognition then follow naturally from classical results in the theory of non-linear filtering. Prior knowledge of articulatory behaviour is assumed to define a probability measure on a function space of physical state trajectories; speech production and perception then involve recursive construction of a conditional probability measure on the same state space by integrating partial sensory measurement data. The contribution of the paper lies in explaining that the concept of an "internal model" is directly reflected in the structure of the optimal filter, as provided by the two main theorems in this field, which have an intuitive and appealing interpretation for speech.

The paper assumes a basic familiarity with stochastic calculus, accounts of which may be found in the references provided [3] [4]. The non-linear filtering results stated in the paper are evidently not original, and have been adapted from [4] [5] [6].

2. HIDDEN MARKOV PROCESSES

Before abstract modelling issues can be addressed, a general mathematical representation is needed that does not overly rely on the details of any specific model. Here the essential problem lies in linking statistical properties of observed measurement data to physical laws governing underlying state trajectories, and a stochastic framework is therefore appropriate.

Assume a complete underlying probability space $(\Omega, \mathcal{F}, P)$ throughout. Let $X = \{X_t : t \in \mathbb{R}_+\}$ be a stochastic process representing the physical state of the vocal tract, taking values in a state space $(\mathcal{S}_X, \mathcal{B}(\mathcal{S}_X))$, and let $Y = \{Y_t : t \in \mathbb{R}_+\}$ be a stochastic process representing partial measurements of the state, taking values in a state space $(\mathcal{S}_Y, \mathcal{B}(\mathcal{S}_Y))$. Define $\mathcal{F}^X = \{\mathcal{F}_t^X : t \in \mathbb{R}_+\}$ and $\mathcal{F}^Y = \{\mathcal{F}_t^Y : t \in \mathbb{R}_+\}$ to be the right-continuous

filtrations generated by X and Y respectively.

Although the evolution of the state of any physical system involved in speech production must usually be modelled by partial differential equations defined on a function space of infinite dimension, most of the important phenomena in speech arise from vibratory systems whose response can be represented by localized eigenmode expansions. By including only the dominant eigenmodes, or by making use of standard numerical simulation techniques, adequate finite-dimensional representations of the underlying physics can be constructed, and little generality is lost by restricting the model structure to a system of non-linear differential equations evolving on a finite-dimensional state space; $\mathcal{S}_X$ and $\mathcal{S}_Y$ can therefore be assumed to be Euclidean vector spaces.

Furthermore, it is sensible to assume that the physical system modelled by the state process is causal, and therefore that the future of the system is independent of its past, given the present state. In a statistical framework, this immediately implies that the state process X must be a Markov process, and if it is assumed that only the measurement process Y can be observed, then X and Y together define a general continuous-time *hidden Markov model*. Under certain technical conditions, it is always possible to represent X and Y as solutions of random integral equations,

$$X_t = X_0 + \int_0^t g(X_s, s)ds + \int_0^t v(X_s, s)dV_s, \quad (1)$$

$$Y_t = Y_0 + \int_0^t h(X_s, s)ds + \int_0^t w(s)dW_s, \quad (2)$$

which are usually written as stochastic differential equations,

$$dX_t = g(X_t, t)dt + v(X_t, t)dV_t, \quad (3)$$

$$dY_t = h(X_t, t)dt + w(t)dW_t, \quad (4)$$

where $V = \{V_t : t \in \mathbb{R}_+\}$ and $W = \{W_t : t \in \mathbb{R}_+\}$ are independent Wiener processes, independent of X_0 and Y_0, and g, h, v, w are appropriate measurable functions. Remark that the definition of the stochastic integrals in (1) and (2) necessarily involves the use of martingale calculus; previous attempts to define continuous-time models of speech appear to overlook this [7].

The stochastic differential equations implicitly define both the sample-path properties of X, Y and their joint probability law. The functions g and h essentially determine the form of the state and measurement trajectories, and can be chosen to constrain the sample paths of X and Y to follow physically-realistic patterns. The functions v and w determine how randomness enters into the system, and can be chosen to reflect the systematic variability that affects the physical evolution of each component of the system state. In order to interpret these equations as a statistical model of speech production, therefore, a suitable state space must be chosen and functions g, h, v, w selected to reflect prior knowledge about the deterministic physics of the vocal tract and the random intentional variability underlying speech motor control. Remark that the model structure is general enough to encompass both standard HMMs and articulatory or acoustic models as special cases; all that is required is a basic state-variable description.

Once the model structure has been defined, the sample paths and statistical properties of the state process can be calculated, and these can be used to examine the behaviour of the model. The essential tool is the functional described in the following definition.

Definition 1
Let p_t be the functional defined by the expectation

$$p_t(\phi) := E\{\phi(X_t)\}, \quad (5)$$

where ϕ is any suitably-regular measurable function on $\mathcal{S}_X$.

Using p_t, the probability law of the process and all of its moments can evidently be obtained by substituting appropriate functions for ϕ. The usefulness of p_t centres on the existence of the recursive representation stated in the theorem below.

Theorem 1
The functional p_t is generated by the integral recursion

$$p_t(\phi) = p_0(\phi) + \int_0^t p_s(L\phi)ds, \quad (6)$$

which can be written as a stochastic differential equation

$$dp_t(\phi) = p_t(L\phi)dt, \quad (7)$$

where L is the operator defined by

$$L\phi = \sum_i g_i \frac{\partial \phi}{\partial x_i} - \frac{1}{2}\sum_{i,j} v_i \frac{\partial^2 \phi}{\partial x_i \partial x_j} v_j. \quad (8)$$

The operator L is the *extended generator* of the Markov process, and describes how the probability mass is transported along the sample paths. Equation (7) describes the evolution of the *unconditional* or *a priori* probability law of the state process for any choice of g,h,v,w, and can be thought of as the "forward model" recursively characterising the statistical dynamics of the system.

The model description is now complete, and can be used for *stochastic articulatory speech synthesis*, by generating Monte Carlo simulations of the state trajectories that arise from solution of equations (3) and (4), and by calculating the prior probability distribution of the state and measurement processes using (7).

3. NON-LINEAR FILTERING

In order to use the model for *stochastic articulatory speech recognition*, procedures must be derived to recover optimal estimates of the physical state from partial or incomplete observations.

The basic definition of the model structure generates the unconditional probability law P_{X_t} for the state process X, for any particular choice of state space and functions g,h,v,w, and this embodies all of the prior knowledge about articulatory behaviour present in the model. If the state process cannot be observed, and no measurements are provided, then the optimal estimate of the hidden state trajectory X_t is given by the unconditional mean $E\{X_t\}$.

When partial observations of the state process are available through the measurement process Y, it can be shown that the optimal (minimum variance) estimate of the hidden state trajectory X_t is provided by the conditional mean $E\{X_t|\mathcal{F}_t^Y\}$. More generally, all of the information supplied by the measurements is embodied in the shape of the conditional probability law $P_{X_t|\mathcal{F}_t^Y}$.

Moreover, since the state and measurement processes are generated recursively in time, it is of considerable interest to derive

recursive formulae for estimates of any function of the state. The solution of the state estimation problem for systems modelled by non-linear stochastic differential equations is provided by two key results, termed the *Kushner-Stratonovich* and *Zakai* filters.

3.1. The Semi-Martingale Approach

The original approach to non-linear filtering was based on a generalisation of the innovations method used to derive the Kalman filter, and is based on constructing the following functional:

Definition 2
Let π_t be the functional defined by the conditional expectation

$$\pi_t(\phi) \quad := \quad E\{\phi(X_t)|\mathcal{F}_t^Y\}, \tag{9}$$

where ϕ is any suitably-regular measurable function on $\mathcal{S}_X$.

Using π_t, the conditional probability law of the process can be obtained. The central result of the *semi-martingale approach* to non-linear filtering is the recursive representation stated below;

Theorem 2 *(Fujisaki-Kallianpur-Kunita) [5]*
The functional π_t is generated by the integral recursion

$$\pi_t(\phi) \quad = \quad \pi_0(\phi) + \int_0^t \pi_s(L\phi)ds + \int_0^t \sigma_s(h,\phi)d\nu_s, \tag{10}$$

which can be written as a stochastic differential equation

$$d\pi_t(\phi) \quad = \quad \pi_t(L\phi)dt + \sigma_t(h,\phi)d\nu_t, \tag{11}$$

where σ_t is the conditional covariance matrix defined by

$$\sigma_t(h,\phi) \quad := \quad \pi_t(h\phi) - \pi_t(h)\pi_t(\phi), \tag{12}$$

and ν_t is the innovations process defined by

$$\nu_t \quad = \quad Y_t - \int_0^t \pi_s(h)ds. \tag{13}$$

The solution of the filtering problem thus consists of a stochastic differential equation evolving on the hidden state space, and this equation has an interesting structure. Examining the *Kushner-Stratonovich filter* (11), the first term reproduces a conditional version of the "forward model" in equation (7) describing the true state process, and predicts the way that the system is believed to evolve given prior knowledge. The innovations process ν_t defines the error between the observed measurement trajectory and the expected measurement trajectory predicted from the internal state of the filter, whereas the conditional covariance σ_t measures the expected size of the discrepancy between observation and prediction. The second term in equation (11) corrects the prediction provided by the forward model, by adjusting the filter state by an amount proportional to the measurement error, weighted by an estimate of how large the model expects this error to be.

Interpreting this result for speech, the non-linear filter implements an intuitive and logical "predictor-corrector" structure, which is based on using an "internal model" of articulatory dynamics to propagate the state estimate, corrected by the perceived observation error calculated recursively from the measurements.

3.2. The Measure-Change Approach

An alternative approach to non-linear filtering is based on transforming the original probability measure into a new measure, under which the state and measurement processes are independent.

Theorem 3 *(Cameron-Martin-Girsanov)*
There exists a measure $\overline{P}$ on $(\Omega,\mathcal{F})$, absolutely continuous w.r.t. P, such that X and Y are independent under $\overline{P}$, Y is a Wiener process under $\overline{P}$, and $\overline{P}$ coincides with P on $\mathcal{F}^X$.

Defining $\Lambda_t := \overline{E}\{dP/d\overline{P}|\mathcal{F}_t^Y\}$, construct another functional;

Definition 3
Let $\overline{\pi}_t$ be the functional defined by the conditional expectation

$$\overline{\pi}_t(\phi) \quad := \quad \overline{E}\{\phi(X_t)\Lambda_t|\mathcal{F}_t^Y\}, \tag{14}$$

where ϕ is any suitably-regular measurable function on $\mathcal{S}_X$.

A simple relationship can be shown to exist between $\overline{\pi}_t$ and π_t;

Theorem 4 *(Kallianpur-Striebel)*
For any suitably-regular measurable function ϕ on $\mathcal{S}_X$,

$$\pi_t(\phi) \quad = \quad \overline{\pi}_t(\phi)/\overline{\pi}_t(1). \tag{15}$$

Using $\overline{\pi}_t$, the "unnormalized" conditional probability law of the process can be obtained, and by applying Theorem 4 this immediately provides the conditional probability law. The central result of the *measure-change approach* to non-linear filtering is the recursive representation stated in the following theorem;

Theorem 5 *(Duncan-Mortensen-Zakai) [6]*
The functional $\overline{\pi}_t$ is generated by the integral recursion

$$\overline{\pi}_t(\phi) \quad = \quad \overline{\pi}_0(\phi) + \int_0^t \overline{\pi}_s(L\phi)ds + \int_0^t \overline{\sigma}_t(h,\phi)dY_s \tag{16}$$

which can be written as a stochastic differential equation

$$d\overline{\pi}(\phi) \quad = \quad \overline{\pi}_t(L\phi)dt + \overline{\sigma}_t(h,\phi)dY_t, \tag{17}$$

where $\overline{\sigma}_t$ is the conditional correlation matrix defined by

$$\overline{\sigma}_t(h,\phi) \quad := \quad \overline{\pi}_t(h\phi). \tag{18}$$

Once again, the solution of the filtering problem consists of a stochastic differential equation evolving on the hidden state space, but now under a different probability measure. Examining the *Zakai filter* (17), the first term employs a conditional version of the "forward model" defined in equation (7) to predict the evolution of the system state, as before. The conditional correlation $\overline{\sigma}_t$ measures the expected agreement between the observed measurement trajectory and the expected measurement trajectory predicted from the system state. The second term in equation (17) corrects the prediction provided by the forward model, by adjusting the filter state according to the correlation between the true observation and the filter prediction.

Interpreting this result for speech, the non-linear filter again implements a "predictor-corrector" structure, based on using an "internal model" of articulatory dynamics to propagate the state estimate, but this time the state estimate is corrected according to the perceived correlation with the measurements.

3.3. The Internal Model Hypothesis

Stochastic filtering theory provides two alternative solutions to the mathematical problem of estimating the conditional probability law of a hidden state process, governed by non-linear stochastic differential equations, from partial measurements. Both of these solutions are themselves non-linear stochastic differential equations, and both possess a simple and intuitive "predictor-corrector" structure. The basic structure of the optimal non-linear filter has been shown to centre inevitably on an "internal model" of the original system, and this is used by the filter to predict the evolution of the system state. The Kushner-Stratonovich filter continuously corrects the prediction by monitoring the covariance of the error between the observed measurement and the measurement predicted from the filter state. The Zakai filter also continuously corrects the prediction, but monitors instead the correlation between observed and predicted measurements.

If the model structure, which is extremely general, is chosen to provide a faithful representation of the physics of the vocal tract, then the logical conclusion must be that the optimal means of recovering an appropriate underlying articulatory state or motor command from sensory measurements necessarily involves implementing an "internal model" of the system dynamics, using the predictions of this model to correct the evolution of the state estimate. This provides a concrete and rigorous justification for many of the basic ideas underlying current theories of speech production and perception (cf. [8] – [15] and references therein).

It is important to realise that the filtering equations do not simply provide optimal estimates for particular functions of the state process, but implicitly define the evolution of the entire conditional probability law of the articulatory state. Many formulations of the internal model hypothesis assume that production or perception systems recover a single optimal trajectory. Within a statistical framework, the "objects of production" and "objects of perception" are conditional probability measures defined over the entire articulatory state space, completely characterising the uncertainty in the articulatory state and the information in any measurements. During speech production, speakers begin with a prior probability distribution of acceptable control signals, and update this distribution by recursively conditioning on sensory feedback, randomly sampling the result to generate a control strategy. During speech perception, listeners begin with a prior probability distribution of perceived articulatory states, and update this distribution by recursively conditioning on observations of the speaker's behaviour.

4. CONCLUSIONS

A statistical framework for modelling speech production and perception has been outlined, based on representing speech by non-linear stochastic differential equations describing a general continuous-time hidden Markov process. The conditional probability law of the process can be determined using the Kushner-Stratonovich or Zakai filters, and the structure of the non-linear filtering equations is consistent with many heuristic theories of speech perception and motor control. In particular, it has been suggested that non-linear filtering provides a useful mathematical justification for the proposal that speakers and listeners make use of an "internal model" in producing and perceiving speech.

5. REFERENCES

1. G. Ramsay and L. Deng. A stochastic framework for articulatory speech recognition. *Journal of the Acoustical Society of America*, 95(5):Pt.2, 2873 (2aSP19), 1994 (abstract).
2. G. Ramsay. A non-linear filtering approach to stochastic training of the articulatory-acoustic mapping using the EM algorithm. In *Proceedings, ICSLP-96*, volume 2, pages 514–517, Philadelphia, U.S.A., 1996.
3. I. Karatzas and S. E. Shreve. *Brownian Motion and Stochastic Calculus*. Springer-Verlag, 1991.
4. E. Wong and B. Hajek. *Stochastic Processes in Engineering Systems*. Springer-Verlag, 1985.
5. M. Fujisaki, G. Kallianpur, and H. Kunita. Stochastic differential equations for the non linear filtering problem. *Osaka Journal of Mathematics*, 9:19–40, 1972.
6. M. Zakai. On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11:230–243, 1969.
7. M. Saerens. A continuous-time dynamic formulation of the Viterbi algorithm for one-gaussian-per-state hidden Markov models. *Speech Communication*, 12:321–333, 1993.
8. M. Halle and K. N. Stevens. Analysis by synthesis. In W. Wathen-Dunn and L. Woods, editors, *Proceedings, Seminar on Speech Compression and Processing*, 1959.
9. B. Lindblom and J. Lubker. Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics*, 7:147–161, 1979.
10. G. Bailly, R. Laboissière, and J. L. Schwartz. Formant trajectories as audible gestures: an alternative for speech synthesis. *Journal of Phonetics*, 19:9–23, 1991.
11. J. S. Perkell, M. L. Matthies, M. Svirsky, and M. I. Jordan. Goal-based speech motor control : a theoretical framework and some preliminary data. *Journal of Phonetics*, 23:23–35, 1995.
12. J. S. Perkell, M. L. Matthies, H. Lane, F. Guenther, R. Wilhelms-Tricarico, and P. Guiod. Speech motor control: acoustic goals, saturation effects, auditory feedback, and internal models. *Speech Communication*, 22:227–250, 1997.
13. M. I. Jordan and D. E. Rumelhart. Forward models : supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.
14. F. H. Guenther. Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, 102(3):594–621, 1995.
15. V. N. Sorokin. The concept of internal model in speech production and speech perception. In *Proceedings of the 1st ETRW on Speech Production Modelling*, pages 129–132, Autrans, France, 1996.

Acknowledgement: The research described in this paper was funded by a Marie Curie Research Fellowship awarded to the author under the Training and Mobility of Researchers Programme of the European Commission, and forms part of a project entitled "Stochastic Modelling of Speech Motor Control," currently carried out at ICP-INPG in collaboration with Rafael Laboissière.