A Non-linear Filtering Approach to Stochastic Training of the Articulatory-Acoustic Mapping Using the EM Algorithm

Gordon Ramsay

Department of Electrical & Computer Engineering
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1.

ABSTRACT

Current techniques for training representations of the articulatory-acoustic mapping from data rely on artificial simulations to provide codebooks of articulatory and acoustic measurements, which are then modelled by simple functional approximations. This paper outlines a stochastic framework for adapting an artificial model to real speech from acoustic measurements alone, using the EM algorithm. It is shown that parameter and state estimation problems for articulatory-acoustic inversion can be solved by adopting a statistical approach based on non-linear filtering.

1. Introduction

Representation of the non-linear mapping between articulatory parameters and acoustic measurements is a long-standing problem in speech research, and is of particular importance for articulatory approaches to speech recognition and synthesis. Analytic solutions are not available, and functional approximations must be derived instead from a set of examples. Ideally, a corpus of training data consisting of articulatory and acoustic measurements is needed, in which case a model can be obtained using standard data-fitting techniques; in practice, however, there is considerable difficulty in obtaining a sufficient quantity of articulatory measurements. Training from partial acoustic data alone is very difficult without precise prior knowledge about the behaviour of the hidden articulators.

Most current approaches rely on artificial simulations to produce large sample sets from articulatory and acoustic models; the results are then used to train a deterministic model of the complete data. A survey of previous work can be found in [1]. However, the performance of current simulations is poor, and the data generated do not usually provide a good match to the characteristics of real speech. What is needed, therefore, is a means of adapting an existing artificial model of the articulatory-acoustic mapping to real speech, using only acoustic measurements.

This paper proposes a general framework for training a particular class of approximations from measurement data, using a stochastic model to supply prior information about dynamic constraints implicit in the behaviour of the unobserved vocal tract. The EM algorithm is used to provide estimates of the model parameters, and it is shown that the problem essentially reduces to an exercise in non-linear filtering. This also provides the basis for a statistical formulation of the inversion problem in articulatory speech recognition.

2. Problem Formulation

Assume a complete probability space $(\Omega, \mathcal{F}, P)$ throughout. Let $X = \{X_n : n \in \mathbb{N}\}$ be a state process representing articulatory trajectories, taking values in a state space $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, and let $Y = \{Y_n : n \in \mathbb{N}\}$ be a measurement process representing acoustic data, taking values in an observation space $(\mathbb{R}^q, \mathcal{B}(\mathbb{R}^q))$. Define $\mathcal{F}^X = \{\mathcal{F}^X_n : n \in \mathbb{N}\}$, $\mathcal{F}^Y = \{\mathcal{F}^Y_n : n \in \mathbb{N}\}$ to be the filtrations generated by X and Y, where $\mathcal{F}^X_n = \sigma(X_1 \dots X_n)$, $\mathcal{F}^Y_n = \sigma(Y_1 \dots Y_n)$, and let $\mathcal{G} = \{\mathcal{G}_n : n \in \mathbb{N}\}$, where $\mathcal{G}_n = \mathcal{F}^X_n V \mathcal{F}^Y_n$.

Suppose that Y is generated from X by a Borel-measurable function $h : \mathbb{R}^p \to \mathbb{R}^q$ representing the articulatory-acoustic mapping, according to

$$Y_n = h(X_n; \theta) + W_n,$$

where θ are parameters taking values in some measurable space $(\Theta, \mathcal{B}(\Theta))$, and $W = \{W_n : n \in \mathbb{N}\}$ is a random process representing modelling errors, assumed to be Gaussian i.i.d. and independent of X, with $W_n \sim N(0, \Sigma_w)$ for some positive-definite covariance matrix $\Sigma_w \in \mathbb{R}^{q \times q}$.

In the absence of a convenient analytic expression for $h(\cdot)$, it is useful in practice to consider local approximations built up from a combination of elementary functions. The main topic of this paper is the general structure given by

$$h(x; \theta) \approx \sum_{k=1}^{K} w_k(x) \, (h_k + H_k x) \quad (\forall x \in \mathbb{R}^p)$$

where $\theta = \{(h_k \in \mathbb{R}^q, H_k \in \mathbb{R}^{q \times p}) : k = 1 \dots K\}$ describes a family of affine functions mapping $\mathbb{R}^p$ into $\mathbb{R}^q$. The predefined weighting functions $w_k : \mathbb{R}^p \to \mathbb{R}$ act as interpolators between the different component mappings, where each

w_k is chosen to restrict the influence of (h_k, H_k) to a limited sub-domain of $\mathbb{R}^p$. For example, if the input space is partitioned into disjoint subsets $\{A_k : k = 1 \ldots K\}$, and $w_k = I_{A_k}$, we have a piecewise-affine representation for $h(\cdot)$. Alternatively, if the w_k form a class of radial basis functions, $h(\cdot)$ describes a smooth interpolation between local affine mappings centred upon particular points in the input space. A linear representation for $h(\cdot)$ can be obtained by defining

$$\begin{aligned} \overline{X} &\triangleq [w_1(X)\ w_1(X)X^T\ \ldots\ w_K(X)\ w_K(X)X^T]^T, \\ \overline{H}_\theta &\triangleq [h_1\ H_1\ \ldots\ h_K\ H_K], \end{aligned}$$

in which case the measurement equation becomes

$$Y_n = \overline{H}_\theta \overline{X}_n + W_n.$$

This provides a basic model for the relationship between the articulatory process X and the observation process Y, and is closely related to a number of previous proposals. Stokbro et al. [2] and Gorinevsky [3] analyse deterministic versions of a similar model where X and Y are known completely during training; Jordan and Xu [4] develop a "mixture of experts" model where X and Y are observed, but the relationship between them is modulated by an additional fixed set of unobserved random variables; Bengio and Frasconi [5] propose a model similar to Jordan's, where the hidden variables follow a Markov chain evolving on a discrete state-space. The framework proposed here can be seen as a parallel result for the case where the input X is unobserved, and the weight functions $w_k(\cdot)$ determining the mapping $h(\cdot)$ between X and Y are controlled by a hidden process propagating in time on a continuous state-space.

Two problems are of immediate interest. Firstly, estimates of the unknown model parameters θ must be derived from real speech data (parameter estimation), using measurements of the Y process in conjunction with prior information about the statistics of the unobserved process X. Secondly, the optimized model must be used to recover information about unseen articulatory trajectories from any available measurements (state estimation). Solutions to both problems can be obtained using results from non-linear filtering.

3. Parameter Estimation

Suppose that a corpus of training data $\mathcal{O} = \{\mathcal{O}_n : n = 1 \ldots N_o\}$ is available, where each $\mathcal{O}_n = \{y_k^n : k = 1 \ldots N(n)\}$ describes a single observed sample path of the Y process, and the $\mathcal{O}_n$ are generated from independent trials. The X process is not observed, but it will be assumed that the specification of the probability space completely describes the possible trajectories of X in advance, as well as their distribution.

Each parameter value $\theta \in \Theta$ induces a corresponding probability measure P_θ on the underlying measurable space $(\Omega, \mathcal{F})$. Suppose that the measures $\{P, P_\theta : \theta \in \Theta\}$ are mutually absolutely continuous; then the likelihood $L(\theta)$ of the observations $\mathcal{O}$ under each measure P_θ relative to the reference measure P is well-defined, and given by

$$L(\theta) = E\left\{ \frac{dP_\theta}{dP} \,\middle|\, \mathcal{O} \right\}.$$

Adopting a maximum-likelihood approach, we wish to find an optimum value $\theta^\star \in \Theta$ maximizing $L(\theta)$. In general, this is a difficult problem, and it may not be possible to obtain a closed-form expression for $\theta^\star$. The *EM algorithm* [6] provides an alternative approach, whereby a sequence of estimates $\hat{\theta} = \{\hat{\theta}_m : m = 1, 2, \ldots\}$ is constructed recursively, by choosing $\hat{\theta}_{m+1}$ to maximize an auxiliary Q-function $Q(\theta, \hat{\theta}_m)$ [7] [8];

$$\begin{aligned} Q(\theta_1, \theta_2) &\triangleq E_{\theta_2}\left\{ \ln \frac{dP_{\theta_1}}{dP_{\theta_2}} \,\middle|\, \mathcal{O} \right\}, \\ \hat{\theta}_{m+1} &= \arg\max_{\theta \in \Theta} Q(\theta, \hat{\theta}_m). \end{aligned}$$

Under suitable regularity conditions on $Q(\cdot, \cdot)$, the sequence $\hat{\theta}$ can be shown to converge to a stationary point of the original likelihood function, and maximization of the Q-function at each iteration is usually easier than direct maximization of $L(\cdot)$. Applying the EM algorithm to the model described above, and ignoring constants, the Q-function is given by

$$\begin{aligned} Q(\theta_1, \theta_2) = E_{\theta_2}\{ \sum_{n=1}^{N_o} \sum_{k=1}^{N(n)} ([y_k^n - \overline{H}_{\theta_2} \overline{X}_k^n]^T \Sigma_w^{-1} [y_k^n - \overline{H}_{\theta_2} \overline{X}_k^n] \\ - [y_k^n - \overline{H}_{\theta_1} \overline{X}_k^n]^T \Sigma_w^{-1} [y_k^n - \overline{H}_{\theta_1} \overline{X}_k^n]) | \mathcal{O} \}. \end{aligned}$$

Neglecting terms not involving θ_1, this can be re-written as

$$Q(\theta_1, \theta_2) = \text{tr}\ \Sigma_w^{-1} (\overline{H}_{\theta_1} \mathcal{Z}_{\theta_2}^{b\ T} + \mathcal{Z}_{\theta_2}^b \overline{H}_{\theta_1}^T - \overline{H}_{\theta_1} \mathcal{Z}_{\theta_2}^a \overline{H}_{\theta_1}^T) + K(\theta_2)$$

where $K(\theta_2)$ is a function of θ_2 alone, and we define

$$\begin{aligned} \mathcal{Z}_\theta^a &\triangleq \sum_{n=1}^{N_o} \sum_{k=1}^{N(n)} E_\theta\{ \overline{X}_k^n \overline{X}_k^{n\ T} | \mathcal{O}_n \}, \\ \mathcal{Z}_\theta^b &\triangleq \sum_{n=1}^{N_o} \sum_{k=1}^{N(n)} y_k^n E_\theta\{ \overline{X}_k^n | \mathcal{O}_n \}^T. \end{aligned}$$

The first term in $Q(\theta, \hat{\theta}_m)$ contains a negative semi-definite quadratic form in H_θ, and can be maximized using standard optimization techniques, provided that $\mathcal{Z}_\theta^a$ is non-singular. Conditions ensuring that the problem is well-posed will not be discussed here. Differentiating $Q(\theta, \hat{\theta}_m)$ w.r.t. each parameter in $\overline{H}_\theta$ and setting the result to zero under this assumption gives the new estimate $\overline{H}_{\hat{\theta}_{m+1}}$ as a (regularized) solution of the following linear system of equations:

$$\overline{H}_{\hat{\theta}_{m+1}} \mathcal{Z}_{\hat{\theta}_m}^a = \mathcal{Z}_{\hat{\theta}_m}^b$$

Here $\mathcal{Z}_\theta^a$ and $\mathcal{Z}_\theta^b$ constitute the sufficient statistics for the parameter estimation problem, and involve evaluating the conditional expectation of certain functionals of the hidden X process from the observed data. More specifically, re-examining the structure of $h(\cdot)$, we require the following

quantities for each time step k and all $i, j = 1 \ldots K$:

$$E_\theta\{w_i(X_k^n)|\mathcal{O}_n\},$$
$$E_\theta\{w_i(X_k^n)X_k^n|\mathcal{O}_n\},$$
$$E_\theta\{w_i(X_k^n)w_j(X_k^n)|\mathcal{O}_n\},$$
$$E_\theta\{w_i(X_k^n)w_j(X_k^n)X_k^n|\mathcal{O}_n\},$$
$$E_\theta\{w_i(X_k^n)w_j(X_k^n)X_k^n X_k^{n\,T}|\mathcal{O}_n\}.$$

All of these are of the form $E_\theta\{\phi(X_k)|\mathcal{O}_n\}$, where ϕ is an integrable $\mathcal{B}(\mathbb{R}^p)$-measurable function taking values in an appropriate range space. Each expectation can be represented as the Lebesgue integral of ϕ over the state space $\mathbb{R}^p$, with respect to the conditional law $P^\theta_{X_k|\mathcal{O}_n}$ of X_k under the measure associated with θ;

$$E_\theta\{\phi(X_k)|\mathcal{O}_n\} = \int_{\mathbb{R}^p} \phi(x)\ P^\theta_{X_k|\mathcal{O}_n}(dx).$$

The sufficient statistics can be interpreted as the "local moments" of different portions of the conditional distribution, determined by the structure of the weighting functions covering $\mathbb{R}^p$. Evaluation of this class of integrals may be difficult analytically, and numerical techniques are required unless the conditional distribution of X and the weighting functions w_k assume a particularly simple form.

The standard affine regression problem is recovered as a special case, when $K = 1$, $w_1 \equiv 1$. If $w_k = I_{A_k}$, calculation of the sufficient statistics involves dividing up the probability mass into regions A_k, fitting a regression to each region separately. If w_k and the conditional probability density of X can both be expressed as a sum of Gaussian functions, the sufficient statistics can be evaluated explicitly as a function of the means and covariances of the Gaussian components.

By adopting a maximum-likelihood approach based on the EM algorithm, the parameter estimation problem reduces to an equivalent non-linear smoothing problem requiring construction of the conditional law of X_k^n from each $\mathcal{O}_n$, under a sequence of different probability measures (cf. [7] [8] [9]).

4. State Estimation

State estimation involves determining the general class of conditional expectations $E\{\phi(X_k)|\mathcal{F}_n^Y\}$ of measurable functionals of the X process. The conditional law of X, $P_{X_k|\mathcal{F}_n^Y} : \mathcal{B}(\mathbb{R}^p) \to [0,1]$, is obtained as a special case by noting that

$$P_{X_k|\mathcal{F}_n^Y}(A) \quad = \quad E\{I(X_k \in A)|\mathcal{F}_n^Y\} \quad : \quad \forall A \in \mathcal{B}(\mathbb{R}^p),$$

and contains all information about the X process that can be extracted from measurement of $Y_1 \ldots Y_n$. It is of particular interest to consider cases where this measure (or the corresponding density function, if it exists) can be propagated recursively from the measurements. Temporal correlation in the X and Y processes makes it difficult to accomplish this directly; by introducing a new probability measure, under which Y is an i.i.d. process independent of X, the problem can be circumvented. This idea is originally due to Brémaud and van Schuppen [10], and was later developed in [11] [12]. Here we duplicate the approach used by Elliott [12], to yield recursive smoothing formulae for our model. First construct processes $\lambda = \{\lambda_n : n \in \mathbb{N}\}$, $\Lambda = \{\Lambda_n : n \in \mathbb{N}\}$, where

$$\lambda_n \quad \triangleq \quad \exp(\frac{1}{2}h(X_n)^T\Sigma_w^{-1}h(X_n) - h(X_n)^T\Sigma_w^{-1}Y_n),$$
$$\Lambda_n \quad \triangleq \quad \Pi_{k\le n}\lambda_k.$$

Under the original measure P, $E\Lambda_n = 1$ and Λ is a uniformly-integrable $\mathcal{G}$-martingale, converging a.s. and in L_1 to some integrable $\mathcal{G}_\infty$-measurable r.v. Λ_∞. A new measure $\overline{P}$, absolutely continuous w.r.t. P on $(\Omega, \mathcal{G}_\infty)$, can be obtained by defining the Radon-Nikodym derivative $d\overline{P}/dP|_{\mathcal{G}_\infty} = \Lambda_\infty$. Using a discrete-time version of the Cameron-Martin-Girsanov theorem, it can be shown that under $\overline{P}$, $\mathcal{F}_\infty^X$ and $\mathcal{F}_\infty^Y$ are independent σ-algebras, Y is an i.i.d. Gaussian process with distribution $N(0, \Sigma_w)$, and the restrictions of P and $\overline{P}$ to $(\Omega, \mathcal{F}_\infty^X)$ coincide. A reverse measure change can be effected by defining similar processes $\overline{\lambda}$, $\overline{\Lambda}$, with $\overline{\lambda}_n = 1/\lambda_n$, and recovers the original probability measure P. Furthermore, conditional expectations under P and $\overline{P}$ are related by the Kallianpur-Striebel formula,

$$E\{\phi(X_k)|\mathcal{F}_n^Y\} = \frac{\overline{E}\{\phi(X_k)\overline{\Lambda}_n|\mathcal{F}_n^Y\}}{\overline{E}\{\overline{\Lambda}_n|\mathcal{F}_n^Y\}} \triangleq \frac{\sigma_n(\phi(X_k))}{\sigma_n(1)} : k \le n.$$

The state estimation problem is solved if the unnormalized quantity $\sigma_n(\phi(X_k))$ can be calculated from the observations. Under appropriate assumptions, there exist forward, backward, and smoothed unnormalized conditional densities α_k, β_k, γ_k for X_k such that for all Borel functions $\phi(X_k)$,

$$\sigma_n(\phi(X_k)) \quad = \quad \int \phi(x)\ \gamma_k(x)\ dx,$$
$$\sigma_k(\phi(X_k)) \quad = \quad \int \phi(x)\ \alpha_k(x)\ dx,$$
$$\beta_k \quad = \quad \overline{E}\{(\Lambda_n/\Lambda_{k-1})|\mathcal{F}_{k-1}^X V \mathcal{F}_n^Y\}.$$

The following results are easily established;

$$\overline{E}\{\phi_k\ \overline{\Lambda}_n|\mathcal{F}_n^Y\} = \overline{E}\{\phi_k\ \beta_{k+1}\ \overline{\Lambda}_k|\mathcal{F}_n^Y\},$$
$$\overline{E}\{\phi_k\ \overline{\Lambda}_k|\mathcal{F}_k^Y\} = \overline{E}\{\overline{E}\{\phi_k\ \overline{\lambda}_k|\mathcal{F}_{k-1}^X V \mathcal{F}_k^Y\}\overline{\Lambda}_{k-1}|\mathcal{F}_k^Y\},$$
$$\beta_k = \overline{E}\{\overline{\lambda}_k\overline{E}\{(\overline{\Lambda}_n/\overline{\Lambda}_k)|\mathcal{F}_k^X V \mathcal{F}_n^Y\}|\mathcal{F}_{k-1}^X V \mathcal{F}_n^Y\}.$$

To develop explicit recursions for α_k, β_k, γ_k, assume that X is a Markov process, admitting a transition density $\rho_k : \mathbb{R}^p \times \mathbb{R}^p \to [0,1]$ such that for all Borel functions $f(X_k)$,

$$E\{f(X_k)|\mathcal{F}_{k-1}^X\} \quad = \quad \int f(x)\ \rho_k(x, X_{k-1})\ dx.$$

Using the independence of X and Y under $\overline{P}$, and the definitions of α, β, γ, ρ, with $\alpha_0 \equiv \beta_n \equiv 1$, we finally obtain

$$\alpha_k(x) \quad = \quad \overline{\lambda}_k(y_k, x)\int \rho_k(x,\xi)\ \alpha_{k-1}(\xi)\ d\xi,$$
$$\beta_k(x) \quad = \quad \int \overline{\lambda}_k(y_k,\xi)\rho_k(\xi,x)\ \beta_{k+1}(\xi)\ d\xi,$$
$$\gamma_k(x) \quad = \quad \alpha_k(x)\beta_{k+1}(x).$$

These results provide the recursions necessary for propagating the filtered and smoothed conditional densities of the unobserved state process X from measurements of Y, in the general framework where X is Markov, and Y is generated from X by the static mapping $h(\cdot)$. Numerical integration techniques are generally needed to evaluate the conditional densities, but the assumptions that have been placed on the structure of $h(\cdot)$ typically permit further simplification. In principle, the parameter and state estimation problems are solved. A related application may be found in [13].

5. Conclusions

The purpose of this paper was to outline a statistical approach to training the articulatory-acoustic mapping from partial acoustic observations, using a functional model of the vocal tract. Applying the EM algorithm, it has been shown that solution of the problem in a maximum-likelihood sense relies on solving an associated state-estimation problem to gather statistics from the measurement data.

It is interesting to note that the state estimation procedure can be interpreted as a probabilistic formulation of articulatory-acoustic inversion. Traditional deterministic approaches to the inverse problem rely on searching for an optimum state trajectory corresponding to the observed acoustic measurements, using a functional approximation to the forward mapping $h(\cdot)$. Since $h(\cdot)$ is non-unique, additional constraints are usually imposed to regularize the problem. However, given that there may not be enough information in the acoustic signal to recover the true articulation precisely, it is not clear that this gives an adequate representation of the phonetic ambiguity inherent in speech production. When the articulation cannot be recovered uniquely, a more natural approach would be to attempt to characterize the class of all possible movements which might have produced the data. In a statistical framework, this corresponds to constructing a conditional probability measure on a function space of articulatory trajectories described by a model. The model specifies a prior (unconditional) distribution on a parameterized description of the underlying complete articulatory and acoustic measurements, which is then modified during recognition to reflect the influence of incoming information in the partial observed data. This is consistent with the view that speech recognition is conducted with reference to an internal model of linguistic behaviour; recovery of articulatory gestures from the speech signal is essentially a model-based non-linear filtering problem.

Recursive estimation formulae have been given for propagating the conditional density of the articulatory state, in the case where the hidden state evolves as a discrete-parameter Markov process and the memoryless mapping $h(.)$ assumes a particular structure. More general methods can be found in the remarkable papers by Zakai [14], and by Fujisaki, Kallianpur, and Kunita [15]; it is suggested that a full solution of the articulatory-acoustic inversion problem will follow from application of these results to speech.

6. REFERENCES

1. Schroeter J. and Sondhi M.M. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1):133–150, 1994.
2. Stokbro K., Umberger D.K., and Hertz J.A. Exploiting neurons with localized receptive fields to learn chaos. *Complex Systems*, 4:603–622, 1990.
3. Gorinevsky D. On the persistency of excitation in radial basis function network identification of nonlinear systems. *IEEE Transactions on Neural Networks*, 6(5):1237–1244, 1995.
4. Jordan M.I. and Xu L. Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8(9):1409–1431, 1995.
5. Bengio Y. and Frasconi P. An input-output HMM architecture. In *Proc. NIPS-7 (preprint)*, 1995.
6. Dempster A.P., Laird N.M., and Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.
7. Dembo A. and Zeitouni O. Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm. *Stochastic Processes and their Applications*, 23:91–113, 1986.
8. Campillo F. and Le Gland F. MLE for partially observed diffusions : direct maximization vs. the EM algorithm. *Stochastic Processes and their Applications*, 33:245–274, 1989.
9. Shumway R.H. and Stoffer D.S. An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis*, 3(4):253–264, 1982.
10. Brémaud P.M. and van Schuppen J.H. SSM Technical Report 7603 : Discrete-time stochastic systems. Technical report, Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130, U.S.A., 1976.
11. Di Masi G.B. and Runggaldier W.J. On measure transformations for combined filtering and parameter estimation in discrete time. *Systems & Control Letters*, 2(1):57–62, 1982.
12. Elliott R.J. A general recursive discrete-time filter. *Journal of Applied Probability*, 30:575–588, 1993.
13. Ramsay G. and Deng L. Optimal filtering and smoothing for speech recognition using a stochastic target model. In *Proc. ICSLP-96 (these proceedings)*, 1996.
14. M. Zakai. On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11:230–243, 1969.
15. M. Fujisaki, G. Kallianpur, and H. Kunita. Stochastic differential equations for the non-linear filtering problem. *Osaka Journal of Mathematics*, 9:19–40, 1972.