SIGNAL PROCESSING VII: Theories and Applications
M. Holt, C. Cowan, P. Grant, W. Sandham (Eds.)
© 1994 European Association for Signal Processing

Parametric Modeling of PDF Using a Convolution of One-Sided Exponentials: Application to HMM

Josep VIDAL, Antonio BONAFONTE, José A.R. FONOLLOSA, Natalia Fdez. de LOSADA

Dept. of Signal Theory and Communications, Universitat Politècnica de Catalunya, Apdo. 30002,
08080 Barcelona, SPAIN, Tel./Fax (+343) 401 6457 / 401 6447, E-mail: pepe@tsc.upc.es

Abstract. We investigate in this paper how the method of moments can be used to estimate the probability density function PDF of a random variable (RV) modeled by a Markov chain. The procedure departs from the computation of the higher-order moments of the data and can be generally used as a linear parametric approach to the estimation of any unimodal PDF. One of the potential applications is speech recognition. Although Hidden Markov Models (HMM) have been proven to be one of the most successful approaches to the problem, its major weakness is that the state duration probability density functions (DPDF) are constrained to be exponential. In order to cope with this impairment it is possible to model the data (representing the duration of an speech event) through another Markov chain, and preserve the features of the training and recognition stages.

1. Introduction[1]

Since the early 80's the success of HMM has been demonstrated in speech recognition. However, in its most simple form, the probability of duration a state turns out to decrease exponentially with time, which is not an adequate representation, in light of the experimental data. In order to cope with this deficiency Russell and Cooke [2] proposed to replace each state of the HMM by another Markov chain (sub-HMM) such that the DPDF for a given state is the overall DPDF of the associated sub-HMM. Thus modeled, the DPDF of the observed RV measuring the duration happens to be the summation of exponentially distributed RV, since the duration in one state is the summation of the duration in each state of the sub-HMM. Henceforth, the use of multiple exponentials as an alternate model allows, from one point of view, improve the fit between the model and the data, and from another point of view, to preserve the Markov chain structure which allows an easy training and recognition task. Other approaches have been used in the past to improve the duration modeling [1]-[2], based on other parametric functions, as the Gamma function. Its main drawback is the complexity of the training as well as the recognition stage.

The goal of the paper is to use the method of moments to estimate the parameters of such a RV from the estimated moments of the data. The desirable feature of this approach is the linear relationship appearing between the moments and the unknowns which allow an easy estimation procedure which do not need non-linear optimization and can be used very generally to obtain parametric PDF estimations. Relations have been found for continuous RV and for discrete-lattice type ones.

2. Parametric PDF estimation

2.1 Continuous RV

We assume first that the observed RV $\{x\}$ is the sum of p exponentially distributed RV, and hence, the PDF of the data is the convolution of the individual exponential PDF:

$$f_j(x) = \lambda_j \exp(-\lambda_j x) \tag{2.1}$$

$$f(x) = f_1(x) * f_2(x) * \ldots * f_p(x) \tag{2.2}$$

The Laplace transform of equation (2.2) (particularizing for $s=j\omega$ the characteristic function is obtained) is the product of the individual Laplace transforms of each PDF. It will be very valuable to recover the parameters p and λ_j Therefore, for the model we are dealing with, we obtain:

$$\Phi(\omega) = \Phi(s)\Big|_{s=j\omega} = \int_0^\infty f(x)\exp(sx)dx\Big|_{s=j\omega} =$$
$$= \prod_{i=1}^{p} \frac{1}{(1-\frac{j\omega}{\lambda_i})} = \frac{1}{1+\sum_{i=1}^{p}(j\omega)^i a_i} \tag{2.3}$$

where the zeros of the denominator are real. According to [5], the characteristic function (CF) can be developed in a Taylor series, where the coefficient of each term of order k corresponds to the kth-order moment m_k of the RV $\{x\}$:

$$m_k = (-j)^k \frac{\partial^k \Phi(\omega)}{\partial \omega^k}\Big|_{\omega=0} = (-j)^k \Phi_k\Big|_{\omega=0} \tag{2.4}$$

[1] This work has been supported by the Spanish National Program TIC 92-0800-C05-04.

where:

$$\Phi_0 = \frac{1}{1+\sum_{i=1}^{p}(j\omega)^i a_i} = \frac{P(0)}{P(j\omega)} \tag{2.5}$$

Although unnecessary, we have introduced the term $P(0)$ in equation (2.5) for reasons that will become apparent in section 2.2. To obtain a general expression of the a_i coefficients in equation (2.3) as a function of the moments of the observed process, it is useful to introduce the second characteristic function (SCF):

$$\begin{aligned} \Psi_0 &= \ln(P(0)) - \ln(P(j\omega)) \\ \Psi_k &= \frac{\partial^k \ln \Phi(\omega)}{\partial \omega^k} = -\frac{\partial^k \ln P(j\omega)}{\partial \omega^k} \end{aligned} \tag{2.6}$$

By developing (2.4) and (2.6) and applying the derivation chain rule for the polynomial model, it is easy to show that the CF and the SCF are related through the recursions:

$$\Phi_k = \sum_{i=0}^{k-1}\binom{k-1}{i}\Psi_{k-i}\Phi_i \tag{2.7}$$

$$\Psi_k = -\sum_{i=0}^{k-1}\binom{k-1}{i}P^{(i+1)}(j\omega)\Phi_{k-i-1} \tag{2.8}$$

where the j order derivative of the polynomial is written as:

$$P^{(i)}(j\omega) = \frac{\partial^i P(j\omega)}{\partial \omega^i} \tag{2.9}$$

Since the polynomial $P(j\omega\ )$ is of finite order p the successive derivatives particularized for $\omega\ =0$ are given by the expression:

$$P^{(n)}(j\omega)\Big|_{\omega=0} = \begin{cases} j^n n!\, a_n & n \le p \\ 0 & n > p \end{cases} \tag{2.10}$$

Now, by combining equations (2.7),(2.8),(2.9) and (2.10) it is easy to derive an expression for the kth-order moment as a function of the coefficients of the polynomial and the moments up to order k-1, in the way:

$$\begin{aligned} m_k &= -\sum_{i=0}^{k-1}\sum_{n=0}^{k-i-1}(n+1)!\binom{k-1}{i}\binom{k-i-1}{n}a_{n+1}m_i\, m_{k-n-i-1} \\ m_0 &= 1 \end{aligned} \tag{2.11}$$

Not much effort is required to construct a linear relation between a_i and m_i.

2.2 Discrete lattice-type RV

This case is specially well suited when trying to estimate the parameters of a Markov chain, as stated in the introduction. Equations (2.1)-(2.11) are strongly related to the continuous nature of the RV assumed in section 2.1. If the data are known to be discretely distributed, the derivations exhibit many similarities but the final counterpart of equation is completely different. The main reason for that is the definition of the discrete PDF and hence, the use of the z-transform to obtain the CF and the SCF. The PDF of the data is still the convolution of the individual exponential PDF:

$$f_j(t) = \left(1-\exp(-\lambda_j)\right)\exp(-\lambda_j t) \tag{2.12}$$

and hence the CF turns out to be:

$$\begin{aligned} \Phi(\omega) &= \Phi(z)\Big|_{z=e^{j\omega}} = \sum_t f(t)z^t\Big|_{z=e^{j\omega}} = \\ &= \prod_{i=1}^{p}\frac{\left(1-e^{-\lambda_i}\right)}{\left(1-ze^{-\lambda_i}\right)}\Bigg|_{z=e^{j\omega}} = \frac{\prod_{i=1}^{p}\left(1-e^{-\lambda_i}\right)}{1+\sum_{i=1}^{p}\left(e^{j\omega}\right)^i a_i} = \frac{P(1)}{P(e^{j\omega})} \end{aligned} \tag{2.13}$$

The relationship between the CF and the SCF given by equation (2.7) still holds. Note however that the first difference arise in the CF, where the term $P(1)$ appears in the numerator. This term is not the unity as in section 2.1, but it is the summation of all a_j coefficients. Therefore, equation (2.8) becomes:

$$\Psi_k = -\frac{1}{P(1)}\sum_{i=0}^{k-1}\binom{k-1}{i}P^{(i+1)}(e^{j\omega})\Phi_{k-i-1} \tag{2.14}$$

where the j derivative of the polinomial with respect to ω and particularized for ω =0 is now given by the expression:

$$P^{(n)}(e^{j\omega})\Big|_{\omega=0} = j^n\sum_{i=1}^{p} i^n a_i \tag{2.15}$$

and by combining equations we obtain the counterpart of (2.11) for the discrete lattice-type RV:

$$\begin{aligned} m_k &= -\frac{1}{P(1)}\sum_{i=0}^{k-1}\sum_{n=0}^{k-i-1}\binom{k-1}{i}\binom{k-i-1}{n}\sum_{r=1}^{p} r^{n+1}a_r m_i\, m_{k-n-i-1} \\ m_0 &= 1 \end{aligned} \tag{2.16}$$

The recursion is different but there is still a linear relation between moments and parameters a. For the particular case p = 2 the linear equations are

$$\begin{pmatrix} 1 & 0 \\ 2m_1 & 2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = -\begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \tag{2.17}$$

for the continuous RV, and

$$\begin{pmatrix} 1+m_1 & 2+m_1 \\ 1+2m_1+m_2 & 4+4m_1+m_2 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = -\begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \tag{2.18}$$

for the discrete lattice-type RV. The proposed estimation procedure may be summarized in the following steps:

1. Set an upper bound of the true order p.
2. Compute p moments from the data by averaging.
3. Solve the a_i using the linear set of equations.
4. Decision on the true order p' based on the a_i.
5. If $p' \neq\ p$ then goto 3.
6. Compute the λ_i from the roots of $P(j\omega\ )$.

Moreover, if the true order p is known, we can even use more than p moments of the estimated data to construct an overdetermined system of linear equations. We could also think of estimating the poles directly, that is by building a relationship similar to (2.16), thus bypassing the estimation of the parameters **a**. Unfortunately, the relation in no longer linear, so an estimation procedure should be based on the iterative optimization of a cost function. As it will be shown in the simulations, the computational burden of this approach do not justify the improvements of the results with respect to the computation of the coefficients **a** and extraction of the roots.

It might seem that this framework is only capable to estimate one sided PDF. Note however that, for the continuous RV case, the poles are not restricted to be positive. Positive poles are the contribution to the right hand side of the PDF while negative poles render the left hand side of the PDF. On the other hand, for the discrete lattice-type RV, if $e^{-\lambda_i}>1$ we obtain a contribution to the left hand side PDF, and vice versa.

The issue of uniqueness of the solution is outside the scope of this communication. It is of course guaranteed in the continuous RV case, because of the triangular nature of the system matrix. For the discrete-lattice type RV, it is always possible to find a PDF whose moments do not specify a unique solution to equation (2.18). A particular solution can be found under a given criterion. It should be stressed here that the moments of a RV do not necessarily specify a PDF. Some examples of different PDF sharing the same set of moments can be found in [6].

3. Theoretic asymptotic performance

In general, the linear systems shown in equations (2.17) and (2.18) can be expressed as:

$$\mathbf{A}(\mathbf{m})\,\mathbf{a} = \mathbf{m} \qquad (3.1)$$

for which the unknowns are the coefficients a_i, arranged in vector **a**. In equation (3.1), **A**(**m**) is supposed to be a K × M full-rank matrix with K ≥ M, and **m** is a K × 1 vector containing the moments. The estimation procedure consist of computing the sample moments from the data to construct an estimate of the vector **m**, and then solve equation (3.2) using standard weighted least squares:

$$\hat{\mathbf{a}}=[[\mathbf{A}(\hat{\mathbf{m}})]^{\mathrm{T}}\mathbf{W}\mathbf{A}(\hat{\mathbf{m}})]^{-1}[\mathbf{A}(\hat{\mathbf{m}})]^{\mathrm{T}}\mathbf{W}\hat{\mathbf{m}} \qquad (3.2)$$

where **W** is a positive definite weighting matrix which compels the estimate of equation (3.2) to be efficient. The analytical study of the performance of the estimates thus obtained has been established in [3]. The normalized asymptotic covariance matrix of the estimates $\hat{\mathbf{a}}$ depends on the covariances of the estimated moments in the following way (Theorem 4, [3]):

$$\mathbf{P}(\mathbf{a}) = \lim_{N\to\infty} N\cdot E\left\{(\hat{\mathbf{a}}-\mathbf{a})(\hat{\mathbf{a}}-\mathbf{a})^T\right\} = \mathbf{G}(\mathbf{a})\Sigma(\mathbf{a})\mathbf{G}^T(\mathbf{a}) \qquad (3.3)$$

where Σ(**a**) is the asymptotic covariance matrix of the vector $\hat{\mathbf{m}}$, and **G**(**a**) is the Jacobian matrix of the parameters with respect to the moments. Matrix **W** is given by Theorem 5 in [3] and depends implicitly of the parameters **a**. Note however, that the construction of the weighting matrix **W** is a function of the (unknown) true parameters **a**. Equation (3.1) becomes nonlinear, and hence, cumbersome to solve. The same authors have proposed in [4] the computation of **W** directly from the data. The algorithm constructed in this way is less efficient, but remains linear. However, it is worth to consider the case K=L (same number of moments and parameters). The expression (3.2) is then independent of the weighting matrix **W**, and the solution remains statistically efficient. In this case the asymptotic covariance of $\hat{\mathbf{a}}$ given by equation (3.3) is also greatly simplified, and can be found using any symbolic mathematical package. For the continuous RV case, and for K=L=2, the corresponding vector of normalized asymptotic covariances of the estimated a_1 and a_2 are given by the diagonal terms of **P**(**a**):

$$N\cdot\mathbf{cov}(\mathbf{a}) = \begin{pmatrix} m_2 - m_1^2 \\ -\dfrac{m_2^2}{4} + 6m_1^2 m_2 - 2m_1 m_3 - 4m_1^4 + \dfrac{m_4}{4} \end{pmatrix} \qquad (3.6)$$

where N represent the number of data used to compute the averaging moments. Figures 1 and 2 show the result of equation (3.6) versus positive values of the poles, for the continuous case of section 2.1.

4. Simulations and results

The performance is illustrated in this section, both in simulated and real data.

4.1. Estimation of continuous RV

Two sets of parameters have been used to generate synthetic realizations of RV fitting the model in equation (2.2). 50 Monte Carlo realizations of 2000 data each have been run. The results can be found in Tables 1 and 2, for an order 2 and order 3 models respectively. Notice that the theoretical variances are close to those estimated. The non-linear procedure exhibits more accurate means and lower variances, but the differences do not justify the computational burden.

In another experience, we have generated 4000 random zero-mean, unity variance, Gaussian data x, and modify them through a non-linear equation of the form:

$$y = (x+3)^2$$

Table 1. Two exponentials model

True Param.	*Linear Estimates*	*Analytic std. dev.*	*Nonlinear Estimates*
$\lambda_1 = 0,8$	0,798 ± 0,124	–	0,793 ± 0,096
$\lambda_2 = 0,2$	0,202 ± 0,009		0,201 ± 0,008
$a_1 = -6,25$	-6,245 ± 0,127	0,115	–
$a_2 = 6,25$	6,328 ± 0,730	0,610	

Table 2. Three exponentials model

True Param.	*Linear Estimates*	*Nonlinear Estimates*
$\lambda\ _1 = 0,9$	0,884 ± 0,241	0,905 ± 0,241
$\lambda\ _2 = 0,5$	0,509 ± 0,085	0,466 ± 0,096
$\lambda\ _3 = 0,3$	0,305 ± 0,025	0,327 ± 0,072
$a_1 = -6,44$	-6,427 ± 0,068	–
$a_2 = 12,59$	12,50 ± 0,381	
$a_3 = -7,40$	-7,460 ± 0,850	

A sixth order (p=7) model has been used to estimate the PDF and simulated data from the estimated model have been generated. Figure 3 shows the histograms of the original data an the synthesized ones. The agreement between both is remarkable.

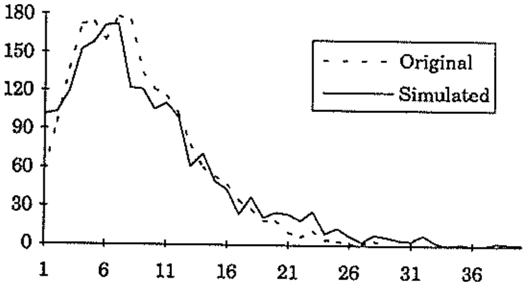

Figure 3. Histograms of 4000 realizations of a continuous type RV (dashed line) and the same number of realizations of the synthesized data (solid line).

4.2. PDF estimation of real data

In order to verify the behavior of the method in real data, we have collected 100 realizations of the RV measuring the duration (in speech frames of 30 ms each) of an HMM state, in a catalan digit data-base. The utterances have been segmented using the Viterbi algorithm and classical HMM methods, without using the transition probabilities. The nature of the RV is discrete and it takes on integer values, so that theory of section 2.2 applies. As it has been said, one-sided exponentials modeling allows the inclusion of the duration information in a Markov chain, and hence in a HMM-based speech recognition system in a very straightforward manner. In a Markov chain, the probability of staying n time steps in at given state is measured as:

$$p(n = N) = (1 - a_{jj})a_{jj}^{N-1} = (1 - e^{-\lambda_j})e^{-\lambda_j(N-1)} \quad N \geq 1$$

where 1-a_{jj} is the probability of jumping from state i to state j, and correspondingly a_{jj} is the probability of staying in state i. Note that the roots of the polynomial $P(e^{j\omega}\ )$ are directly $e^{-\lambda}\ j$, that is, the probability a_{jj}. Due to this feature of the Markov chain the PDF of the duration at a given state is constrained to be exponential. This is in contradiction with observed data histogram (figure 4). The parameters obtained through the proposed estimation procedure bring a model that fits the data very conveniently (figure 3), for as few as 100 realizations of the observed data. Results involving larger data bases are to be generated in the future, as well as its application in a HMM-based recognition system.

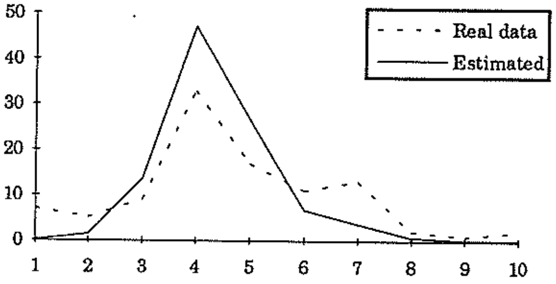

Figure 4. Histogram of the observed duration of the phoneme /θ/ through along 100 realizations of the utterance *zero* (dashed line) and the estimated PDF (solid line).

References

[1] **S. E. Levinson (1986),** "Explicit Modeling of State Occupancy in Hidden Markov Models for Automatic Speech Recognition", *Proc. ICASSP'85*, Tampa, FL, pp. 5-8, March 1985..

[2] **M. J. Russell and A. E. Cook (1987),** Experimental Evaluation of Duration Modeling Techniques for Automatic Speech Recognition", *Proc. of ICASSP'87*, pp. 2376-2379.

[3] **B. Porat and B. Friedlander (1989),** Performace Analysis of Parameter Estimation Algorithms based on High-Order Moments", *Int. Journal of Adapt. Control and Signal Proc.*, vol. 3, pp. 191-229.

[4] **B. Friedlander and B. Porat (1990),** "Asymptotically Optimal Estimation of MA and ARMA Parameters of Non-Gaussian Processes from Higher-Order Moments", *IEEE Trans. AC*, vol. 35, no. 1, Jan, 1990.

[5] **A. Papoulis (1984),** *Probability, Random Variables and Stochastic Processes*, Ed. McGraw-Hill, second edition, 1984.

[6] **G. E. Johnson (1994),** "Constructions of Particular Random Processes", *Proc. of the IEEE*, vol. 82, no. 2, pp. 270-285, Feb. 1994.