Rational Characteristic Functions and Markov Chains

Josep VIDAL, Antonio BONAFONTE, Natalia F. De LOSADA,
José A. R. FONOLLOSA, Javier R. FONOLLOSA

*Dept. of Signal Theory and Communications, Universitat Politècnica de Catalunya,Campus Nord, Modul-D5,
Gran Capità s/n, 08720 Barcelona, SPAIN, Fax (+343) 401 6447, E-mail: pepe@tsc.upc.es*

Abstract[1]

We investigate in this paper how to estimate the density function of a random variable using a parametric ARMA model for its characteristic function. The choice of this model is motivated by the fact that this type of density characterizes the duration of staying at an N-states Markov chain, but the approach is general enough to be applied to many practical problems. Both ML and moment-based linear estimates are derived, the former being based on the optimization of a highly non-linear function.

1. Introduction

Many statistical problems can be solved by modeling the density function (DF) of a random variable (RV). Among them we can quote classification, detection, estimation or simulation. Practical applications range from radar clutter classification or remote sensing to multipath channels characterization [5] and, in general, those problems in which the Central Limit Theorem does not apply. In many cases, the underlying distribution is theoretically known and a physically significant model can be applied. In other cases, the mechanics for the variable generation are unknown and then a model has to be imposed following some paradigm or heuristic points of view. On the later case it is usual to apply the maximum entropy paradigm when some of the moments are available. The resulting distribution turns out to be a product of Gaussian and sub-Gaussian distributions [6], which does not always fit the histogram of the data, because of decaying rates of the tails or assymetry of the DF. Other approaches tend to fit a series of orthogonal functions to the DF of the data. The approach followed here is to assume that our distribution is a sum of exponentially density distributed random variables, and hence its characteristic function (CF) turns out to be rational. As a result, the parameters of the CF are linearly related to the moments of the RV. Lacking of space, we will only deal with discrete-lattice type RV, although the principles can be extended to continuous-valued RVs.

2. Parametric Density Function estimation

We assume that the CF of the lattice-type observed random variable $\{x_t\}$ follows a rational form as in equation (1).

$$\Phi(\boldsymbol{w}) = \sum_n f(n)z^n \Bigg|_{z=\exp(j\boldsymbol{w})} = A(0)\frac{b_0 + \sum_{i=1}^{M} z^i b_i}{1 + \sum_{i=1}^{N} z^i a_i}\Bigg|_{z=\exp(j\boldsymbol{w})} \quad (1)$$

$$A(0) = 1 + a_1 + \ldots + a_N$$

The poles of the denominator must be real in order to get a positive DF. The unit integral property of the DF is satisfied thanks to the term $A(0)$ and the condition of unity in the sum of the b_i. Note that $\{x_t\}$ is assumed to be the sum of N exponentially distributed random variable, each having a DF:

$$f_j(n) = (1 - \boldsymbol{a}_j)\boldsymbol{a}_j^n \quad (2)$$

and an $(M+1)$-valued RV. In this sense, this work is a generalization of the equally valued all-poles approximation to a DF found in [4], chapter 11. The rational behind this model will become apparent in section 4, but for the moment, note that it is general enough to accomodate many DF types.

The CF can be expressed as:

$$A(\boldsymbol{w})\Phi(\boldsymbol{w}) = A(0)B(\boldsymbol{w}) \quad (3)$$

By continuously deriving this expression with respect to ω, we obtain the correspondent coefficients of its Taylor series, and hence the moments m_k of $\{x_t\}$ as functions of the parameters b and a.

$$m_k = \sum_{i=1}^{M} i^k b_i - \sum_{j=0}^{k-1}\sum_{s=1}^{N}\binom{k}{j}s^{k-j}v_s m_j \quad (4)$$

where:

$$v_k = \frac{a_k}{A(0)} \qquad A(0) = 1 + a_1 + \ldots + a_N$$

Note that each moment can be expressed as a sum of terms depending only on the MA part of the CF and terms depending on the AR part. From equation (4) it is possible to write down a linear matrix equation relating moments and coefficients of the CF:

$$\mathbf{m} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{TW}_1 \\ \mathbf{W}_2 & \mathbf{RW}_2 \end{pmatrix}\begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \quad (5)$$

where **m** is a vector containing the first M+N moments, $\begin{pmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{pmatrix}$ contain the first M columns of an (M+N)×(M+N) Vandermonde matrix, **T** is an N-size lower triangular matrix containing moments and **R** is a matrix with no other particularity of having the moments of the RV on it. In particular, for the M=N=2 case, the set of equations is as follows:

$$\begin{pmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \end{pmatrix} = \begin{pmatrix} 1 & 2 & -1 & -2 \\ 1 & 4 & -1-2m_1 & -4-4m_1 \\ 1 & 8 & -1-3m_1-3m_2 & -8-12m_1-6m_2 \\ 1 & 16 & -1-4m_1-6m_2-4m_3 & -16-32m_1-24m_2-8m_3 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ v_1 \\ v_2 \end{pmatrix}$$

The unicity of the solution can be shown in the following way. If the rank of the matrix is not complete, then two different sets of parameters **v**, **b** could yield the same set of moments. This is not possible because the rational function in (1) is analytic and because of the unicity of the Fourier transform.

It is worth mentioning that in general, not every RV can be model in this way. Only those moments yielding a denominator for the CF of only real valued poles can render a positive-definite DF, and hence, a DF. On the other hand, it may seem that this framework is only capable to estimate one sided DFs. General enough, this framework is capable to parametrize two sided exponentials DFs and thus to adapt to the nature of the problem. In fact, stable $(0 < a_j < 1)$ poles are the contribution to the right hand side of the DF while unstable poles render the left hand side of the DF. As a final consideration, the approach can be easily developed for continuous valued RV, by taking the Laplace transform in lieu of Z-transform (see [9] for the particular case of AR characteristic functions).

The approach presented above stresses the linear relation between the moments of the RV and the coefficients of the AR and MA polynomials. Note that the poles of the underlying DF have to be real and positive, in order to get a positive DF, and the moment-based estimation is not free of the "hard failure" of finding negative or even complex poles. A similar consideration can be done on the coefficients b: they have to be bounded between 0 and 1, and its sum must be the unity. The reasons for these failures are the finite-data length estimation of the moments or maybe because the data does not fit the model. In those cases, one has to resort to ML estimation of the poles. Assuming independent observations of the RV, the ML approach is obtained by maximizing the probability of the L observations with respect to the parameters as:

$$f(\mathbf{x}; \mathbf{a}, \mathbf{b}) = \prod_{n=1}^{L} \left[\left[\sum_{i=0}^{M} b_i \delta(x-i) \right] \otimes (1-\alpha_1)\alpha_1^{-x} \otimes \cdots \otimes (1-\alpha_N)\alpha_N^{-x} \right]_{x=x_n} \quad (6)$$

with respect to the coefficients b and the poles $\mathbf{a}$. The symbol $\otimes$ denotes convolution. The optimization process needs a good first approach to the solution that can be obtained by rooting the the linear approach estimation. In the simulations shown below, the Nelder-Mead algorithms has been used to compute the ML estimates with excelent results.

3. Theoretic performance of linear estimates

Having shown the consistency of the linear approach, it is interesting to stablish the efficiency of the estimation, that is, its normalized variance. It has to be into account for instance, if the intended application is signals classification based on the ARMA set of parameters. It is also interesting so as to be sure that good initial values are provided to the non-linear optimization procedure. The linear system shown in equation (5) can be expressed as:

$$\mathbf{A}(\mathbf{m})\,\mathbf{w} = \mathbf{m} \quad (7)$$

for which the unknowns are b_i and the normalized coefficients v_i arranged in vector **w**. The estimation procedure consist of computing the sample moments from the data to construct an estimate of the vector **m**, and then solve (7). Note that the system can be overdetermined and hence, using standard least squares:

$$\tilde{\mathbf{w}} = [[\mathbf{A}(\mathbf{m})]^T \mathbf{C}\mathbf{A}(\mathbf{m})]^{-1}[\mathbf{A}(\mathbf{m})]^T \mathbf{C}\mathbf{m} \quad (8)$$
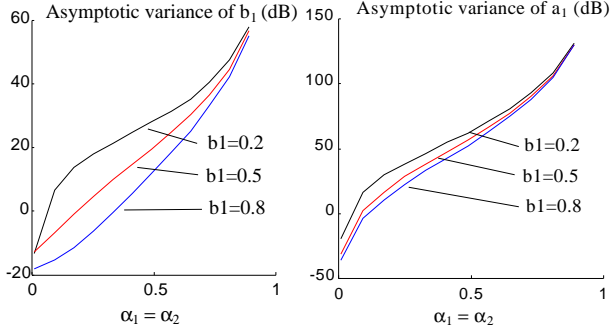
where **C** is a positive definite weighting matrix. The analytical study of the performance of the estimates thus obtained has been established in [3]. The normalized asymptotic covariance matrix of the estimates $\tilde{\mathbf{w}}$ depends on the covariances of the estimated moments in the following way (Theorem 4, [3]):

$$\mathbf{P}(\mathbf{w}) = \lim_{L \to \infty} L \cdot E\left\{ (\tilde{\mathbf{w}} - \mathbf{w})(\tilde{\mathbf{w}} - \mathbf{w})^T \right\} = \mathbf{G}(\mathbf{w})\mathbf{S}(\mathbf{w})\mathbf{G}^T(\mathbf{w}) \quad (9)$$

where $\mathbf{S}(\mathbf{w})$ is the asymptotic covariance matrix of the vector **m**, and $\mathbf{G}(\mathbf{w})$ is the Jacobian matrix of the parameters with respect to the moments. Matrix **C** is given by Theorem 5 in [3] and depends implicitly of the parameters **w**. Note however, that the construction of the weighting matrix **C** is a function of the (unknown) true parameters **w**. Equation (7) becomes nonlinear, and hence, cumbersome to solve. However, it is worth to consider the case of same number of moments and parameters. The expression (8) is then independent of the weighting matrix **W**, and the solution remains statistically efficient. In this case the asymptotic covariance of $\tilde{\mathbf{w}}$ given by equation (9) is also greatly simplified. It can be computed using any symbolic mathematical package. Results are not written here because, even for the M=1, N=2 case the corresponding vector of normalized asymptotic covariances of the estimated $\tilde{\mathbf{w}}$ contain a sum of more than 60 terms. However, in order to clarify the behaviour of the estimates, figures 1 and 2 display the variances of the parameters b_1 and a_1 for different values of b_1 (= 0.2, 0.5 and 0.8) versus values of the poles (assumed to be equal, that is $a_1 = a_2$) between 0 and 1. Although it is not shown, the variance of a_2 is quite similar to the a_1 one. The trends of the curves show that the parameters exhibit larger performance as the values of the poles approaches 1. In other words, the longer the tails of the DF, the larger the variance of the estimated parameters.

4. States duration model in a Markov chain

Since the early 80's, Hidden Markov Models are widely used to represent words in a speech recognition system (see [10] for a review on the subject). Each state in the Markov chain (see figure (3)) is characterized by the probabilities of observation of a set of AR spectra taken from the AR spectra of the speech frames. In this way, a word is modeled as a temporal sequence of spectra.

Figures 1&2. Theoretic asymptotic variances of the parameters b_1 and a_1 in an ARMA(2,1) model, versus the values of the poles. Three diferent values for b_1 have been used. Both poles are equal.

The construction of an HMM-based recognition system is done in two stages. Suppose you want your computer to recognize a set of words spoken by many different people. First, in the training stage, you should be given a sufficient number of realizations of those words uttered by different speakers. Each word is segmented into pieces or frames, and an AR spectrum is computed in each. Generally speaking, different type of spectra are associated to different phonemes, and hence, this is a suitable magnitude to be associated to each state in the Markov chain. An HM Model is built for each word to be recognized. Second, in the recognition stage, as a word is available, it is framed and the sequence of AR spectra that characterizes the word are computed. Then, the decision is based on the likelihood of each model to produce the observed word. This framework has demonstrated great improvement with respect to other approaches. However, observe that the probability of having n speech frames associated to a state (or in other words, the duration in speech frames of a given phoneme) is one-sided exponential, which, for many actual data, is inappropriate.

Some efforts have been done in the past to allow some flexibility in the inherent exponential model. Other approaches have been used in the past to improve the duration modeling [1], based on other parametric functions, as the Gamma function. Its main drawback is the complexity of the training as well as the recognition stage, which becomes unpractical as the number of states increases. 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 duration for a given state is the sum of the duration RVs associated to each sub-HMM (see figure 3). Thus modeled, the DF of the observed the duration happens to be the convolution of exponentially distributed RV. We will show next that if we also consider direct paths from the substates 1 to state 2, the CF is ARMA and the methods seen in previous sections become very useful.

Briefly, 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 the use of the standard algorithms to carry on the training and recognition tasks [10].
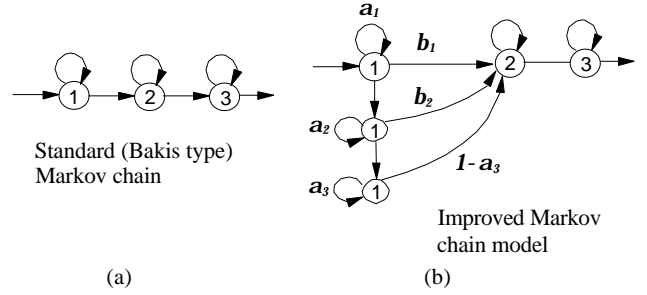

(a) (b)

Figure 3. Every single state in the standard Markov chain (a) can be substituted by another Markov chain (b). In this way, the duration modeling of state 1 is improved. The CF of the duration random variable in the sub-chain turns out to be ARMA.

Let us proceed with model (b) in figure 3 for ilustrative purposes, and then we will generalize to the most general case of N states. The succesive probabilities of remaining in state 1 during n speech frames are easily derived by induction as the probabilities of remaining at each state:

$$
\begin{aligned}
f_3(0) &= 0 \\
f_3(1) &= b_1 \\
f_3(2) &= a_1 b_1 + (1 - a_1 - b_1) b_2 \\
f_3(3) &= a_1^2 b_1 + (a_1 + a_2)(1 - a_1 - b_1) b_2 + \\
&\quad + (1 - a_1 - b_1)(1 - a_2 - b_2)(1 - a_3) \\
f_3(4) &= a_1^3 b_1 + (a_1^2 + a_1 a_2 + a_2^2)(1 - a_1 - b_1) b_2 + \\
&\quad + (a_1 + a_2 + a_3)(1 - a_1 - b_1)(1 - a_2 - b_2)(1 - a_3) \\
&\quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
f_3(n) &= X_1^{n-1} b_1 + X_{1,2}^{n-2}(1 - a_1 - b_1) b_2 + \\
&\quad + X_{1,2,3}^{n-3}(1 - a_1 - b_1)(1 - a_2 - b_2)(1 - a_3)
\end{aligned}
\tag{11}
$$

where the terms $X_{1,\dots,M}^k$ represent a multiple convolution of exponential responses with time constants equal to a_j, that is:

$$
X_{1,\dots,M}^p = \sum_{i=0}^{p}\sum_{j=0}^{p-i}\dots\sum_{l=0}^{p-i-j-\dots} a_1^i a_2^j \dots a_M^{p-i-j-\dots-l}
\tag{12}
$$

Thus, in general, for an N states sub-chain the DF of the duration RV is given by:

$$
f_N(n) = \sum_{i=1}^{N} b_i X_{1,\dots,i}^{n-i} \prod_{j=0}^{i-1}(1 - a_j - b_j) \qquad b_N = 1 - a_N
\tag{13}
$$
$$
a_0 = b_0 = 0
$$

By Z-transforming this expression, and proceeding by induction, we obtain the characteristic function, which turns out to be ARMA(N,N), with a zero at $z=0$. This zero accounts for the right displacement of $f_N(n)$ ($f_N(0)=0$). By comparing this Z-transform with equation (1) we can relate the coefficients b_i with the transition probabilities:

$$\Phi(z) = \frac{A(0)\sum_{i=1}^{N} b_i z^i}{\prod_{i=1}^{N}\left(1-\alpha_i z\right)}$$

$$\text{where}\quad A(0)b_i = \sum_{j=1}^{i}(-1)^{i-j}\beta_j\Psi_{j+1,...,N}^{i-j}\prod_{k=0}^{j-1}(1-\alpha_k-\beta_k)$$

$$\alpha_0 = \beta_0 = 0$$

$$A(0) = 1+a_1+...+a_N = \prod_{i=1}^{N}\left(1-\alpha_i\right)$$

$$\Psi_{s,...,N}^0 = 1$$

$$(14)$$

and the term $\Psi_{1,...,N}^p$ contains the sum of all possible products of p terms among the coefficients $a_1, a_2,..., a_N$, that is:

$$\Psi_{1,2,3}^2 = a_1 a_2 + a_1 a_3 + a_2 a_3$$

In this way, once the terms b and a are computed from the available data using the approach of section 3, the recovery of the transition probabilities a_j is possible by rooting the AR polinomial [7], and then, recovery of b_j using (14) is straightforward, since b_1 depends on b_1 and, further on, b_k depends on b_k and $b_1,..., b_{k-1}$.

Finally, note that if we do not allow direct transitions from sub-states to the next state (that is $b_k = 0$), then the CF is purely AR, but then the duration RV cannot take lower values than N.

5. Simulations and practical results

5.1. Estimation on simulated data

Two sets of parameters have been used to generate synthetic realizations of RV fitting the model in equation (1), corresponding to ARMA(3,1) and ARMA(3,2) CFs. A set of 10 Monte Carlo realizations of 250 data each have been run. The results can be found in Tables 1 and 2 for different models. As expected, the ML procedure exhibits more accurate means and lower variances, although it is bothering to compute. Graphic results of the second case are shown in figure 4, displaying three plots, one for the moment-based estimation, one for the ML estimation and one for the theoretical DF.

5.2. DF estimation on real data

In order to verify the goodness of fit on real data, we have collected two sets of measures (170 and 750 observations each) corresponding to the duration (in speech frames of 25 ms each) of the \t\ and \u\ phonemes from the Spanish EUROM.1 database [11]. The utterances have been segmented using the Viterbi algorithm and classical HMM methods in a Bakis-type Markov chain, without using the transition probabilities. Then, according to the parameters estimated from the data, we substitute each state by a sub-chain as in figure 3. Note that without this substitution, the probability of

staying n time steps in state 1 given by the one-sided exponential function:

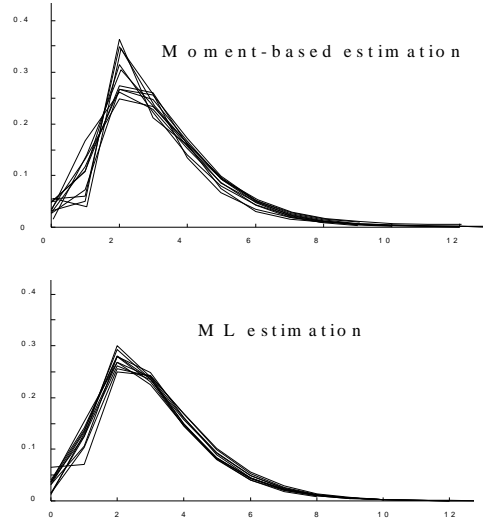$$p(n = N) = (1-a_j)a_j^{N-1}\quad N \geq 1$$

which is far from being a good approximation to the histogram of the data (see solid lines in figures 5 and 6). We have tested both the moment-based and the ML approaches to both data sets, having chosen an ARMA(3,3) model, corresponding to a Markov chain of three sub-states. The values of the transitions probabilities obtained using both the moment-based and ML estimations are shown in Tables 3 and 4. As it was said above, the moment-based procedure may render improper transition probabilites. In the first case, it was found that the poles of $A(z)$ where complex, but with a quite small imaginary part. The real part was then taken as first approximation to the ML. The parameters thus obtained fit the data very conveniently .

Table 1. ARMA(3,1) CF model

True Param.	Moment based Estimates	ML Estimates
$a_1 = -0{,}2463$	$-0{,}2493 \pm 0{,}0875$	$-0{,}2422 \pm 0{,}0471$
$a_2 = 0{,}0202$	$0{,}0161 \pm 0{,}0382$	$0{,}0117 \pm 0{,}0102$
$a_3 = -0{,}0006$	$-0{,}0007 \pm 0{,}0052$	$0{,}000 \pm 0{,}0010$
$b_0 = 0{,}5$	$0{,}5431 \pm 0{,}0806$	$0{,}6265 \pm 0{,}0690$
$b_1 = 0{,}5$	$0{,}4569 \pm 0{,}0806$	$0{,}3735 \pm 0{,}0691$

Table 2. ARMA(3,2) CF model

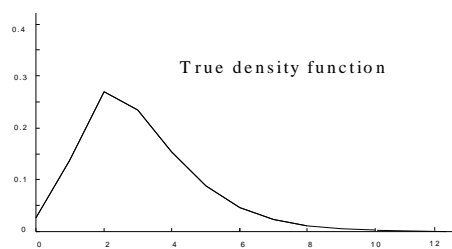| True Param.	Moment based Estimates	ML Estimates
$a_1 = -1{,}0503$	$-0{,}9232 \pm 0{,}2201$	$-1{,}0182 \pm 0{,}0683$
$a_2 = 0{,}3619$	$0{,}2575 \pm 0{,}2202$	$0{,}3362 \pm 0{,}0591$
$a_3 = -0{,}0410$	$-0{,}0217 \pm 0{,}0604$	$0{,}0361 \pm 0{,}0123$
$b_0 = 0{,}1$	$0{,}1193 \pm 0{,}3121$	$0{,}1152 \pm 0{,}2102$
$b_1 = 0{,}4$	$0{,}2315 \pm 0{,}1630$	$0{,}3172 \pm 0{,}1172$
$b_2 = 0{,}5$	$0{,}6490 \pm 0{,}1610$	$0{,}5676 \pm 0{,}0806$

Moment-based estimation

ML estimation

Figure 4. Comparison of both moment based and ML methods with the true DF corresponding to table 2.

Table 3. ARMA(3,3) fit of data from phoneme \t\

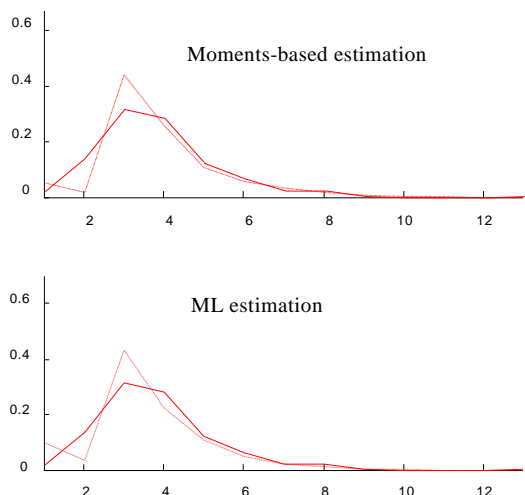| Moment-based estimation		ML estimation	
b transition probabilities	*a* transitions probabilities	*b* transition probabilities	*a* transitions probabilities
b_1=0,0513 b_2=0,0058	a_1=0,5341 a_2=0,0237 a_3=0,0237	b_1=0,1073 b_2=0,0089	a_1=0,4329 a_2=0,0191 a_3=0,0191

Figure 5. Histogram of the observed duration of the phoneme \t\ using 175 observations (solid line) and the estimated DF (dashed line) using a three-states Markov sub-chain.

Table 4. ARMA(3,3) fit of data from phoneme \x\

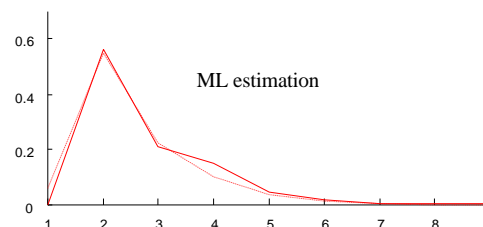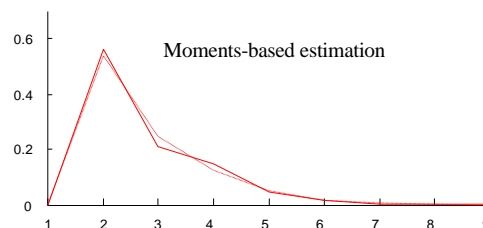| Moment-based estimation		ML estimation	
b transition probabilities	*a* transitions probabilities	*b* transition probabilities	*a* transitions probabilities
b_1=0,1682 b_2=0,3613	a_1=0,1975 a_2=0,1975 a_3=0,2658	b_1=0,2040 b_2=0,2500	a_1=0,1778 a_2=0,1778 a_3=0,2392

Figure 6. Histogram of the observed duration of the phoneme \x\ using 750 observations. Histogram (solid line) and the estimated DF (dashed line) using a three-states Markov sub-chain.

6. Conclusions

The approach presented above is a general framework for DF estimation, particularly if they present slow decaying tails. Both linear and non-linear approaches have been presented, the first exhibiting low variance in many practical cases. It may serve as well as first initialization to the ML method. Application to real phoneme duration data has been presented. Further work aims to develop criteria for order determination and the inclusion of the estimated models in a speech recognition system.

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] **A. Papoulis (1962),** *The Fourier Integral and its Applications*, McGraw-Hill, 1962

[5] **H. Hashemi (1993)**, "Impulse Response Modeling of Indoor Radio Propagation Channels", *IEEE J. On Selected Areas in Comm.,* vol. 11, no. 7, Sept. 1993.

[6] **A. Wragg, D.C. Dowson (1970),** "Fitting Continuous Probability Density Functions Over [0,∞) Using Information Theory Ideas", *IEEE Trans. on IT*, March 1970.

[7] **M. Lang, B. Frenzel (1994),** "Polynomial Root Finding", *IEEE Signal Processing Letters,* vol. 1, no. 10, October 1994.

[8] **E. Masry (1983)**, "Probability Density Estimation from Sampled Data", *IEEE Trans. on IT*, vol. IT-29, no. 5, September 1983.

[9] **J. Vidal, A. Bonafonte, J.A.R. Fonollosa, N.F. de Losada (1994)**, "Parametric Modeling of PDF using a Convolution of One-Sided Exponentials: Application to HMM", *EUSIPCO'94,* pp. 54-57, Edinburg, U.K. 13-16th September 1994.

[10] **L.R. Rabiner, B.H. Juang (1986)**, "An Introduction to Hidden Markov Models", *IEEE ASSP Magaz.*, pp. 4-16, January 1986.

[11] **A. Moreno (1994)**, "EUROM.1 Spanish Database", *Speech Technology Assesment in Multilingual Applications,* ESPRIT Project 6919, Report D6, 1994.