Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ

The Application of Bayesian Inference to Linear Prediction of Speech

G.M.K. Saleh[1] M. Niranjan[2] W.J.Fitzgerald[3]

CUED/F-INFENG/TR.205 1994

15th December 1994

This report is available by anonymous ftp from svr-ftp.eng.cam.ac.uk in /pub/reports/saleh_tr205.ps.Z

[1]email:gs113@eng.cam.ac.uk

[2]email:niranjan@eng.cam.ac.uk

[3]email:wjf@eng.cam.ac.uk

Abstract

The analysis of a speech segment is conventionally performed through linear prediction and the subsequent minimisation of a data error term in the least squares sense. The parameters derived as such maximise the likelihood of the data. In a learning problem, the addition of penalty terms, or regularisers, to the data term facilitates the estimation of the Maximum a Posteriori , or MAP, parameters. A direct equivalence can be drawn between the type of regulariser used and the prior assumptions regarding the solution. The Bayesian evidence procedure provides a framework for MAP parameter estimation and model order selection. In this paper, the use of suitable quadratic regularisers for the determination of linear prediction MAP parameters is addressed. The application of continuity constraints across successive speech segments will be demonstrated to enhance the tracking of formants for speech embedded in gaussian noise. The use of variable order models for speech analysis-synthesis is also addressed and its apparent benefits discussed.

1 Introduction

The efficient, reliable and sufficiently accurate representation of the information held within an acoustical speech waveform is of paramount importance in the various speech analysis ,coding and processing applications that are currently in use. To date, the technique of linear prediction is the most widely used and easily implementable for speech analysis purposes. [1],[2] [3],[4].

The motivation for performing linear prediction for speech analysis stems from our understanding of the speech production process. At the acoustic level, the speech waveform is produced as a result of the excitation of the vocal tract due to the glottal waveform emanating from the constriction of the vocal folds. The vocal tract is characterised by its resonances, or formants, which can be considered constant over a short length of time. The shape of the vocal tract, at any one time, and the positions of the articulators, determine the frequencies at which the vocal tract resonates.

The speech production model proposed by Fant [5] models the speech waveform as the convolution of excitation,glottal, vocal tract and lip radiation models. In the z-domain, the output speech is expressed as :

$$S(z) = E(z)G(z)V(z)L(z) \tag{1}$$

The spectral effects of the glottal and lip radiation are combined with those of the vocal tract in order to produce the simplified all-pole model for speech production shown in Figure (1). The main advantage of the all-pole model is that it separates the excitation from the vocal tract , paving the way for the parametric representation of a speech segment in terms of corresponding linear prediction parameters.

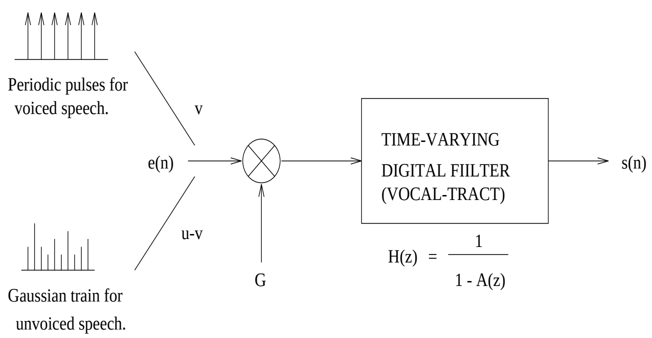

Figure 1: All-pole model for speech production characterised by gain, G, filter A(z), and v/u-v switch

The transfer function of the all-pole filter is written as :

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}} \tag{2}$$

which in the time domain translates to expressing the current speech sample as a linear combination of the past p samples and the current input:

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gu(n) \tag{3}$$

Speech is broadly classified as belonging to one of two categories, voiced or unvoiced. The excitation to the filter takes the form of either a series of pulses for voiced speech or random gaussian noise for unvoiced speech.The frequency response of the filter signifies the broad spectrum of speech whilst the variations in pitch are controlled by the excitation. In analysing a speech segment, the aim is thus to spectrally match the spectrum of the estimated all pole filter to that of the speech segment under consideration through optimising the predictor or all-pole filter parameters, a_i. It can be be shown that spectral matching is achieved by minimising the total squared error between the predicted samples of the speech signal and their true values [6].

The parameter estimation methods conventionally used with linear prediction are based on minimising the total squared error between the predicted speech sample and its actual value over a short segment of speech. A review of the three most popular methods; the covariance, autocorrelation and lattice methods will be found in [1],[3]. These least squares methods do not computationally cater for any prior assumptions with regards to the nature of the particular problem under consideration. As such, it is not possible for any contextual knowledge about a specific speech segment or prior assumptions about the nature of the parameters to be embodied within the parameter estimation process. To be specific, the redundancies that exist between one speech frame and the next, and within the same speech frame are not exploited in conventional least squares methods.

The modelling of *a priori* assumptions can be achieved through the use of standard regularisation, which has its roots in function approximation, or learning theory [7]. Techniques of regularisation have been widely used for the imposition of constraints on functions which approximate given mappings. In the case of linear prediction, a penalty functional of the parameters, the regulariser, is added to the squared error term prior to minimising it.

The idea of regularisation is solidly grounded within Bayesian theory whereby a direct correspondence between the type of regulariser used and the probability distributions which are assumed to govern the model parameters can be drawn [8]. The regularisation of a solution, or inclusion of prior information, amounts to maximising the posterior probability of the parameters given the data under consideration. The model parameters arrived at thus are termed the *maximum a posteriori*, or MAP, parameters. Least squares estimation amounts to prefect prior ignorance and as such maximises the likelihood of the parameters given the data. The linear prediction parameters arrived at through least squares estimation techniques are termed *maximum likelihood*, or ML, parameters.

$$Posterior \propto Prior \times Likelihood \tag{4}$$

$$P(\mathbf{w}|D) \propto P(\mathbf{w})\, P(D|\mathbf{w}) \tag{5}$$

Through our choice of a regulariser which is deemed relevant for a particular task, or application, we can arrive at parameters which would otherwise seem less suited to our application.

This report is divided into seven sections. Section 1 introduces least squares parameter estimation. In Section 2, the type of regulariser that is used is given and a general expression for the MAP parameters is derived. Section 3 provides a statistical interpretation of ML and MAP parameters. In Section 4 the bayesian evidence framework [8] which is used for estimating the parameters and performing model comparison is briefly over-viewed. In section 5, the use of MAP parameter estimates for tracking of formants is demonstrated for a synthetic waveform and for speech embedded in gaussian noise. In section 6, the use of variable order models to perform analysis-synthesis is compared and contrasted with fixed order models . The data reduction rates and listening tests results are also given.Finally, the last section offers a discussion of the results and explores areas for future work.

2 Least Squares Parameter Estimation

In order to derive a least squares expression for the linear predictor parameters, we consider the following mapping, where the n-dimensional space is that of lagged inputs :

$$S = (\mathbf{x_i}, \mathbf{y_i}) \in \Re^n \times \Re, \qquad \mathbf{i} = 1, 2, ...N \tag{6}$$

where, with reference to equation (2), $\mathbf{y_i}$ corresponds to the current speech sample, $s(n)$, and $\mathbf{x_i}$ is a vector of the past speech samples :

$$\mathbf{x_i} = \left[\begin{array}{cccc} s_{n-p} & s_{n-p-1} & \cdots & s_{n-1} \end{array} \right] \tag{7}$$

We wish to estimate the function $f(\mathbf{x_i})$ which best satisfies the mapping :

$$f(\mathbf{x_i}) = \mathbf{y_i}, \qquad \mathbf{i} = 1, 2, \cdots, N \tag{8}$$

The data error between the current output, $\mathbf{y_i}$, and its estimated value, $f(\mathbf{x_i})$, is :

$$E_D = \frac{1}{2} \sum_i \left(\mathbf{y_i} - f(\mathbf{x_i})\right)^2 \tag{9}$$

Within the context of linear prediction, we consider the mapping as being performed by a single perceptron with weight vector $\mathbf{w}$ and a linearly varying activation unit. The function, $f(\mathbf{x_i})$, can thus be written as :

$$f(\mathbf{x_i}) = \mathbf{w^T x_i} \tag{10}$$

and the data error is re-expressed as :

$$E_D = \frac{1}{2} \sum \left(\mathbf{y_i} - \mathbf{w^T x_i} \right)^2 \tag{11}$$

The gradient of the data error is given by :

$$\nabla E_D = \mathbf{Rw} - \mathbf{q} \tag{12}$$

where $\mathbf{R}$ is the hessian matrix of the data error whose elements are given by :

$$\mathbf{R}_{lm} = \sum_{i=1}^{N} \mathbf{x}_l^i \mathbf{x}_m^i \tag{13}$$

and $\mathbf{q}$ is a vector of correlations whose elements are given by :

$$\mathbf{q}_l = \sum_{i=1}^{N} \mathbf{y}_l^i \mathbf{x}_l^i \tag{14}$$

Setting ∇E_D in equation (12) to zero, we obtain the predictor parameters which coincide with the global minimum of the quadratic error surface defined by E_D :

$$\mathbf{Rw_{ml}} = \mathbf{q} \tag{15}$$

3 Imposing Continuity Constraints

We wish to further utilise the information that we have about the speech waveform in estimating the linear prediction parameters. To this end, the techniques of standard regularisation which exploit smoothness constraints in function approximation problems are used [7].

In standard regularisation, the function, $f(\mathbf{x_i})$, is estimated after minimising a cost function consisting of two terms. The first term is a data error term, as that appearing in equation (11). The second term, or the regulariser, is a penalty functional of $f(\mathbf{x_i})$ which embodies our a priori beliefs about the characteristics of $f(\mathbf{x_i})$. The cost function to be minimised is thus of the following form :

$$M = \frac{1}{2}\sum\left(\mathbf{y_i} - f\left(\mathbf{x_i}\right)\right)^2 + \lambda\|Pf\|^2 \tag{16}$$

where all the relationships given in the last section still hold and $\|Pf\|$ defines the type of constraint to be applied.

The regularisers that are used for speech analysis in this paper are all special cases of the following :

$$E_{\mathbf{w}} = \left(\mathbf{w} - \mathbf{u}\right)\mathbf{\Phi}\left(\mathbf{w} - \mathbf{u}\right)^{\mathbf{T}} \tag{17}$$

Where $\mathbf{\Phi}$ is a diagonal matrix that holds the regularisation parameters, $\mathbf{w}$ is a vector which holds the model parameters and $\mathbf{u}$ is a vector whose elements can take either zero or non-zero values.

The cost function to be minimised is thus:

$$M = \left(\mathbf{w} - \mathbf{u}\right)\mathbf{\Phi}\left(\mathbf{w} - \mathbf{u}\right)^{T} + E_D \tag{18}$$

Introducing the substitution $\mathbf{\Psi} = \frac{\mathbf{\Phi}}{\beta}$, the cost function is re-expressed as :

$$M = \left(\mathbf{w} - \mathbf{u}\right)\mathbf{\Psi}\left(\mathbf{w} - \mathbf{u}\right)^{\mathbf{T}} + \beta E_D \tag{19}$$

where $\mathbf{\Psi}$ is a diagonal matrix whose elements are the same of those of $\mathbf{\Phi}$ but are scaled with the parameter β.The purpose of introducing $\mathbf{\Psi}$ and β will be apparent in the next section where a statistical interpretation of the assumptions embodied in using the type of quadratic regulariser appearing in equation (17), β and of the related model parameters will be given.

The gradient of M is given by :

$$\nabla M = \left(\beta\mathbf{R} + \mathbf{\Psi}\right)\mathbf{w} - \beta\mathbf{q} - \mathbf{\Psi}\mathbf{u} \tag{20}$$

For known β and Ψ, the parameters obtained by setting ∇M to zero, give the minimum of the error surface defined by M. Substituting the expression for $\mathbf{w_{ml}}$ given in equation (15) in equation (20), and setting the gradient to zero, we get the following expression for the model parameters:

$$\mathbf{w_{mp}} = \left(\beta\mathbf{R} + \mathbf{\Psi}\right)^{-1}\left(\beta\mathbf{R}\mathbf{w_{ml}} + \mathbf{\Psi}\mathbf{u}\right) \tag{21}$$

4 Statistical Interpretation

As briefly mentioned in the Introduction, the $\mathbf{w_{ml}}$ and $\mathbf{w_{mp}}$ parameters can be viewed as those that maximise the likelihood of the data and the posterior probability of the parameters respectively. This nomenclature falls within the statistical interpretation of estimating a function to learn a mapping between two data sets.

With reference to equation (9), the predicted values, $f(\mathbf{x_i})$, are assumed to deviate from their actual values, $\mathbf{y_i}$, according to a gaussian distribution with variance $\frac{1}{\beta}$. Under such an assumption, the probability of the data, D, given the model parameters, $\mathbf{w}$, is governed by the following proportionality :

$$P(D|\mathbf{w}, \beta) \propto \exp(-\beta E_D) \tag{22}$$

where E_D has appeared previously in equation (11).

In order to express the posterior probability of the parameters, $P(\mathbf{w}|D)$, we need to initially make assumptions about the prior probability of our model parameters. Throughout this paper, the assumption is that the parameters of the linear prediction model are derived from independent gaussian distributions. This view is compatible with the type of quadratic regulariser that has been given in the last section. Consider the regulariser given in equation (17) which was used to derive the expression for $\mathbf{w_{mp}}$. Assuming that each of the model parameters, $\mathbf{w_i}$, is derived from a gaussian distribution of mean u_i and variance $(1|\psi_i)$, the prior over all the model parameters follows the relationship :

$$P(\mathbf{w}|\mathbf{\Psi}) \propto \exp\Big(-(\mathbf{w}-\mathbf{u})\,\mathbf{\Psi}(\mathbf{w}-\mathbf{u})^{\mathbf{T}}\Big) \tag{23}$$

Now, given the prior and likelihood, the posterior probability is expressed as :

$$\begin{aligned} P(\mathbf{w}|D, \Psi, \mathbf{u}, \beta) &\propto P(D|\mathbf{w}, \beta)\,P(\mathbf{w}|\mathbf{\Psi}) \\ P(\mathbf{w}|D, \Psi, \beta) &\propto \exp(-M) \end{aligned} \tag{24}$$

where M is the function to be minimised which has appeared in equation (19) and the minimisation of M corresponds to the maximisation of the posterior probability, for known $\mathbf{\Psi}$ and β.

The parameters $\mathbf{\Psi}$ and β reflect the relative importance given to the prior with respect respect to the data. If $\mathbf{\Psi}$ is set to zero, the optimisation of the model parameters will rely entirely on the data set under consideration and the model parameters will correspond to $\mathbf{w_{ml}}$ (see equations (21) and (24). As $\mathbf{\Psi}$ becomes larger, the prior plays a bigger role in the determination of the model parameters. In the limiting case where the components of $\mathbf{\Psi}$ are too large with respect to β, the parameter estimates will simply be the means of the gaussian distributions given in their corresponding priors, which are the components of $\mathbf{u}$.

We wish to find the model parameters that maximise their posterior probability, $P(\mathbf{w}|D)$. The interpretation given to $\boldsymbol{\Psi}$ and β so far leads to the suggestion that optimum values of $\boldsymbol{\Psi}$ and β exist which, when substituted in equation (21), give an estimate of the required MAP model parameters. The approach adopted in this paper is based on the evidence framework [8] , the implementation of which is reviewed in the next section. The posterior probability of the parameters, $P(\mathbf{w}|D)$, is written as :

$$P(\mathbf{w}|D) = \int P(\mathbf{w}|D, \Psi, \beta,) \, P(\Psi, \beta|D) \, \mathrm{d}\Psi d\beta \tag{25}$$

The second term in the integral above, $P(\Psi, \beta|D)$, is assumed to be sharply peaked around optimum values of $\boldsymbol{\Psi}$ and β, denoted by $\hat{\boldsymbol{\Psi}}$ and $\hat{\beta}$. As such, the maximum of $P(\mathbf{w}|D)$ is approximated by $P\left(\mathbf{w}|D, \hat{\boldsymbol{\Psi}}, \hat{\beta}\right)$. An outline of all the approximations and assumptions which the evidence framework is based on will be found in [8],[9].

It is worth noting at this point that alternative approaches for the calculation of MAP parameters can be based on marginalising $\boldsymbol{\Psi}$ out of the prior distribution $P(\mathbf{w}|\boldsymbol{\Psi})$ and obtaining an expression for the exact posterior distribution $P(\mathbf{w}|D)$. The maximum of the posterior distribution is then seeked for the MAP parameters (see [10], [11]).

5 Bayesian Evidence Framework

The Evidence framework utilises bayesian inference in order to perform parameter estimation and model comparison in a unified and consistent manner. In the first level of inference, that of parameter estimation, the values of $\boldsymbol{\Psi}$ and β are optimised by maximising their *evidence*, which gives a measure of their posterior probability. The second level of inference deals with model comparison to choose the most plausible model, given the data mapping used. Again, the *evidence* of a model is the criterion used to determine the goodness of the model.

For a given model,$\mathcal{M}$, the posterior probability of the parameters, $\mathbf{w}$, given $\boldsymbol{\Psi}$, β, is written fully as :

$$P(\mathbf{w}|D, \boldsymbol{\Psi}, \beta, \mathcal{M}) = \frac{P(D|\mathbf{w}, \beta, \mathcal{M}) \, P(\mathbf{w}|\boldsymbol{\Psi}, \mathcal{M})}{P(D|\Psi, \beta, \mathcal{M})} \tag{26}$$

where the denominator, $P(D|\boldsymbol{\Psi}, \beta, \mathcal{M})$, is the evidence for $\boldsymbol{\Psi}$ and β.

We wish to find $\hat{\boldsymbol{\Psi}}$ and $\hat{\beta}$, which maximise the posterior probability $P(\mathbf{w}|\mathbf{D}, \boldsymbol{\Psi}, \beta, \mathcal{M})$. To this end, and with reference to equation (25), the values of $\boldsymbol{\Psi}$ and β, that maximise their posterior probability, $P(\boldsymbol{\Psi}, \beta, |D)$, are obtained.

The posterior probability of $\boldsymbol{\Psi}$ and β is given by :

$$P(\boldsymbol{\Psi}, \beta|D, \mathcal{M}) = \frac{P(D|\boldsymbol{\Psi}, \beta, \mathcal{M}) \, P(\boldsymbol{\Psi}, \beta|\mathcal{M})}{P(D|\mathcal{M})} \tag{27}$$

Within the first level of inference, $P(D|\mathcal{M})$ is constant.$P(\boldsymbol{\Psi}, \beta, \mathcal{M})$ is assumed to be a uniform non-informative prior.As such, the evidence for $\boldsymbol{\Psi}$ and β, $P(D|\boldsymbol{\Psi}, \beta, \mathcal{M})$ is evaluated as a measure of their posterior probability. The maximum of the evidence is used to denote $\hat{\boldsymbol{\Psi}}$ and $\hat{\beta}$, which will maximise the posterior probability of $\mathbf{w}$.

Now, $P(D|\boldsymbol{\Psi}, \beta, \mathcal{M})$ is the normalising constant in equation (26) :

$$\begin{aligned} P(D|\boldsymbol{\Psi}, \beta, \mathcal{M}) &= \int P(D|\mathbf{w}, \beta, \mathcal{M}) P(\mathbf{w}|\boldsymbol{\Psi}, \mathcal{M}) d\mathbf{w} \\ &= \frac{Z_M}{Z_W Z_D} \end{aligned}$$

where

$$Z_M = \int \beta E_D + (\mathbf{w} - \mathbf{u}) \Psi (\mathbf{w} - \mathbf{u})^{\mathbf{T}} d\mathbf{w} \tag{28}$$

$$Z_D = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}} \tag{29}$$

$$Z_W = \prod_j \left(\frac{2\pi}{\psi_j}\right)^{\frac{1}{2}} \tag{30}$$

Z_M is evaluated in closed form after performing a second order Taylor expansion on M around the $\mathbf{w_{mp}}$ parameters and evaluating the resulting expression as a gaussian integral.

$$Z_M = \frac{exp(-M_{mp})(2\pi)^{\frac{k}{2}}}{\sqrt{\det \boldsymbol{\Psi} + \beta \mathbf{R}}} \tag{31}$$

The log evidence for Ψ and β is thus written as :

$$\begin{aligned} \log P(D|\boldsymbol{\Psi}, \beta) &= -\left(\beta E_D + (\mathbf{w_{mp}} - \mathbf{u}) \boldsymbol{\Psi} (\mathbf{w_{mp}} - \mathbf{u})^{\mathbf{T}}\right) \\ &\quad -\frac{1}{2} \log\det(\boldsymbol{\Psi} + \beta \mathbf{R}) + \frac{k}{2} \log(2\pi) \\ &\quad -\frac{N}{2} \log \frac{2\pi}{\beta} + \frac{1}{2} \sum_j \log(\psi)_j \end{aligned} \tag{32}$$

The derivatives of the log evidence, $\frac{d}{d\psi_j} \log P(D|\Psi, \beta, \mathcal{M})$ and $\frac{d}{d\beta} \log P(D|\Psi, \beta, \mathcal{M})$, when set to zero yield the following expressions:

$$
\begin{aligned}
2\psi_j E_{Wj} &= \gamma_j && (33)\\
2\beta E_D &= N - \sum_j \gamma_j && (34)\\
\gamma_j &= 1 - \psi_j \mathrm{tr}\left((\mathbf{\Psi} + \beta \mathbf{R})^{-1} \mathbf{I}_j \right) && (35)
\end{aligned}
$$

where $E_{Wj} = \frac{1}{2}(w_j - u_j)^2$, $\mathbf{I}_j(j,j) = 1$ and all other elements of $\mathbf{I}_j$ are zero.

For the case when a single regularisation parameter, ψ, is used , equations (33)-(35) are rewritten as :

$$
\begin{aligned}
2\psi E_W &= \gamma && (36)\\
2\beta E_D &= N - \gamma && (37)\\
\gamma &= k - \psi \mathrm{tr}(\psi \mathbf{I} + \beta R)^{-1} && (38)
\end{aligned}
$$

where $\mathbf{I}$ is the identity matrix and $E_W = \frac{1}{2}\sum_j (w_j - u_j)^2$

The above expressions (33),(34) and (35), coupled with equation (21), can be solved recursively in order to arrive at $\hat{\mathbf{\Psi}}$, $\hat{\beta}$ and the corresponding MAP model parameters. The flowchart in figure (2), depicts the stages involved in computing the MAP parameters for a linear predictor with multiple regularisation constants. For cases when a single regularisation constant is used for all the parameters, the search for maximum takes the same form as in the flow-chart, taking into account the changes in equations (36) to (38).

The second level of inference is concerned with model comparison. The evidence for a model, $P(D|\mathcal{M})$, is used to assign preferences to different models. This follows from the following relationship, where the posterior probability of a model is expressed as :

$$
P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{M})\, P(D|\mathcal{M})}{P(D)} \tag{39}
$$

The evidence, $P(D|\mathcal{M})$, is evaluated by marginalising Ψ and β from $P(D|\mathbf{\Psi}, \beta, \mathcal{M})$ which appeared in equation (27) :

$$
P(D|\mathcal{M}) = \int P(D|\mathbf{\Psi}, \beta, \mathcal{M})\, P(\mathbf{\Psi}, \beta|\mathcal{M})\, \mathrm{d}\mathbf{\Psi} \mathrm{d}\beta \tag{40}
$$

$P(\mathbf{\Psi}, \beta|\mathcal{M})$ is assumed to be a uniform non-informative prior. The evidence for a model is thus obtained after evaluating error bars on $\log \psi_j$ and $\log \beta$:

$$
\sigma_{\log \psi_j} = \sqrt{\frac{2}{\gamma_j}} \tag{41}
$$

$$\sigma_{\log\beta} = \sqrt{\frac{2}{N - \sum_j \gamma_j}} \tag{42}$$

$$\tag{43}$$

The log evidence expression, used to assign preference to different models, $\log P\left(D|\mathcal{M}\right)$, is written in terms of $\log P\left(D|\hat{\Psi}, \hat{\beta}\right)$ as :

$$\log P\left(D|\mathcal{M}\right) = \log P\left(D|\hat{\Psi}, \hat{\beta}\right) + \log 2 - \frac{1}{2}\log\left(N - \sum_j \gamma_j\right) - \frac{1}{2}\sum \log \gamma_j \tag{44}$$

6 Order Selection for Linear Prediction

One important factor which is concerned with the representation of a speech segment is the order of the linear predictor model used in analysing it. The model order should be large enough to cater for all the formants that were used in the production of the original speech sequence, together with the source excitation and lip radiation effects. Ideally,the smallest possible adequate model order should be used in the analysis stage. Various empirical order selection methods have been suggested for selecting the order of a linear predictor. Commonly, a density of 2 poles per KHz is assumed to represent the vocal tract contribution. A further 3 to 4 poles are added in order to cater for the source excitation spectrum and radiation load [1]. We note that , in the estimation process, not all the poles are in complex conjugate pairs. As such, they would not have all necessarily contributed to the resonances of the vocal tract, or formants.

Some of the popular model selection techniques are Akaike's Information Criterion, (AIC), and Akaike's Final Prediction error, (FPE) ,[12],[13],[14]. The derivation of AIC for a model order k, $AIC\left(k\right)$, relies on the Taylor expansion of the log likelihood of a model around the maximum likelihood estimate of its parameters. For a model of order k, the relationship can be written as :

$$\begin{aligned} AIC(k) &= -2(\text{maximum log likelihood of the model}) \\ &\quad +2(\text{ number of free parameters}). \end{aligned}$$

The maximum likelihood of the model can be regarded as a biased estimator of the mean expected likelihood of the model with bias equal to k, the number of free parameters. The mean expected log likelihood is used to give a measure of the goodness of the parameters of a model.

If the noise in the model is assumed to follow a gaussian distribution, $AIC(k)$ can be written

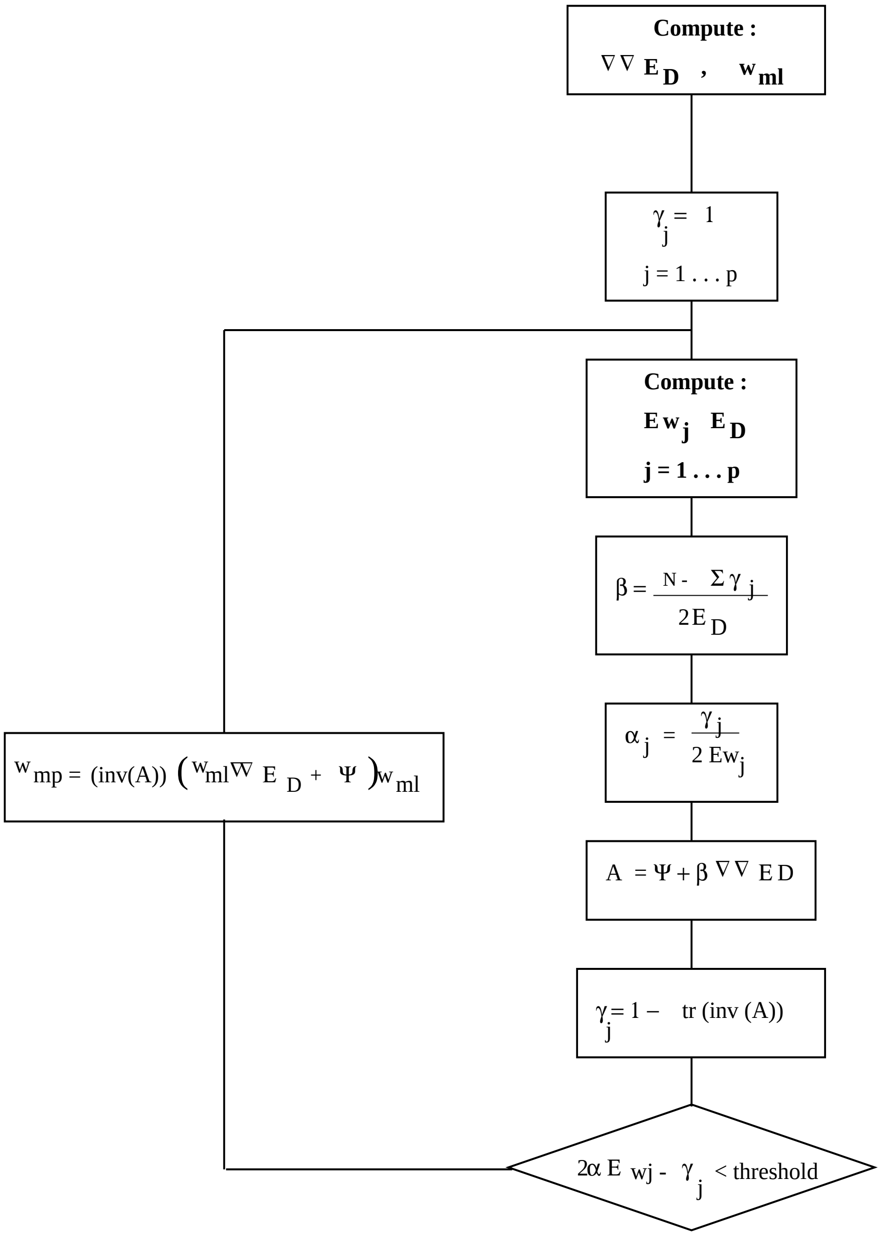

Figure 2: Flow chart depicting stages involved in the calculation of MAP parameters for a linear predictor using the evidence procedure.

as :

$$AIC(k) = n\log{\sigma_\epsilon}^2 + 2k \tag{45}$$

where σ^2 is the mean squared error incurred in prediction, which is performed over n samples.

FPE is a special case of AIC [12]. The final prediction error is given by :

$$FPE(k) = \frac{n+k}{n-k}{\sigma_\epsilon}^2 \tag{46}$$

Both AIC and FPE aim to achieve a trade-off between the residual error and the model size . This is clearly apparent in equation (45), which is composed of a data error term and a term referring to model size, or complexity.

For large n, $AIC(k)$ and $FPE(k)$ are asymptotically equivalent :

$$AIC(k) = n\log FPE(k) \tag{47}$$

In contrast to AIC and FPE, Bayesian methods for order selection make provision for prior assumptions about the solution within the parameter estimation process.The plots shown in Figure (3), depict the variation in AIC, FPE and log evidence for a linear predictor vs model order. The three criteria provide the same cue as to the model order to be used. The AIC and FPE plots are over identical parameters ML, whilst log evidence was evaluated for the equivalent MAP parameters.

7 Formant Tracking

Conventionally, the automatic tracking of formants in continuous speech is achieved through performing peak-picking on linear prediction spectra [15], [16].

Other methods that have been used include analysis-by-synthesis methods [17] , filter-bank analysis [18], log cepstra analysis [19], auditory modelling methods [20] and Kalman filtering techniques [21].

Regardless of the method used in extracting the necessary features for formant picking , it is sometimes necessary to impose a smoothness constraint in going from one speech frame to the next one. The imposition of such a constraint should be such that a closer picture of the true variation of formants is achieved . The simplest such constraint is the widely used overlap between successive speech segments, where an inherent correlation between the parameters representing successive speech segments is achieved. A successful technique that is utilised in the xwaves speech analysis package relies on Viterbi alignment as a

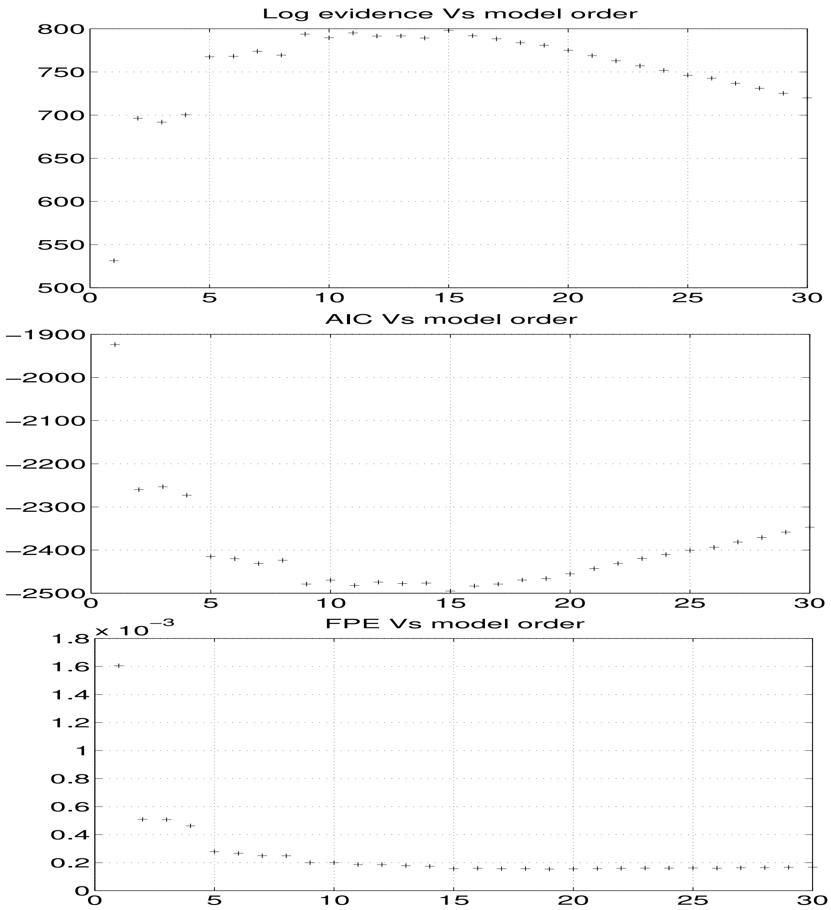

Figure 3: Plots depicting the variation of log evidence, AIC and FPE with linear predictor model order.

means of imposing continuity constraints across candidate formants [22]. The candidate formants are derived after solving the relevant all-pole filter equations.

In this section, the use of suitable MAP parameters for formant picking on linear prediction spectra is illustrated. The results are also compared with those based on ML parameter estimation. We consider the following 3 cases, with reference to the general form of the regulariser given in equation (17) :

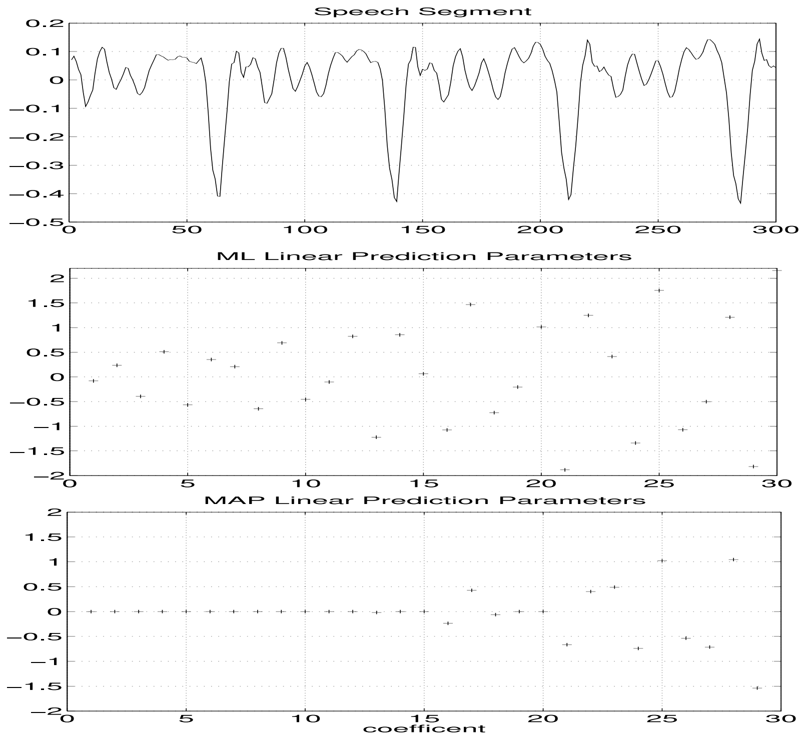

Figure 4: ML and MAP parameter estimates for a linear predictor.

(a) $\mathbf{u} = \mathbf{0}$ and the components of Ψ are distinct (zero mean gaussian priors with multiple variances).

(b) $\mathbf{u} = \mathbf{w}^{\mathbf{n}-1}$, where $\mathbf{w}^{\mathbf{n}-1}$ are the parameters estimated in the last speech segment and $\Psi = \psi \mathrm{I}$ (non-zero mean gaussian priors with a single variance, $\frac{1}{\psi}$).

(c) $\mathbf{u} = \mathbf{w}^{\mathbf{n}-1}$, where $\mathbf{w}^{\mathbf{n}-1}$ are the parameters estimated in the last speech segment and the components of Ψ are distinct (non-zero mean gaussians with distinct variances).

The aim of using regulariser in (a) is to exploit the redundancies within a waveform segment in order to set the unwanted linear predictor parameters to zero. An example of the use of multiple variance zero mean gaussian priors for estimating the linear predictor parameters is shown in Figure (4).

The values of MAP and ML coefficients are plotted against their indices for a section of a vowel of speech, sampled at 10 KHz. In this particular example, a time-delay threshold of

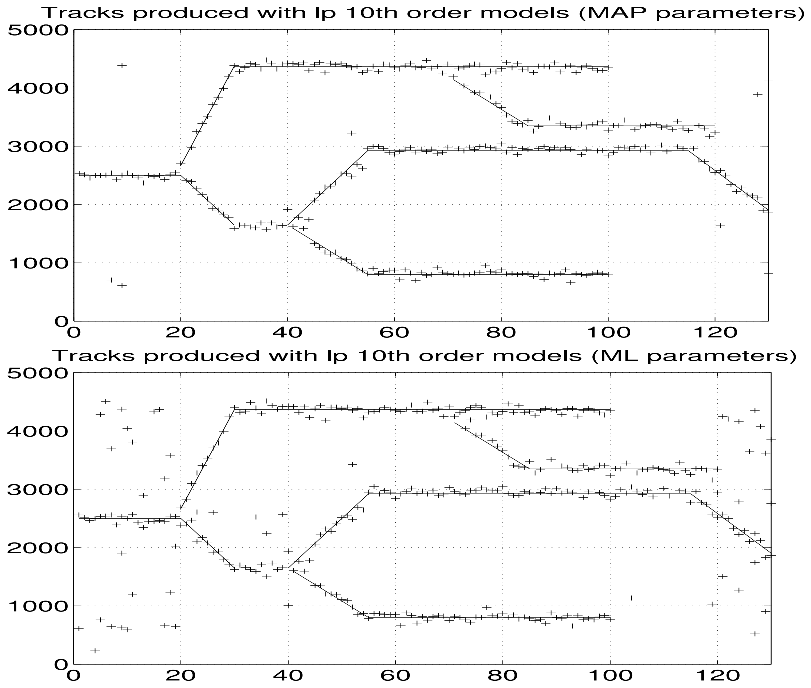

Figure 5: Plots depicting formant tracks obtained by performing peak-picking on linear prediction spectra. The two cases are for ML model parameters and MAP parameters where the priors are zero-mean gaussians with distinct variances.

15 exists beyond which all parameters are set to zero. The order of the linear predictor is thus effectively reduced to 15, instead of 30 as would have been the case under maximum likelihood analysis techniques.

A further demonstration of the ability to set unwanted parameters to zero was illustrated through the tracking of formants in a synthetic waveform simulating a variable order autoregressive process. As the waveform is purely autoregressive, the effects of glottal shaping, lip radiation and nasalisation that are assumed in analysing real speech segments are not present here. The order of the model used in generating the waveform is thus exactly twice the number of resonances that is shown in solid lines. The parameters of the generating system were updated on a block-by-block basis after 100 samples. Figure (5), depicts the frequency values estimated for the generating system after using ML estimates and MAP estimates for linear prediction. The frequency values were obtained after performing peak-picking on the linear prediction spectra.The prior for each of the model parameters was a zero mean gaussian distribution whose variance, $1/\psi$, is set independently of the others (case a above).

The use of regularisation in order to encourage inter-frame smoothness can be utilised to enhance the tracking of formants in noisy speech. This can be achieved by using regularisers of type (b) or (c). The priors for the parameters assume gaussians which are centred on the corresponding parameter estimates from the previous analysis segment. Figure (6) shows the utterance “tell me more”, embedded in gaussian noise, together with the spectrogram of its original clean version and the formant tracks derived after using regularisers (b), (c) and ML estimates. As can be seen, the use of MAP parameters facilitated the

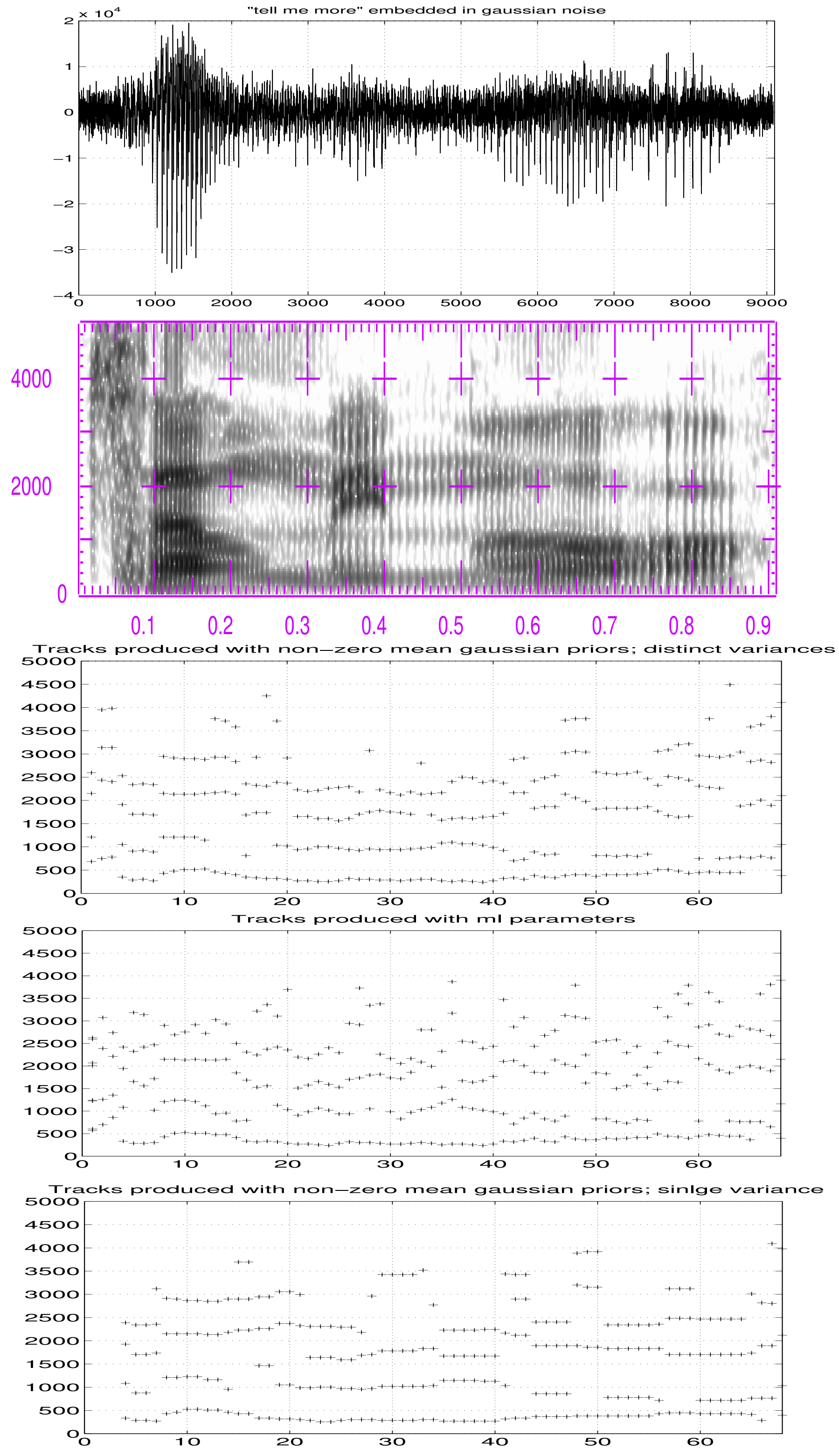

Figure 6: Plots of the noisy speech waveform, the spectrogram before adding the noise, and formant tracks produced using different prior configurations and a fixed model order of 16.

derivation of smoother tracks in comparison to the ML parameters. The tracks produced with multiple variance priors are closest to those shown in the spectrogram. With ML estimates, the absence of a smoothness constraint results in scattered tracks which are harder to trace through the speech waveform. The use of single variance non-zero mean priors, on the other hand, produced over-smoothed tracks which do not exhibit adequate variation from one frame to the next one. This can be intuitively attributed to the fact that different coefficients are forced to take the same variance value although their corresponding distributions are assumed to posses different means. As such, their variation becomes more restricted than case (b), which affects the mapping to their corresponding frequency values accordingly.

8 Analysis-synthesis Demonstrations

As mentioned in the last section, the order of a linear predictor used in analysing a speech segment is essential in ensuring its accurate representation. For cases where speech is re-constructed from its linear predictor parameters, the use of a linear predictor whose order varies according to the segment under investigation could lead to an increase in the overall efficiency when parameterising the waveform.

The following analyses of the speech utterance: "*France became the first decimal country in Europe. Germany followed eight years later and the Scandinavian states and Russia changed in 1875*", sampled at 16KHz, were performed :

i. Maximum Likelihood with 20th order models.
ii. Maximum a Posteriori with model order selection. For each frame, a search over models varying in order from 2 to 20 was performed and the model with the highest evidence was picked. Prior is a single variance zero mean gaussian distribution (regulariser (a)).
iii. Maximum Likelihood with 8th order models.

With reference to case (ii), Figure (7) shows the variation in the model order with speech frame for the utterance "*France became the first decimal country in Europe*".The model orders which are chosen are low for unvoiced and silenced speech segments and higher for voiced speech segments. For each continuous speech segment, the average number of parameters per frame was calculated and rounded up to the nearest whole number. The average number of parameters per frame was then calculated over all the continuous speech segments and was found to be 8. In contrast, the average number of parameters per frame, including silence speech is 6.05.

An 8th order system was thus used in order to re-synthesise the speech waveform and to compare the output with that from a variable-order synthesiser based on the parameterisation given in (ii). Although the total number of parameters used is the same in both cases, the distribution of parameters per frame is non-uniform in the variable synthesiser case. The aim was to gain an insight into the effect of the non-uniform distribution of parameters and assess its usefulness in improving the performance of a synthesis system.

The system used for the synthesis of speech here is based on the speech production model

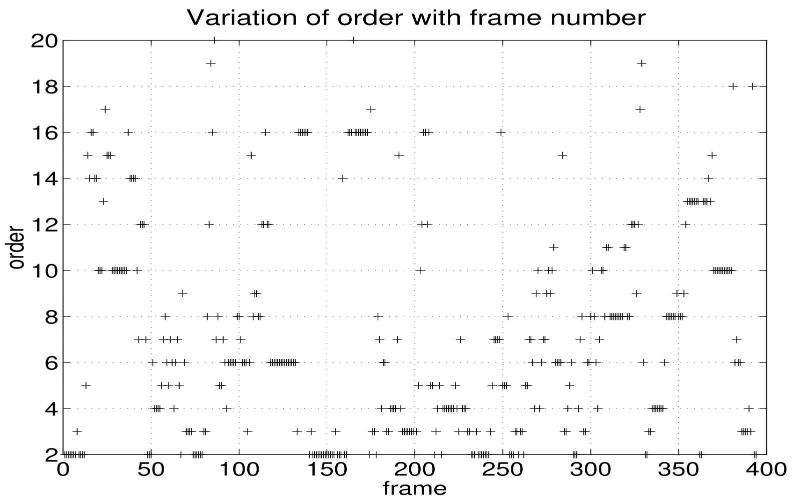

Figure 7: Variation in linear predictor order with frame for a speech segment.

that was shown in Figure (1), [23]. The waveforms representing the glottal excitation were used for the extraction of pitch and voicing decisions. The filter parameters and the gain, G, were then updated at the start of every pitch period for voiced speech and every frame length for unvoiced speech. For each synthesised speech segment, the gain of the system, G, was determined by equalising the mean squared energy in the speech signal with that in the synthesised signal [2]. No pre-emphasis was performed on the input speech and the output synthesised speech was not filtered.

A total of 5 subjects were asked to listen to speech synthesised using 8th order models and variable order models. The synthesised speech was sectioned into 5 utterances and the subjects were asked to rank the synthesised utterances according to their naturalness and closeness to their original versions. For each utterance, the original segment was played, followed by it's two synthesised versions. The synthesised segments were not played in the same order for the different utterances. Table (1), shows the preferences that were made by the subjects. Overall, the variable-order synthesiser was preferred 90 % of the times, the ML 8th order synthesiser in one occasion (5 %) and no preference could be made in one occasion.

Utterance	8th order ML	Variable order MAP	No preference
I	-	5	-
II	-	5	-
III	-	4	1
VI	1	4	-

Table 1. Ranking of synthesised speech quality made by 5 subjects on 4 different speech utterances. The numbers I-IV stand for :
I : France became the first decimal country in Europe.
II : Germany's decision followed eight years later.
III: and the Scandinavian States and Russia.
IV : Changed in 1875.

9 Conclusions

This report dealt with the usage of MAP parameter estimates within the linear prediction paradigm. The viability of their usage, in comparison to conventional ML parameters, was also assessed.

The Bayesian evidence framework was utilised in deriving MAP parameters and performing model order selection. Depending on the particular application, suitable gaussian priors were utilised with the aim of achieving a better parameterisation of speech. The linear prediction estimates were subsequently used in two applications, formant tracking and analysis-synthesis.

Formant tracking was performed by peak-picking on the linear prediction spectra. The usage of gaussian priors, with distinct variances, on the parameters, was found to result in spectra which are more representative of the speech segments under consideration. As such, the formant estimates depicted a more accurate representation of the true variation of formants in the signals that were investigated. Zero mean gaussian priors were used in order to exploit the redundancies that are present within a synthetic autoregressive waveform. On the other hand, non-zero mean gaussians were used in order to encourage parameter smoothness in going from one frame to the next.

For the analysis-synthesis application, zero mean gaussians with a single variance were used in the parameter estimation process. For each speech frame, an ensemble of linear prediction models were evaluated and the model with highest evidence was utilised in the synthesis stage. The performance of the resulting variable-order synthesiser was compared to a fixed order synthesiser which uses ML parameter estimates. The total number of parameters for the fixed order synthesiser was kept the same as for the variable order synthesiser. The quality of synthesised speech, in comparison to original versions was assessed by subjects whose judgements preferred the variable rate synthesiser on 90 % of the occasions. This result should pave the way to the use of MAP parameters in the design of low bit-rate variable speech coders.

In general, it was found that the choice of priors is a critical factor in the derivation of suitable MAP parameters. The parameter estimation process used achieves a proper balance in determining the relative importance of the prior with respect to the data. To this end, a useful insight into the relevance of the priors was gained as a result of the optimisation of the hyper-parameters, ($\boldsymbol{\Psi}$). Future work will explore the marginalisation of the hyper-parameters, as opposed to their optimisation, and the effect on the MAP parameter estimates. This should also allow the freedom of using priors which are other than gaussian in the parameter estimation process. The direct application of continuity constraints in the frequency domain will also be investigated, and the effect on formant tracking assessed.

References

[1] L.R Rabiner and R.W Schafer. *Digital Processing of Speech Signals.* Prentice-Hall, 1978.

[2] A. H. Gray and J. D. Markel. *Linear Prediction of Speech.* Springer Verlag, 1976.

[3] John Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 4(4):561–579, April 1975.

[4] Shuzo Saito and Kazuo Nakata. *Fundamentals Of Speech Signal Processing.* Academic Press, 1985.

[5] G. Fant. *Acoustic Theory of Speech Production.* Mouton, 1970.

[6] F. Fsllside and W. A. Eoods. *Computer Speech Processing.* Prentice Hall International, 1985.

[7] Tomaso Poggio and Federico Girosi. A theory of networks for approximation and laerning. Technical Report 1140, MIT. AI Laboratory, 1989. Paper No. 31.

[8] D.J.C Mackay. *Bayesian Methods for Adaptive Models.* PhD thesis, California Institute of Technology, 1992.

[9] D.J.C Mackay. Hyperparameters: Optimise, or itegrate out ? Submitted to *Neural Computation*, 1994.

[10] W. L Buntine and A. S Weigend. Bayesian back-propagation. *Complex Syetems*, 5:603–643, 1991.

[11] David H. Wolpert. On the use of evidence in neural networks. In *Advances in Neural Information Processing Systems 5*, pages 1352–1355, 1993.

[12] M.B Priestley. *Spectral Analysis and Time Series, Volume 1: Univariate Series.* Academic Press, 1981.

[13] H. Akaike. Statistical predictor identification. *Ann. Inst. Stat. Math*, 22:203–217, 1970.

[14] H. Akaike. A new look at the statistical model identification. *IEEE Tansactions on Automatic Control*, 19:716–723, 1974.

[15] Stephanie S. McCandless. An algorithm for automatic formant extraction using linear prediction spectra. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-22:135–141, 1974.

[16] J.D Markel. Application of a digital inverse filter for automatic formant and fo analysis. *IEEE transactions on Audio and Electrostatics*, AU-21(3):149–153, 1973.

[17] J. Olive. Automatic formant tracking in a newton raphson technique. *J. Acoust. Soc. Am.*, 50:661–670, August 1971.

[18] J. Flanagan. Automatic extraction of formant frequencies from continuous speech. *J. Acoust. Soc. Am.*, 28:110–118, 1956.

[19] R. W. Schafer and L. R Rabiner. System for automatic formant analysis of voiced speech. *J. Acoust. Soc. Am.*, 47(2):634–648, February 1970.

[20] Li Deng and Issam Kheirallah. Dynamic formant tracking of noisy speech using temporal analysis on outputs from a nonlinear cochlear model. *IEEE Transactions on Biomedical Engineering*, 40(5):456–465, May 1993.

[21] G. Rigoll. A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended kalman filter. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, volume ., pages 1229–1232. ., 1986.

[22] *Entropic Signal Processing System (ESPS 5); waves+ (Version 3.1)*, 1993.

[23] B. S. Atal and S. L Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *J. Acoust. Soc. Am.*, 50(2):637–655, 1971.

[24] M.B Priestley. *Non-Linear and Non-Stationary Time Series Analysis.* Academic Press, 1991.

[25] Hans Henrik Thodberg. Ace of bayes : Application of neural networks with pruning. Technical Report 11 32 E, Danish Meat Research Institute, 1993.

[26] I. Kenter Y. LeCun and S.A.Solla. Second order properties of error surfaces. In *Advances in Neural Information Processing Systems*, volume 3. Morgan Kaufmann, 1991.

[27] B.D.Ripley. Statistical aspects of neural networks. Invited Lectures for Semsat (*Seminaire European de Statistique*), Sandbjerg, Denmark. To appear in proceedings publised by Chapman and Hall, 25-30 April 1992.