Variational Bayesian Method for Time-Varying Linear Prediction Modelling

Hongjie Wan\textsuperscript{1*}, Haiyun Zhang\textsuperscript{2}

\textsuperscript{1} Information Engineering Dept, Beijing University of Chemical Technology, Beijing, 100029, China
\textsuperscript{2} College of Information Science and Engineering, Fujian University of Technology, Fuzhou, 350118, China
\*Corresponding author’s e-mail: wanhj@mail.buct.edu.cn

Abstract. The speech signal is produced by vocal tract of human beings. The characteristics of the vocal tract can be modelled by time varying linear prection analysis. In this paper, a hierarchical Bayesian model is proposed to model the time varying linear prediction model. Then Variational Bayesian method is used to estimate the parameters of the model. Experiments are carried out on real speech to demonstrate the performance of the proposed method.

1. Introduction

The speech signal which is produced by human beings can be modelled as excitation signal passing through a vocal tract model [1]. The excitation can be modelled as a mixture of pulse train and random noise [2]. The vocal tract is changing slowly as the speech is being produced by people. One approach is assuming the vocal tract is fixed in a very short time, e.g. 10 to 30 ms. Then linear prediction model is applied to the signal producing process, and the parameters of the model are fixed during the short time due to the fixed vocal tract assumption. However, this approach has been demonstrated in speech enhancement that this will yield un-continuous listening as the changing of the vocal tract is not considered [3]. Therefore, to consider the time varying characteristics of speech signal, time varying linear prediction is necessary.

In [4], the author proposed the time-varying linear prediction modelling of speech signal. The study is still based on segment, but the segment used is much longer than that of short time fixed assumption. At every point of the segment, the linear prediction coefficients are different, and each of the items is expanded by a type of base. The bases can be Fourier series, power series, cosine and sine series, Legendre polynomials [5] and wavelets [6]. The author in [7] assumed each item of a vector of linear prediction coefficient changes linearly with time increasing, so the bases are chosen as a linear line function series with a slope variable. In [8], the bases are also Legendre polynomials, where the author proposed likelihood ratio test to judge whether the time varying analysis or short time stationary analysis should be chosen for a segment of speech signal. The main problem of these researches is that the excitation modelling of the speech signal is not considered.

In this paper, a hierarchical Bayesian model is built for the time varying linear prediction model. On every time point, the coefficients are expanded using discreet sine function series; each item of the expansion coefficients is modelled using a normal distribution with precision modelled by a gamma distribution. By this approach, the problem is transformed into the estimation of the expansion
coefficients. As for the modelling of noise, a single normal distribution is used to model the excitation on every time point. The precision of the noise model is different to each other, so this can model varying noise characteristics. It can also be further noted that the mixture excitation of speech signal can be modelled by this type of modelling. The Variational Bayesian approach has been demonstrated to be a better method for parameter estimation, so it is adopted in this paper to estimate the parameters of the model.

The structure of the paper is organized as the following. The second section gives out the Bayesian model of time varying linear prediction. The third section describes the Variational Bayesian method for the estimation of the model parameters. The experiments are carried out in the fourth section. The last section concludes the paper.

2. Hierarchical Bayesian Time Varying Linear Prediction model

The speech signal can be modelled by the following time varying linear prediction equation:

\[ x(n) = \sum_{i=1}^{p} a_i(n)x(n-i) + e(n) \] (1)

Where \( n \) is the discreet time index of the digital speech signal; \( e(n) \) represents the noise at time \( n \); \( p \) represents the model order; \( a_i(n) \) represents the \( i \)-th element of the linear prediction coefficient at time \( n \). Each \( a_i(n) \) can be expanded by some base functions \( u_q(n) \) as follows

\[ a_i[n] = \sum_{k=0}^{q} a_{ik}u_k(n) \] (2)

Where \( q+1 \) is the number of the function used in the series; \( u_q(n) \) is typically chose to be 1; \( a_{ik} \) is the expansion coefficient. It can be seen that all the base functions are different at different time \( n \). Assuming there are \( N \) points in the segment of speech, then the relationship between the linear prediction coefficient \( a(n) \), the base functions and the expansion coefficients can be explained using the following matrix form

\[ \mathbf{A} = \mathbf{A}\mathbf{U} \] (3)

Where each column of \( \mathbf{A} \) is the linear coefficient vector \( a(n)=[a_1(n) \cdots a_p(n)]^T \) at time \( n \); each row of \( \mathbf{A} \) is \( \mathbf{a}_{i}=[a_{i0} \cdots a_{ia} \cdots a_{iq}] \) and \( i \) ranges from 1 to \( p \); each column of \( \mathbf{U} \) is \( \mathbf{u}(n)=[u_0(n) \cdots u_k(n) \cdots u_q(n)] \). Substituting Eq. (2) into Eq. (1), the \( N \) time points’ relationship can be represented by a matrix form as following

\[ x = \mathbf{x}a + e \] (4)

In above equation, each row in \( \mathbf{x} \) is \( \mathbf{x}^T = [x(n-1) \cdots x(n-i) \cdots x(n-p)] \otimes \mathbf{u}(n)^T \), where \( \otimes \) denotes Kronecker product; \( \mathbf{x} \) is a column vector containing the speech signal on every point; \( \mathbf{a} \) is formed by concatenating all the rows of \( \mathbf{A} \) into a column vector; \( e \) is the column vector containing the noise at every point.

After deriving the above equations, the problem is evident that the expansion coefficient vector \( \mathbf{a} \) and the noise should be estimated. There are \( p^*(q+1) \) elements in \( \mathbf{a} \). To be concise, let \( M = p(q+1) \). A normal distribution is placed on every element in \( \mathbf{a} \), and the precision \( \delta_m \) of each normal distribution is different. This can be written into a single distribution using matrix form as following

\[ p(\mathbf{a} | \text{diag}(\delta)^{-1}) = (2\pi)^{M/2} |\text{diag}(\delta)|^{M/2} \exp \left\{ -\frac{\mathbf{a}^T \text{diag}(\delta) \mathbf{a}}{2} \right\} \] (5)

The precision vector is \( \delta = [\delta_1, \cdots, \delta_p, \cdots, \delta_M]^T \). A gamma distribution on each element \( \delta_m \) as follows

\[ p(\delta_m) = \mathcal{G}(\delta_m | a_{\delta}, b_{\delta}) \] (6)

The noise on each time point is modelled using a normal distribution.
\[ p(\varepsilon_n) = \mathcal{N}(0, \lambda_n^{-1}) \]  \hspace{1cm} (7)

Where the precision is modelled using a gamma distribution:

\[ p(\lambda_n) = \mathcal{G}(\lambda_n | a_\lambda, b_\lambda) \]  \hspace{1cm} (8)

The parameters of Eq. (6) and (8) are set as \( a_\lambda = b_\lambda = a_\delta = b_\delta = 1 \) to get uninformative priors.

3. Parameter Estimation using Variational Bayesian

After building the hierarchical Bayesian model, all the parameters can be written in a single set as \( \theta = \{ \mathbf{a}, \mathbf{b}, \mathbf{\lambda} \} \). The parameter vector \( \mathbf{\lambda} \) represents all the precision of the noise on every time point. Then the full probability of the model can be obtained as

\[ p(\mathbf{x}, \theta) = p(\mathbf{x} | \mathbf{a}, \mathbf{\lambda}) p(\mathbf{a} | \mathbf{b}) p(\mathbf{\lambda}) p(\mathbf{\delta}) \]  \hspace{1cm} (9)

In Variational Bayesian, every parameter is approximated by a single distribution. Because there are three parameters, they can be marked as \( Q(\mathbf{a}) \), \( Q(\mathbf{\delta}) \) and \( Q(\mathbf{\lambda}) \). In Variational Bayesian method [9,10], the approximating distributions can be obtained by minimizing the following KL divergence

\[ KL(Q | p) = -\int Q(\mathbf{\theta}) \log \frac{p(\mathbf{\theta} | \mathbf{y})}{Q(\mathbf{\theta})} d\mathbf{\theta} \]  \hspace{1cm} (9)

For the expansion coefficient \( \mathbf{a} \), the approximate distribution is

\[ Q(\mathbf{a}) \propto \exp \left\{ \log \left[ p(\mathbf{x} | \mathbf{a}, \mathbf{\lambda}) p(\mathbf{\lambda}) p(\mathbf{\delta}) \right] \right\} \]  \hspace{1cm} (10)

The subscript of the right hand side of Eq. (10) means the distributions that the expectation will be taken with respect to. The sign \( <•> \) is used to denote the expectation of the function in the angle brackets. The subscript of Eq. (10) will be omitted hereafter. Substituting Eq.(4),(5),(7) into (10), the approximate distribution can be computed as a normal distribution with covariance \( \Sigma \) and mean \( \mu \) as following:

\[ \Sigma = \left( \mathbf{X}^T \text{diag}(\langle \mathbf{\lambda} \rangle) \mathbf{X} + \langle \mathbf{\delta} \rangle \right)^{-1} \]  \hspace{1cm} (11)

\[ \mu = \Sigma \mathbf{X}^T \text{diag}(\langle \mathbf{\lambda} \rangle) \mathbf{x} \]  \hspace{1cm} (12)

Similarly, \( Q(\mathbf{\delta}) \) can be computed with respect to \( Q(\mathbf{a}) \), and the result is a gamma distribution for each item of \( \mathbf{\delta} \) which has the following shape and rate parameters

\[ \alpha_\delta = a_\delta + \frac{1}{2} \]  \hspace{1cm} (13)
\[ \beta_\delta = b_\delta + \frac{1}{2} \langle \mathbf{a}^2 \rangle \]  \hspace{1cm} (14)

As for the precision parameter vector \( \mathbf{\lambda} \), each item of which can be derived using the same procedure and is also gamma distributed with shape and parameters as following

\[ \alpha_\lambda = a_\lambda + \frac{1}{2} \]  \hspace{1cm} (15)
\[ \beta_\lambda = b_\lambda + \frac{1}{2} \left\{ (x(n) - \mathbf{x}^T \mathbf{a})^T (x(n) - \mathbf{x}^T \mathbf{a}) \right\} \]  \hspace{1cm} (16)

It should be noted that Eq. (13) and (15) does not change as the index changes. The expectation of a gamma distribution, e.g. \( Q(\lambda_n) \) can be calculated using the following equation

\[ \langle \lambda_n \rangle = \alpha_\lambda / \beta_\lambda \]  \hspace{1cm} (17)

After estimating the parameters of Eq. (11) to (16), the parameters of the model can be approximated. This can be done by iteratively computing Eq. (11) to (16). The stopping criterion of the algorithm can be evaluated by computing the lower bound \( L(Q) \) which is defined as

\[ L(Q) = \int Q(\mathbf{\theta}) \log \frac{p(\mathbf{x}, \theta)}{Q(\mathbf{\theta})} d\mathbf{\theta} \]  \hspace{1cm} (18)
When the relative change of lower bound in two subsequent iterations does not exceed a threshold, e.g. 1e-4 the algorithm can be stopped.

4. Experiments
In this section, the algorithm will be tested on real speech signals to demonstrate the performance. The bases of the expansion are selected from the first $q+1$ items of discrete sine transform, and $q$ is set as 10. The speech signals are selected from TIMIT database, and is down sampled to 8 KHz.

![Figure 1. The lower bound change with the iteration step](image1)

Firstly, one segment of speech is given as an example to demonstrate the performance of the algorithm. The length of the signal is 450 ms. The signal contains voiced and unvoiced parts. All the samples are throwing into the algorithm. The typical linear prediction method (denoted as LPC) is compared with the proposed algorithm (denoted as TVLP), which is denoted LPC in the following contents. The frame length of LPC is set to be 20ms. The lower bound changing curve is firstly shown in Fig.1. The figure shows that the lower bound can be used as a criterion to judge whether the algorithm can be stopped. For linear prediction method, the prediction ability should be checked, so the one time prediction result is displayed in Fig. 2 and the spectrograms are displayed in Fig.3. The results show that the prediction signal is almost same as the original signal. To test the continuity of the algorithm, the predicted signal is compared with the original signal using PESQ [11] and

![Figure 2. The original signal and the predicted signal by TVLP and LPC](image2)
weighted-slope spectral distance (WSS) [12]. Higher PESQ score and small WSS means indicate better performance. The PESQ score of proposed algorithm is 3.35, but that of LPC is 2.55; the WSS of TVLP is 6.08, but that of LPC is 15.87.

![Figure 3. The spectrograms of original signal and the predicted signal by TVLP and LPC](image)

Secondly, 10 sentences are selected from TIMIT database to evaluate the performance of the algorithm. The signals are also down sampled to 8 KHz. The segment length of the signal for TVLP is 500 ms and 20 ms for LPC. The predicted signals are evaluated by PESQ and WSS. The PESQ scores and WSSs are averaged and listed in Table 1. From the table, it can be shown that the algorithm can give better results than LPC method. This result also demonstrates that the continuity of TVLP is really better than that of segmental LPC. As for WSS, the difference of TVLP and LPC is more obvious than PESQ. The results proved that the proposed algorithm can give closer spectral estimate than the short time static LPC method. In addition, the length of the TVLP is 25 times of one frame of the LPC, which means that the number of the parameters of TVLP is fewer than that of LPC.

|                | LPC  | TVLP |
|----------------|------|------|
| WSS            | 10.10| 7.32 |
| PESQ           | 2.91 | 3.10 |

5. Conclusion

In this paper, a Bayesian model is proposed for time varying linear prediction model. Normal distribution priors are placed over the expansion coefficients of the linear prediction coefficients, with individual precision for each normal distributed prior. The noise modelled using normal distribution with individual precision on every time point. Variational Bayesian is used to estimate the parameters of the model. Experiments on real speech signal show that the proposed algorithm can give better continuity than usually used short time static assumption method.

References

[1] Gold, B., Morgan, N., Ellis, D. (2011) Speech and Audio Signal Processing: Processing and Perception of Speech and Music, 2nd edition. Wiley–Interscience, New York.
[2] Mccree, A.V., Barnwell, T. P. (2002) A mixed excitation LPC vocoder model for low bit rate speech coding. IEEE Transactions on Speech and Audio Processing, 3(4):242-250.
[3] Vermaak, J., Andrieu, C., Doucet, A., Godsill, S.J. (1999) Non-Stationary Bayesian Modelling and Enhancement of Speech Signals. Signal Processing Group, Eng. Dept., Cambridge Univ., Cambridge, U.K., Tech.Rep. CUED/F-INFENG/TR.351.

[4] Hall, M.G., Oppenheim, A.V., Willsky, A.S. (1983) Time-varying parametric modeling of speech. Signal Processing, 5(3):267-285.

[5] Liporace, L. A. (1975) Linear estimation of non-stationary signals. J. Acoust.Soc. Am., 58: 1268–1295.

[6] Tsatsanis, M. K., Giannakis, G. B. (1993) Time-varying system identification and model validation using wavelets. IEEE Trans. Signal Process., 41:3512–3523.

[7] Schnell, K., Lacroix, A. (2007) Time-Varying Linear Prediction for Speech Analysis. In: 15th European Signal Processing Conference. Poland. pp. 2045-2049.

[8] Rudoy, D., Quatieri, T.F., Wolfe, P.J. Time-Varying Autoregressions in Speech: Detection Theory and Applications. (2011) IEEE Transactions on Audio Speech & Language Processing, 19(4):977-989.

[9] Bishop,C.M.. (2006) Pattern Recognition and Machine Learning. Springer.

[10] Wan, H., Ma, X., Li, X. (2018) Variational Bayesian learning for removal of sparse impulsive noise from speech signal. Digital Signal Processing, 73:106–116.

[11] Hu,Y., Loizou, P.C . (2008) Evaluation of Objective Quality Measures for Speech Enhancement. IEEE Transactions on Audio, Speech & Language Processing, 16(1):229-238.

[12] Klatt, D.H. (1982) Prediction of perceived phonetic distance from critical-band spectra: A first step. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. pp. 1278–1281.