Bayesian Modelling of the Speech Spectrum Using Mixture of Gaussians

Parham Zolfaghari, Shinji Watanabe, Atsushi Nakamura, Shigeru Katagiri

Speech Open Lab., NTT Communication Science Labs., NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
{zparham,watanabe,ats,katagiri}@cslab.kecl.ntt.co.jp

Abstract

This paper presents a method for modelling the speech spectral envelope using a mixture of Gaussians (MOG). A novel variational Bayesian (VB) framework for Gaussian mixture modelling of a histogram enables the derivation of an objective function that can be used to simultaneously optimise both model parameter distributions and model structure. A histogram representation of the STRAIGHT spectral envelope, which is free of glottal excitation information, is used for parametrisation using this MOG model. This results in a parameterisation scheme that purely models the vocal tract resonant characteristics. Maximum likelihood (ML) and variational Bayesian (VB) solutions of the mixture model on histogram data are found using an iterative algorithm. A comparison between ML-MOG and VB-MOG spectral modelling is carried out using spectral distortion measures and mean opinion scores (MOS). The main advantages of VB-MOG highlighted in this paper include better modelling using fewer Gaussians in the mixture resulting in better correspondence of Gaussians and formant-like peaks, and an objective measure of the number of Gaussians required to best fit the spectral envelope.

1. Introduction

A number of parametrisation methods exist for speech spectral modelling. These include linear predictive and cepstral coefficients which result in a smooth spectral representation. These spectral models have been used in various applications such as in speech coding and recognition with great success. A remaining objective however, is a spectral parametrisation scheme that can give an objective measure for an appropriate model order and allow the use of prior information over the data. Moreover, the data being spectral information should have little affect from signal periodicity which is not the case in LPC or cepstral based parametrisation schemes.

In this paper, we model the STRAIGHT spectral envelope [2] using mixture of Gaussians. STRAIGHT spectral envelope is not a parametric model of the spectrum but has all signal periodicity effectively removed. Our goal is to ease the correspondence between the resonant characteristics of the vocal tract and the parametric Gaussians. By considering a simple physical model of the speech production system being a non-uniform acoustic tube with transmission loss, this loss results in the divergence of spectral characteristics around the resonant frequencies. As a consequence of this divergence a less peaky structure is obtained. A MOG model could represent this characteristic by the divergence of Gaussian distributions.

Previously, an ML based MOG model in a sinusoidal analysis/synthesis framework was developed [6, 5] by one of the authors. However, when learning a mixture model there are two main complications. These include the local optima problem where the learning algorithm is trapped near an initial parameter value, and the problem of determining the appropriate model structure. In spectral modelling, the mixture model should be able to resolve formant like peaks using the least number of Gaussians. This would allow easier correspondence between Gaussian and formant. The local optima problem associated with ML-MOG learning make this difficult. Also, the use of prior information over the data is not effectively used in an ML approach.

A Bayesian modelling approach however, with the aid of prior uncertainty, can theoretically soften the local optima problem, and also it can determine the model structure through a posterior distribution over the model structure, conditional on the training data. The variational Bayes method [4] was proposed for solving the Bayesian computational difficulties with latent variables by incorporating the variational approximation technique into Bayesian learning and was extended by Attias [1] to perform model selection by introducing a posterior over the model structures. Somervuo [3] used the VB-MOG algorithm in speech modelling where several mixtures of Gaussians were trained for representing cepstrum vectors. In contrast, this paper extends VB-MOG for modelling a histogram representation of the STRAIGHT spectral envelope as a means for spectral envelope parametrisation.

2. Variational Bayes MOG Model

In mixture modelling, VB model selection facilitates the estimation of the number of Gaussian distributions in the mixture to model the data. This property is often essential in speech spectral applications such as speech coding and recognition. The reason for choosing to use VB rather than a numerical Bayesian method such as Markov chain Monte Carlo method is that, for larger corpus based applications or for real-time processing of this spectral representation the choice of VB framework was inevitable.

The normalised STRAIGHT spectral envelope (refer to section 2.2) is as shown in figure 1 where $P_t(X)$ is a continuous distribution for a single frame t. In order to represent this model on a discrete level, the spectra is viewed as a histogram where each bin x_k is associated with a uniform density function. Each bin intensity is then described by[1] $P(x_k)$. Here, x_k represents the observed incomplete data[2] and (x_k, y_k) is the

[1] Note that for simplicity, we have removed the frame index t as intra frame independence is assumed in this model.

[2] In the implementation of this model x_k is represented as the mean of the bin number

complete data, where y_k is an unobservable integer indicating the number of the Gaussian component density.

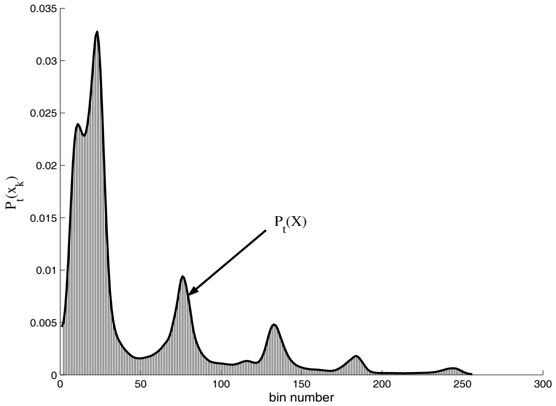

Figure 1: Normalised STRAIGHT spectrum represented as a histogram with intensity $P_t(x_k)$. $P_t(X)$ is the continuous representation of the spectral envelope.

The log-likelihood of the complete data set $X = \{x_1, \cdots, x_K\}$ and $Y = \{y_1, \cdots, y_K\}$ conditioned by parameter set Φ is given by

$$f(X, Y|\Phi) = \prod\nolimits_k \omega_{y_k} f(x_k|y_k, \Phi)^{P(x_k)}$$

where $\omega_{y_k} \equiv P(y_k|\Phi)$ is a mixture weight and $f(x_k|y_k, \Phi)$ is represented by a Gaussian

$$f(x_k|y_k = i, \Phi) = \mathcal{N}(x_k|\mu_i, \sigma_i^2) \equiv (2\pi\sigma_i^2)^{-1/2} e^{-\frac{(x_k-\mu_i)^2}{2\sigma_i^2}}$$

where i is a Gaussian component index, μ_i is the mean and σ_i is the standard deviation.

Based on the complete-data likelihood $f(X, Y|\Phi)$ a corresponding conjugate prior distribution is given by

$$f(\Phi) = \mathcal{D}(\{\omega_i\}|\{\varphi_i^0\}) \prod\nolimits_i \mathcal{N}(\mu_i|\nu_i^0, \sigma_i^2/\xi_i^0)\mathcal{G}(\sigma_i^{-2}|\alpha_i^0, \beta_i^0)$$

where $\mathcal{D}$ is a Dirichlet distribution and $\mathcal{G}$ is a Gamma distribution. The prior distribution is parameterised by hyper-parameters $\varphi_i^0, \xi_i^0, \nu_i^0, \alpha_i^0$ and β_i^0.

By introducing VB posteriors $q(\Phi|X)$ and $q(Y|X)$ the VB objective function $\mathcal{F}$, which is an objective function for not only the VB posteriors but also the model structure, is represented as

$$\mathcal{F}[q(\Phi|X), q(Y|X)] \equiv \left\langle \log \frac{f(X, Y|\Phi)f(\Phi)}{q(Y|X)q(\Phi|X)} \right\rangle_{q(Y|X)q(\Phi|X)} \qquad (1)$$

Using a variational calculation, the VB posterior of $q(\Phi|X)$ is given by

$$q(\Phi|X) = \mathcal{D}(\{\omega_i\}|\{\tilde{\varphi}_i\}) \prod\nolimits_i \mathcal{N}(\mu_i|\tilde{\nu}_i, \sigma_i^2/\tilde{\xi}_i)\mathcal{G}(\sigma_i^{-2}|\tilde{\alpha}_i, \tilde{\beta}_i)$$

where $\tilde{\varphi}_i, \tilde{\xi}_i, \tilde{\nu}_i, \tilde{\alpha}_i$ and $\tilde{\beta}_i$ are distribution parameters of $q(\Phi|X)$ and are defined as

$$\left\{ \begin{array}{lcl} \tilde{\varphi}_i & \equiv & \varphi_i^0 + \sum_k P(x_k)q(y_k = i|X) \\ \tilde{\xi}_i & \equiv & \xi_i^0 + \sum_k P(x_k)q(y_k = i|X) \\ \tilde{\nu}_i & \equiv & (\xi_i^0\nu_i^0 + \sum_k P(x_k)q(y_k = i|X)x_k)/\tilde{\xi}_i \\ \tilde{\alpha}_i & \equiv & \alpha_i^0 + \sum_k P(x_k)q(y_k = i|X) \\ \tilde{\beta}_i & \equiv & (\beta_i^0 + \xi_i^0(\nu_i^0 - \tilde{\nu}_i)^2 \\ & & \quad + \sum_k P(x_k)q(y_k = i|X)(x_k - \tilde{\nu}_i)^2)/\tilde{\alpha}_i \end{array} \right. .$$

The VB posterior $q(y_k = i|X)$ is required in the calculation of the distribution parameters of $q(\Phi|X)$ and is also obtained by a variational calculation given by

$$\begin{aligned} q(y_k = i|X) &= \frac{e^{\left\langle \log f(x_k, y_k = i|\Phi)^{p(x_k)} \right\rangle_{q(\Phi|X)}}}{e^{\left\langle \log f(x_k|\Phi)^{p(x_k)} \right\rangle_{q(\Phi|X)}}} \\ &= \frac{\tilde{\omega}_i \tilde{f}(x_k|y_k = i)}{\sum_{i'} \tilde{\omega}_{i'} \tilde{f}(x_k|y_k = i')} \end{aligned}$$

where

$$\begin{aligned} \tilde{f}(x_k|y_k = i) &\equiv e^{-\frac{1}{2}\{\log 2\pi + \frac{1}{\tilde{\xi}_i} - \Psi(\frac{\tilde{\alpha}_i}{2}) + \log(\frac{\tilde{\beta}_i}{2}) + \frac{\tilde{\alpha}_i}{\tilde{\beta}_i}(x_k - \tilde{\nu}_i)^2\}} \\ \tilde{\omega}_i &\equiv e^{\Psi(\tilde{\varphi}_i) - \Psi(\sum_i \tilde{\varphi}_i)} \end{aligned}$$

and $\Psi(.)$ is the digamma function. A locally optimal $q(\Phi|X)$ is then obtained by an iterative computation algorithm which always increases the value of $\mathcal{F}$. Moreover, an appropriate model structure can be automatically selected by searching the maximum value of $\mathcal{F}$.

2.1. Initialisation and Priors

Two methods of initialising the Gaussian mixture parameters were employed. These include:

(i) Initialising the means uniformly over the interval. The variances are made significant with respect to the interval and the number of Gaussians in the mixture. The mixture weights are set equal values. These initial parameter settings are then substituted as the prior distribution parameters $\nu_i^0, \beta_i^0, \varphi_i^0$, respectively.

(ii) Previous frame parameters are used for initialising the current frame. The VB prior distribution parameters are set as above (i).

The second case (ii) is used in conjunction with voicing decisions made available by the STRAIGHT speech analysis/synthesis system (refer to section 2.2). Whether in a voiced or an unvoiced region, for the initial frame of either region the simple initialisation method of (i) is used and for consecutive voiced or unvoiced frames, method (ii) is used. This results in a simple method for exploiting the voicing dependent correlation between adjacent frames. An alternative to this prior settings is to use formant labelled training data to enable better initialisation of mixture parameters and to devise mixture parameter learning schemes.

A Bayesian mixture estimation framework is more flexible than ML as it allows the use of hyper-parameters over the data. The optimal hyper-parameters can be estimated using the objective function given in equation 1. For specific number of distributions in the mixture, a set of hyper-parameters (ξ_i^0, α_i^0) were selected.

2.2. STRAIGHT Spectral Envelope

A smooth spectral envelope is essentially required as we are not modelling harmonic structure in this paper. STRAIGHT is fundamentally a source filter type vocoder designed for high quality analysis/modification/synthesis of speech [2]. It uses a pitch-adaptive spectral analysis scheme combined with a surface reconstruction method in the time-frequency plane in order to remove signal periodicity. This results in a smooth spectral representation free of glottal excitation information. The separation of the vocal tract and glottal excitation information is the key to this system and enables highly natural modification of pitch, vocal tract length and speaking rate. The excitation source is designed based on phase manipulation methods. Figure 2 shows a comparison between the original short-time Fourier transform (STFT) and the smoothed STRAIGHT representation for a vowel segment. Throughout this paper speech is sampled at 8 kHz, and FFT length of 512 samples was used.

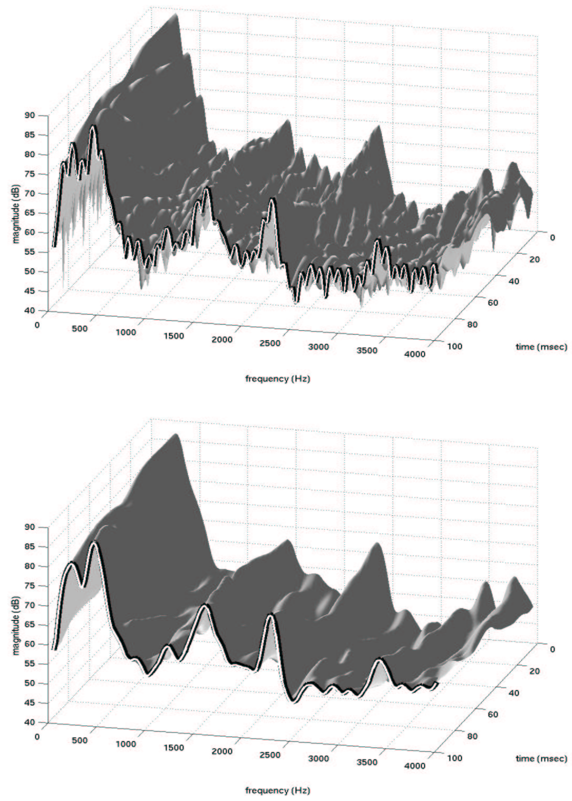

Figure 2: Comparison of STFT magnitude spectrum (Top) and the STRAIGHT smoothed spectrum (bottom) for a vowel segment of speech.

Furthermore, the spectral tilt is removed from the spectrum. This allows better fitting using the MOG model as the area under the spectra is reduced and the formant peaks are emphasised. Spectral tilt is removed from the log-spectral envelope data using the first two cepstral coefficients c_0 and c_1 computed by the following equation

$$c_m = \frac{1}{\pi}\int_0^{\pi} \log S(\omega)\cos(m\omega)d\omega, \qquad m = 0, 1$$

where m is the coefficient index and $S(\omega)$ is the spectral envelope over frequency range ω. The spectral tilt $T(\omega)$ is then given by

$$\log T(\omega) = c_0 + 2c_1\cos(\omega).$$

This spectral tilt is then removed from the measured envelope to give a flattened residual envelope $R(\omega)$ as

$$\log R(\omega) = \log S(\omega) - \log T(\omega).$$

3. Evaluations

Since VB obtains posteriors over the parameters, we regard MAP estimates of the VB posteriors as MOG spectral parameters. The VB objective function is as shown in figure 3(a) for varying number of Gaussians. To enable comparison of ML-MOG and VB-MOG based parametric modelling of spectra, average log spectral distortions between the original spectra in a vowel segment and the estimated MOG spectra was calculated. This procedure was carried out for varying number of Gaussians in the mixture and figure 3(b) illustrates this comparison.

By referring to the peak in the objective function (figure 3(a)), it can be concluded that the optimal number of Gaussians (9) given by this function in modelling the spectra corresponds clearly to the lowest VB-MOG log spectral distortion (figure 3(b)). The VB objective was found to have a similar characteristic over a number of other vowels and consonants tested and the optimal number of Gaussians seems to fall around 8 or 9 Gaussians.

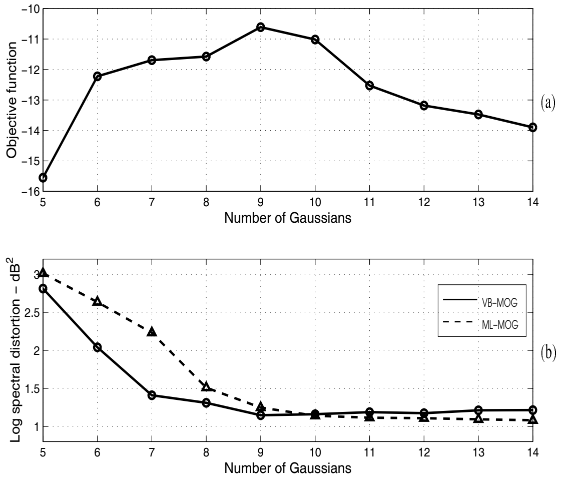

Figure 3: Plot of Bayes objective function (a) and the log spectral distortion between the original smoothed spectrum and the fitted MOG using ML and VB (b) for varying number of mixtures.

It can also be observed from the spectral distortion curves (figure 3(b)) that VB-MOG modelling with lower number of Gaussians (between 6 and 9) in a mixture results in lower spectral distortions as compared to ML-MOG. The estimated ML-MOG mean spectral parameters seem to be highly dependent on the initialisation of the parameters, as during iterative optimisation, the spread across these initial parameters is small. In a spectral estimation technique, intuitively, the mean should be relaxed further and this is one of the advantages of the Bayesian framework and the use of priors resulting in these lower spectral distortions with smaller number of Gaussians in the mixture. Figure 4 illustrates MOG modelling of the smoothed STRAIGHT spectrum using eight Gaussians. As can be seen VB-MOG performs considerably better in modelling the spectral peaks in the 1500 Hz to 3000 Hz frequency region in this specific case.

MOS subjective evaluation results for the ML-MOG and the VB-MOG spectral models using STRAIGHT analysis/synthesis are shown in figure 5. These MOS tests were carried out in an anechoic room using six male subjects. Three test sentences uttered by two males and a female were used. The

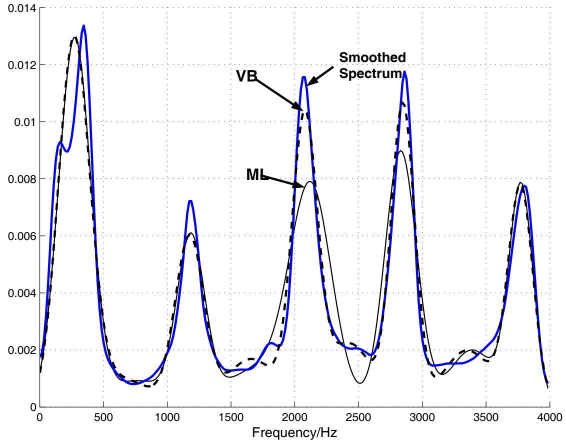

Figure 4: Comparison of fitting the smoothed spectrum with ML and VB estimated mixture Gaussians (8 Gaussians).

original and STRAIGHT re-synthesised utterances were also used in the test. The results clearly follow the average spectral distortion trend shown in figure 3. High quality speech is obtained by using over 8 Gaussians in both ML-MOG and VB-MOG. However, in the region of 6 to 8 Gaussians VB-MOG significantly outperforms ML-MOG. This clearly shows the advantage of using a Bayesian approach in STRAIGHT spectral parametrisation using mixture of Gaussians. Sound examples using this method will be given during the presentation of this paper.

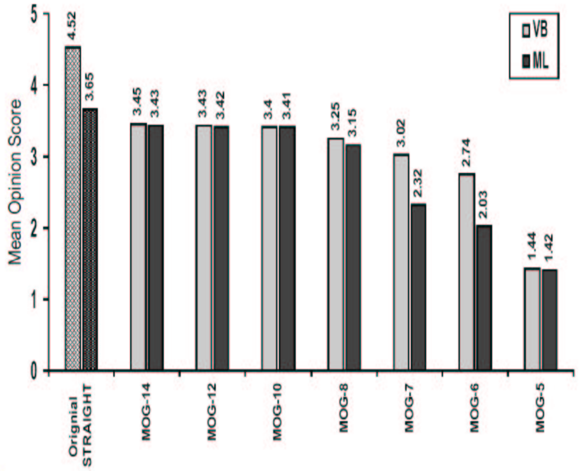

Figure 5: Comparison of mean opinion scores (MOS) over varying number of Gaussians in the mixture using VB and ML.

4. Discussion

In this paper, we have assumed that formants do not have second order resonant characteristics and can be represented by normal distributions. Through close inspection of the STRAIGHT smoothed spectrum, it is possible to infer that such an assumption is valid; MOS tests evaluating the quality of speech synthesised from the mixture of Gaussians spectral model supports this belief. Further theoretical groundwork on this matter is required to really understand and relate Gaussians and formant-like peaks in the STRAIGHT spectrum.

Though precise formant location estimation has not been emphasised in this paper, it is important to note that for future work, the VB-MOG method of spectral parameterisation should enable easier selection of Gaussians corresponding to formant like peaks in the spectrum. By allowing a better fit to the STRAIGHT spectral envelope with less Gaussians in the mixture (as compared to an ML approach) and also by using appropriate priors, we have a model that is suited for such an application. Traditional linear predictive based formant estimation techniques vary the number of poles in the model in order to find the correct number of formants. There are usually more poles than formants in the spectra and these methods always require some form of dynamic programming over time to increase their formant labelling accuracy. Our goal is to include inter and intra frame spectral statistics in the spectral model eliminating the need for such smoothing in time. The VB-MOG approach for spectral representation could realise a solution to this problem based on the Bayesian advantages described in this paper and by using labelled formant training data to devise appropriate parameter initialisation schemes. An application currently being realised is the use of state-space models with this MOG model for studying and understanding infant speech development.

5. Conclusions

A Bayesian approach for modelling a histogram representation of the STRAIGHT spectral envelope was developed. The STRAIGHT spectrum has all glottal excitation information removed resulting in a parameterisation scheme free of excitation information. A Bayesian approach results in better modelling of the spectrum with smaller number of Gaussians than ML easing correspondence of Gaussians and formant-like peaks. It also results in an objective measure of the number of Gaussians required to best fit the spectral envelope. MOS tests suggest that this VB-MOG spectral model is capable of very high quality synthesised speech. Further work will concentrate on formant frequency estimation, a segmental representation of this VB-MOG model and application to infant speech development.

6. References

[1] ATTIAS, H. Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conference on Uncertainty in Artificial Intelligence* (1999).

[2] KAWAHARA, H. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In *Proc. ICASSP* (Munich, April 1997), vol. 2, pp. 1303–1306.

[3] SOMERVUO, P. Speech modeling using variational Bayesian mixture of gaussians. In *Proceedings of the International Conference on Spoken Language Processing* (Denver, USA, Sept. 2002), pp. 1245–1248.

[4] WATERHOUSE, S., MACKAY, D., AND ROBINSON, T. Bayesian methods for mixture of experts. In *Advances in Neural Information Processing Systems 8* (1995), MIT Press.

[5] ZOLFAGHARI, P. *Sinusoidal Model Based Segmental Speech Coding*. PhD thesis, Cambridge University, 1998.

[6] ZOLFAGHARI, P., AND ROBINSON, A. Formant analysis using mixtures of Gaussians. In *Proceedings of the International Conference on Spoken Language Processing* (Philadelphia, USA, Oct 1996), vol. 2, pp. 1229–1232.