Particle Filtering Approach to Bayesian Formant Tracking

Yanli Zheng, Mark Hasegawa-Johnson

Department of Electrical Engineering
University of Illinois at Urbana-Champaign
{zheng3,jhasegawa}@uiuc.edu

This paper presents Particle Filtering Approach to Bayesian Formant Tracking. Explicit nonlinear formulas have been developed to map psd (power spectral density) of speech signal to formant frequencies. Formant tracking is formulated as a nonlinear Bayesian tracking problem and solved by particle filtering approach.

1. Introduction

In many phonological systems, sounds are classified by the action of articulators [1, 2, 3]. A reliable detection of formants and front cavity resonances over time should be capable of recognizing most of the linguistic information carried by the tongue and lips, including consonant place and vowel quality. But few speech recognizers use formant information because gross formant tracking mistakes (typically 3-5% of all voiced frames) invariably cause mistakes in phoneme recognition [4].

Recently, hidden dynamic models [4, 5] have been proposed to incorporate formant information into speech recognition by modeling the formant frequencies as hidden random variables. In these models, the relationship between formant frequencies and the mel-frequency cepstral coefficients (MFCCs) is a nonlinear function modeled by an MLP (multilayer perceptron). Optimizing the parameters of such a model is difficult because the likelihood function has a very large number of spurious local maxima [6].

This paper derives an explicit nonlinear mapping between the formant frequencies and the cepstrum. Using the derived nonlinear mapping, we demonstrate that it is possible to extract formant information from the cepstral coefficients using partile filter approach.

2. Derivation of Nonlinear Mapping Function

Cepstral coefficients have been widely used in the speech recognizer as input features. Using the Taylor series of the function $log(1-x)$, other authors have shown that the cepstrum of a vowel is the sum of exponentially decaying sinusoids at the frequencies of the formants [7]. The exponential decay properties of the cepstrum are convenient, on the one hand, because only 10-30 coefficients are necessary to encode the information about the vocal tract, but on the other hand, the rapid exponential decay of the cepstrum makes it difficult to extract the formant information coded in it. In this section, we show that by considering the correlations between frames, it is possible to extract formant information from this decay sequence.

Assume that the vocal tract transfer function can be modeled by the following function:

$$T(e^{jw}) = \prod_{m=1}^{M} \frac{1}{[(1-e^{-\sigma_m - jw})(1-e^{-\sigma_m^* - jw})]^{\eta_m/2}} \quad (1)$$

where $\sigma_m = \frac{B_m}{2} + j\omega_m$, $0 < \eta_m \leq 2$, B_m is the bandwidth, ω_m is the m^{th} formant frequency, and η_m is a scaling term that models inaccuracies in the all-pole spectral model. During vowel production, $\eta_m \approx 2$. During consonant production, $\eta_m \approx 2$ for fully excited formants, but $\eta_m \approx 0$ for formants canceled by spectral zeros or by nulls in the excitation. The log magnitude spectrum, $-\log|T(e^{j\omega})|^2$, is the sum of $4M$ different terms of the form $\log(1-z)$, where z takes the values of $e^{-\sigma_m - j\omega}$, $e^{-\sigma_m^* - j\omega}$, $e^{-\sigma_m + j\omega}$, and $e^{-\sigma_m^* + j\omega}$. Using the standard Taylor expansion of $\log(1-z)$, and sampling at the frequencies $\omega_1, \cdots, \omega_k$, we obtain:

$$\vec{y} = W\vec{x} + \vec{\mu} + \vec{e} \quad (2)$$

where

$$\vec{y} = [-log|T(e^{jw_1})|^2, \cdots, -log|T(e^{jw_k})|^2]^T$$

$$W = \begin{bmatrix} cos(\omega_1) & cos(2\omega_1) & \cdots & cos(n\omega_1) \\ cos(\omega_2) & cos(2\omega_2) & \cdots & cos(n\omega_2) \\ \vdots & \vdots & \cdots & \vdots \\ cos(\omega_k) & cos(2\omega_k) & \cdots & cos(n\omega_k) \end{bmatrix}$$

$$\vec{x} = \begin{bmatrix} \eta_1 g_{(1,1)} cos(\omega_{f_1}) + \cdots + \eta_M g_{(M,1)} cos(\omega_{f_M}) \\ \eta_1 g_{(1,2)} cos(2\omega_{f_1}) + \cdots + \eta_M g_{(M,2)} cos(2\omega_{f_M}) \\ \vdots \\ \eta_1 g_{(1,n)} cos(n\omega_{f_1}) + \cdots + \eta_M g_{(M,n)} cos(n\omega_{f_M}) \end{bmatrix}$$

where $(g_{(m,1)}, \ldots, g_{(m,n)})$ is found by clustering the coefficients in the Taylor expansion. Empirically, we find that $g_{(m,n)} \approx e^{-n\alpha_m}$ for a bandwidth-dependent decay parameter α_m. The decay factor α_m is a monotonically increasing function of the bandwidth B_m. Note that $\vec{x}$ is the inverse DCT of $\vec{y}$, and $\vec{g}_m$ is the cepstrum corresponding to a single complex pole pair at frequency ω_m, where $\vec{g}_m$ is defined as

$$\vec{g}_m = [e^{-\alpha_m} cos(2\pi f_m), \cdots, e^{-n\alpha_m} cos(2\pi f_m n)]^T$$

Given the above definitions of $\vec{x_t}$ and $\vec{g}(t)_m$, $\vec{x}_t$ and let $\varsigma = [\eta_1, \eta_2, \cdots, \eta_m]^T$,then:

$$\vec{x_t} = \vec{f_t} + \vec{e}_t \approx \sum_{m=1}^{M} \eta_m \vec{g}_m + \vec{e}_t \quad (3)$$

3. Particle Filtering Approach to Bayesian Formant Tracking

Assuming that vocal tract changes slowly with time, and that therefore the formant frequencies change little over a time interval on the order of 10ms to 30 ms, a hidden dynamic model can be formulated as follows:

$$\vec{f}_t = \vec{f}_{t-1} + \vec{v}_{t-1}, \quad \vec{v}_{t-1} \sim N(0, \sigma_f^2 I) \tag{4}$$

$$\vec{\alpha}_t = \vec{\alpha}_{t-1} + \vec{w}_{t-1}, \quad \vec{w}_{t-1} \sim N(0, \sigma_\alpha^2 I) \tag{5}$$

$$\vec{y}_t = C(\vec{f}_t, \vec{\alpha}_t)\vec{\varsigma}_t + \vec{e}_t, \quad \vec{e}_t \sim N(0, \sigma_y^2 I) \tag{6}$$

$$\text{where} \quad C(\vec{f}_t, \vec{\alpha}_t) = [\vec{g}_1(f_1, \alpha_1), \cdots, \vec{g}_M(f_M, \alpha_M)] \tag{7}$$

In the formant tracking problem, we are interested in finding $p(F_t|Y_t)$ and $\hat{F}_t = \underset{F_t}{argmax} \quad p(F_t|Y_t)$, where $F_t \triangleq \vec{f}_{0:t}$ and $Y_t \triangleq \vec{y}_{1:t}$.

3.1. Review of Particle Filtering: Sequential Importance Sampling (SIS) and Resampling(SIR) [8]

By sampling technique, it is to approxmate $p(F_t|Y_t)$ by $\hat{p}(F_t|Y_t)$:

$$p(F_t|Y_t) \approx \hat{p}(F_t|Y_t) = \sum_{i=1}^{N_s} \tilde{w}_t^i \delta(\vec{f}^i - \vec{f}_s^i) \tag{8}$$

where N_s is the number of samples

Obviously, as N_s goes to infinity, $p(F_t|Y_t)$ can be approximated by $\hat{p}(F_t|Y_t)$ arbitrarily well. The idea of important sampling is to sample from a easy-to-sample function $q(F_t|Y_t)$, compare it to $p(F_t|Y_t)$ at sample point, and scale $q^i(F_t|Y_t)$ to find normalized weight $\tilde{w}_t^i$ to approximate $p^i(F_t|Y_t)$ at sample point (particles) i ($i = 1, 2, \cdots, N_s$).

Knowing that it is hard to sample from $p(F_t|Y_t)$, [8] provided a way to circumvent this difficulty by sampling from a easy-to-sample function $q(\vec{f}_t|\vec{f}_{0:t-1}, \vec{y}_{1:t})$ for the Markov model in Eq.4 and Eq. 6. Some important equations were given below, for detailed derivations please see [8].

$$q(\vec{f}_t|\vec{f}_{0:t-1}, \vec{y}_{1:t}) \mathrel{\overset{\circ}{=}} p(\vec{f}_t|\vec{f}_{t-1}) \tag{9}$$

$$w_t^i = w_{t-1}^i \frac{p(\vec{y}_t|\vec{f}_t^i)p(\vec{f}_t^i|\vec{f}_{t-1}^i)}{q(\vec{f}_t|\vec{f}_{0:t-1}, \vec{y}_{1:t})} = w_{t-1}^i p(\vec{y}_t|\vec{f}_t^i) \tag{10}$$

$$\tilde{w}_t^i = \frac{w_t^i}{\sum\limits_{j=1}^{N_s} w_t^j} \tag{11}$$

$$w_0^i = p_0(\vec{f}^i) \tag{12}$$

where p_0 is the prior of $\vec{f}$, $\quad i = 1, 2, \cdots, N_s$

It has been proved in [9] that the variance of the important weights $\tilde{w}_t$ increased stochastically over time in SIS. To avoid the degeneracy of SIS simulation method, sequential importance resampling (SIR) was proposed [10]. By SIR, region with high probability were sampled more frequently than region with low probability. A more refine method to improve the quality of resampled samples is to implement a Markov chain Monte Carlo (MCMC) step after the SIR step.

3.2. Bayesian Formant Tracking

To solve the problem by Particle Filtering, the first thing that we need to figure out is how many particles we need for the formant tracking problem. Assume that

1. Four formants will be tracked,
2. The particles were put based on the mel-frequency scale,
3. According to (2) and constrained that $Fmt_1 > Fmt_2 > Fmt_3 > Fmt_4$ and the distance between any adjacent formants is at least 300 Hz.

Supposed that the phoneme size is 43, and 30 formant particles for each phoneme. We can put 7 samples for Fmt_1 in the range of 300 Hz to 1200 Hz, 8 samples Fmt_2 is in the range of 800 Hz to 3000 Hz, 6 samples Fmt_3 is in the range of 1600 Hz to 3700 Hz, and 4 samples Fmt_4 is in the range of 3200 Hz to 4600 Hz. Then for all the combinations, we have $7 \times 8 \times 6 \times 4 = 1344 \approx 2^{10.4}$ particles to approximately uniformly sampled the formants subspace. [11] At any given frame t, the posterior probability only concentrated in a small region T known as its typical set T, whose volume is given by $|T| \approx 2^{H(\vec{f}_t|\vec{f}_{t-1})}$, where $H(\vec{f}_t|\vec{f}_{t-1}) \approx log_2 4^4 = 8$ is the Shannon-Gibbs entropy of the probability distribution $P(\vec{f}_t|Y_t)$, then the number of samples required to hit the typical set once is this of order $R \approx 2^{N-H} \approx 4$. Then if we want to hit the typical set 10 times, approximately 40 particles are needed. Considering the sampling of $\vec{a}lpha$, approximately 160 particles is enough. So the particle size is manageable for our formant tracking problem.

Given Eq. 6, likelihood function $p(\vec{y}_t|\vec{f}_t, I)$ is (where I is the prior information of the formant frequency):

$$p(\vec{y}_t|\vec{f}_t, I) \propto \sigma^{-n} exp(-\frac{nQ}{2\sigma^2}) \tag{13}$$

$$\text{Where} \quad Q = \frac{1}{n}\|\vec{y}_t - C(\vec{f}_t, \vec{\alpha}_t)\vec{\varsigma}_t\|_2^2 \tag{14}$$

In our experiment, $\sigma_f = 50Hz$, $\sigma_\alpha = 0.05$, and particles size ($N_s = 150$). An example of formant tracking results were shown in Fig 3.2 and Fig 3.2 . Prior distribution of formants is uniform.

From the experiment, we shown that particle fitlering approach is able to given useful result of the formant for uniform prior on formants. It is obviously that the precision of the result will depend on number of particles, and with a much informative informative prior on formants, this method will be able to give more accuarate result. The next experiment will be phoneme base formant tracking.

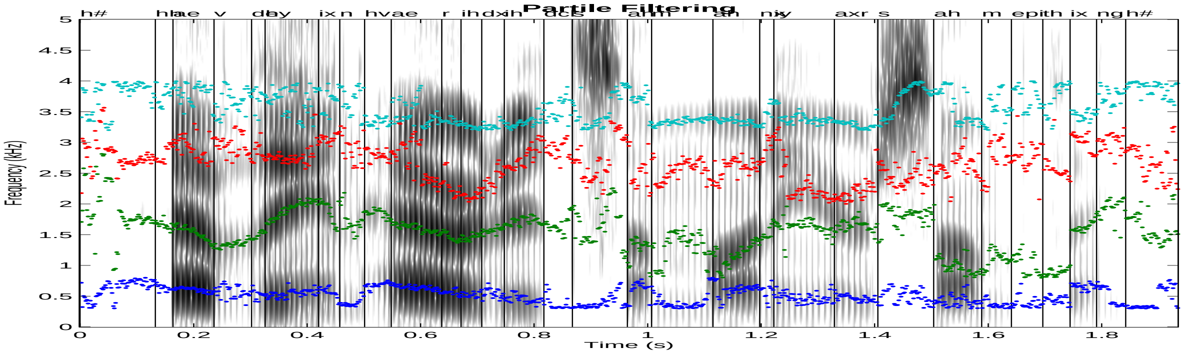
4. References

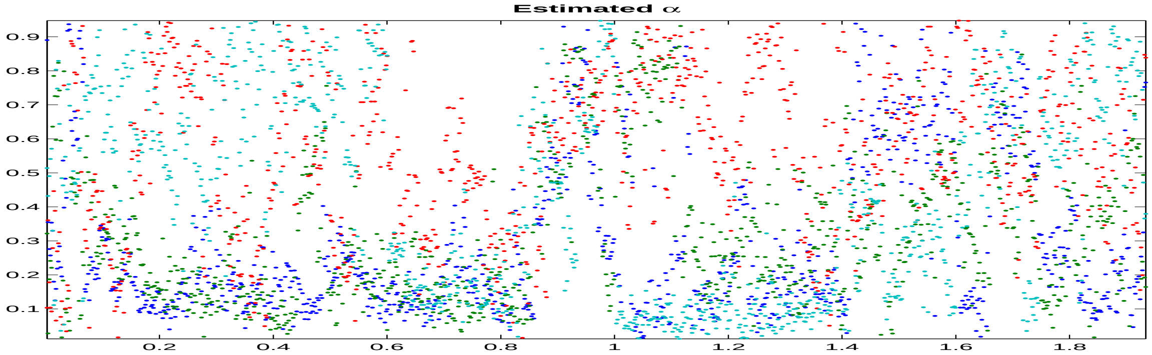
[1] Chomsky, N. and Halle, M., *The Sound Pattern of English*, Harper and Row, New York, NY, 1968.

[2] Browman, C. and Goldstein, L., “Articulatory Phonology: An Overview,” *Phonetica* 49:155-180, 1992.

[3] Stevens, K., *Acoustic Phonetics*, MIT Press, Cambridge, MA, 1999.

[4] Hasegawa-Johnson, M., *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification*, unpublished Ph.D. thesis, MIT, 1996.





[4] Deng, L. and Ma, J., “Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics,” *J. Acoust. Soc. Amer.*, Vol. 108, 2000, 3036-3048 .

[5] Togneri, R., Ma, J., and Deng L., “Parameter estimation of a target-directed dynamic system model with switching states,” *Signal Processing* 81 (2001) p975-987.

[6] Zheng, Y. and Hasegawa-Johnson, M., “Acoustic Segmentation Using Switching State Kalman Filter,” *Proc. ICASSP*, 2003.

[7] Rabiner, L. and Juang, B. H., *Fundamentals of speech recognition*, Prentice-Hall International, Inc, 1993

[8] Merwe, R.V., Doucet, A., Freitas, N., Wan, E. “ The unscented particle filter”, Technical Report TR380, Cambridge University Engineering Department

[9] Doucet, A. and Gordon, N.J. (1999), “Simulation-based optimal filter for manoeuring target tracking,” SPIE signal and Data Processing of Small Targets, Vol. SPIE 3809

[10] Doucet, A., “On sequential simulation-based methods for Bayesian filtering,” Technical Report CUED/F-INFENG/TR310, Cambridge University Engineering Department

[11] Mackay, D.J.C., *Introduction to Monte Carlo Methods*,