FORMANT TRACKING BY MIXTURE STATE PARTICLE FILTER

Yanli Zheng and Mark Hasegawa-Johnson

ECE Department, University of Illinois at Urbana-Champaign
{zheng3,jhasegaw}@uiuc.edu

ABSTRACT

This paper presents a mixture state particle filter method for formant tracking during both vowels and consonants. We show that mixture state particle filter model is able to incorporate prior information about phoneme class into the system, which helps the system to find global optimal solutions. Formant frequencies are defined as eigenfrequencies of the vocal tract in this paper, and by exploring this fact using spectral estimation techniques, the observation PDF of the particle filter can be simplified. We show that by using this likelihood function in the importance weights, the system is able to track the formants using a small number of particles.

1. INTRODUCTION

Estimating formant frequencies during consonant closure is difficult for two reasons. First, consonant spectra are characterized by both poles and zeros, thus the autoregressive (AR) spectral model typically used for formant estimation during vowels is theoretically inapplicable during consonant production. Second, when a spectral zero is close to the frequency of a pole, pole-zero cancellation occurs, and the formant frequency becomes unobservable. The first problem can be solved using an iterative autoregressive moving-average (ARMA) spectral estimation algorithm, provided that the number of spectral zeros is known, and provided that pole-zero cancellation has not occurred [6]. The second problem has not been addressed in the formant tracking literature. In practice, most formant frequencies are canceled by zeros most of the time during production of consonants [10], thus ARMA spectral estimation is not sufficient for accurate formant estimation during consonant closure.

A hidden dynamic model of speech production in order to estimate formant frequencies during consonant closures has been proposed in [11]. In [11], we show that a generic particle filter method can be used to track formants and their amplitudes. In this paper, we propose a mixture state particle filter, which incorporates prior knowledge about 10 phoneme classes in order to effectively sample the typical space of each phoneme class; and by using a problem specific optimized likelihood function, we show that the number of particles can be greatly reduced.

The article is organized as follows. Section 2 reviews the EWAR model, and introduces the problem specific likelihood function. Section 3 demonstrates the mixture state particle filter method. Section 4 reviews conclusions.

2. EWAR MODEL AND LIKELIHOOD FUNCTION

2.1. EWAR Model [11]

Because of the difficulty of estimating spectral zeros, as well as the apparent inability of human listeners to perceive spectral zeros, we proposed to model the spectra of both vowels and consonants using an exponentially-weighted autoregressive (EWAR) spectrum [11]. The EWAR model is not a physical model of speech production; rather it is a function capable of accurately representing the amplitudes and frequencies of poles in an ARMA spectrum without explicitly modeling zeros. Specifically, assume that the vocal tract transfer function can be modeled by the following function:

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

where $\sigma_m = \pi \frac{b_m}{f_s} + j2\pi \frac{f_m}{f_s}$ is the complex frequency of the mth formant, b_m is the formant bandwidth measured in Hz, f_m is the formant frequency in Hz, f_s is the sample frequency in Hz and M is the number of formants to track, and η_m is a coupling coefficient that models inaccuracies in the all-pole spectral model. During vowel production, $\eta_m \approx 1$. During consonant production, $\eta_m \approx 1$ for fully excited formants, but $\eta_m \approx 0$ for formants canceled by spectral zeros or by nulls in the excitation.

Given equation 1,the log spectrum $\log T(z)$ is the sum of $2M$ different terms of the form $\log(1-y)$, where x takes

This work was supported by NSF award number 0132900. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF.

the values of $e^{-\sigma_m}z^{-1}$ and $e^{-\sigma_m^*}z^{-1}$.Using the standard Taylor expansion of $\log(1-y)$, we obtain:

$$logT(z) = logG + \sum_{n=1}^{\infty} c_n z^{-n} \quad (2)$$

$$c_n = \sum_{m=1}^{M} 2\eta_m \frac{e^{-\pi n \frac{b_m}{f_s}}}{n} cos(2\pi n \frac{f_m}{f_s}) \quad (3)$$

Writing c_n in the vector form and taking the first N elements, we obtain:

$$\vec{c} = 2\sum_{m=1}^{M} \eta_m \vec{g}_m \quad (4)$$

where $\vec{g}_m$ is the cepstral component corresponding to a single complex pole pair at frequency f_m with bandwidth b_m, i.e.,

$$\vec{g}_m = [e^{-\pi \frac{b_m}{f_s}} cos(2\pi \frac{f_m}{f_s}), \cdots, \frac{1}{N} e^{-\pi N \frac{b_m}{f_s}} cos(2\pi N \frac{f_m}{f_s})]^T \quad (5)$$

Equations 4 and 5 are similar to the formant decomposition of the cepstrum given in most speech signal processing textbooks (e.g., [8, 9]) and in [3], but with one important difference: each of the formant resonators is scaled by a coupling coefficient η_m. Equation 4 is a closed-form parameterized mapping from the formant frequencies to the cepstrum; the parameters, η_m, allow the model to represent a wide range of speech spectra, from vowels to fricative consonants to silence.

Rabiner [9] notes that the ringing of the cepstrum decays quickly, and that therefore, only a small number of cepstral coefficients contain information relevant to the task of formant frequency discrimination. In our experiment, N=16 and $f_s = 16KHz$

2.2. Likelihood Function

Assuming that the 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:

$$\vec{f}_t = \vec{f}_{t-1} + \vec{v}_{t-1}, \quad \vec{v}_{t-1} \sim N(0, \Sigma_f) \quad (6)$$

$$\vec{y}_t = C_t \vec{\eta}_t + \vec{e}_t, \quad \vec{e}_t \sim N(0, \sigma_y^2 I) \quad (7)$$

where $C_t = [2\vec{g}_1(t), \cdots, 2\vec{g}_M(t)]$,$\vec{f}_t = [f_1(t), \cdots, f_M(t)]^T$, $\vec{\eta}_t = [\eta_1(t), \cdots, \eta_M(t)]^T$, and $\vec{g}_m$ was defined in Equation 5.

We drop the $\vec{b}_t$ here for two reasons: first, the likelihood function which we will use in the importance weights is insensitive to the accuracy of the bandwidth, so we set $b_m = 100Hz$ for all formant frequencies; second, EWAR model is able to capture the amplitude of formants by η in Eq. 1.

The likelihood function of the observation is defined as follows, (notation of t has been dropped):

$$p(\vec{y}|\vec{f},\sigma) = \int d\eta_1 \cdots d\eta_M \frac{1}{(2\pi\sigma^2)^{N/2}} e^{[-\frac{1}{2\sigma^2}||\vec{y}-C(\vec{f})\vec{\eta}||^2]} \quad (8)$$

In the following, we show that marginalization techniques proposed in [1] can be used to simplify the likelihood function in Eq. 8.

By SVD, let $C = \Psi\Lambda\Gamma^T$, $\vec{y}$ can be represented as linear combinations of orthogonal basis functions (i.e. columns in Ψ). By changing the base, Eq. 8 can be written as follows:

$$p(\vec{y}|\vec{f},\sigma) = \int \cdots \int d\gamma_1 \cdots d\gamma_M \frac{1}{(2\pi\sigma^2)^{N/2}} e^{[-\frac{1}{2\sigma^2}||\vec{y}-\Psi\vec{\gamma}||^2]} \quad (9)$$

where

$$\vec{\gamma} = \Lambda\Gamma^T\vec{\eta}$$

By assuming uniform distribution of γ_m and integrating out γ_m, Eq. 9 can be simplified as follows [1]:

$$p(\vec{y}|\vec{f},\sigma) \propto \sigma^{-N+m} e^{-\frac{1}{2\sigma^2}(||\vec{y}||^2 - ||\vec{h}||^2)} \quad (10)$$

$$\vec{h} = \Psi^T\vec{y}$$

And given C, σ will be estimated as follows [2]:

$$\hat{\sigma}^2 = \frac{\vec{y}^T M \vec{y}}{N-1} \quad (11)$$

where $M = C(C^TC)^{-1}C^T$.

The likelihood function in Eq. 10 is very selective of the target formant and very insensitive to the choice of bandwidth and time decay factor $1/n$. An example of normalized likelihood of a simulated signal with a singal complex pole pair at 2000 Hz and bandwidth 250 Hz is shown in Fig. 1.

3. FORMANT TRACKING BY MIXTURE STATE PARTICLE FILTER

A method for formant tracking using a generic particle filter is presented in [11]. In this section, we show how to solve the problem by using a mixture state particle filter.

3.1. Mixture State Particle Filter

In addition to time continuity constrains on the formants enforced by Eq. 6, a discrete phone-dependent state s_t is introduced. The discrete state further constrains the dynamics

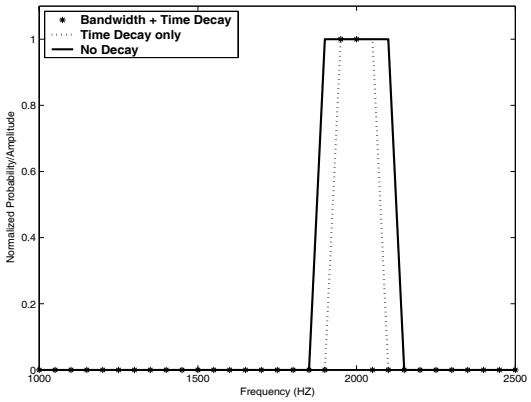

Fig. 1. Illustration of normalized likelihood function for cepstrum of one single frequency at 2000 Hz with 250 Hz bandwidth. For all three cases, the maximum value of likelihood or periodogram is normalized to one. Star line is the likelihood derived from base function with bandwidth 100 Hz and time decay $1/n$, dotted line is the likelihood derived from base function with bandwidth set to 0 Hz and with time decay, solid line is the likehood derived from base function with only cosine terms.

of the formants. The formulation of the problem is as follows:

$$Pr(s_t = k|s_{t-1} = j) = T_{jk} \tag{12}$$

$$P(\vec{f}_t|s_t) \sim N(\vec{\mu}_s, \Sigma_s) \tag{13}$$

$$\vec{f}_t(s_t) = \vec{f}_{t-1}(s_{t-1}) + \vec{v}_{t-1}, \quad \vec{v}_{t-1} \sim N(0, \Sigma_f) \tag{14}$$

$$\vec{y}_t = C_t\vec{\eta}_t + \vec{e}_t, \quad \vec{e}_t \sim N(0, \sigma_y^2 I) \tag{15}$$

The mixture state particle formant tracking algorithm is as follows:

1. *Sampling Step:*

FOR $t = 1 : \tau$

a. SIS Step
(Generating samples for each state)
FOR $s = 1 : K$
$\bar{\Sigma}_s = (\Sigma_f^{-1} + \Sigma_s^{-1})^{-1}$

FOR $i = 1 : N_s$
$\bar{\mu}_s(t) = \bar{\Sigma}_s[\Sigma_f^{-1}\vec{f}_{t-1}^{(i)} + \Sigma_s^{-1}\vec{\mu}_s]$

Sample $f_t^{(i,s)}$ from $N(\bar{\mu}_s(t),\bar{\Sigma}_s)$

$l_t^{(i,s)} \propto p(y|f_t i, s, \sigma)T_{s_{t-1}^{(i)}k}$
END

END

(Choose the particles from K candidates)
FOR $i = 1 : N_s$
$s^* = \underset{s=1,2,\cdots,K}{argmax}\, l_t^{(i,s)}$
$f_t^{(i)} = f_t^{(i,s^*)}$
END

b. SIR Step [7]
Compute $q_t^i = \frac{l_t^{(i)} q_{t-1}^{(i)}}{\sum_{j=1}^{N_s} l_t^{(i)} q_{t-1}^{(j)}}$

Resampling Step:
Resample N_s times from the discrete distribution $q_t^{(i)}$,generating vectors $\vec{f}_{t|t}^{(j)}$ such that for any j,
$\Pr(\vec{f}_{t|t}^{(j)} = \vec{f}_{t|(t-1)}^{(i)}) = q_t^i$.

END

2. *State Estimation*
FOR $t = 1 : \tau$

$$\hat{f}_t = \frac{1}{N_s}\sum_{j=1}^{N_s} f_t^{(j)}$$

$$\hat{\eta}_t = \arg\min \|\vec{y}_t - C(\vec{f}_t)\vec{\eta}_t\|^2$$

END

3.2. Experiment

Experiments reported in this section use a four-dimensional formant state vector $\vec{f}_t$. $N_s = 20$ particles per time are used. $K = 11$ states are used. Vectors $\vec{f}_t$ are predicted and resampled according to the algorithm given in Section 3.1.

In experiments reported in this section, all covariance matrices are assumed to be diagonal, square roots of the diagonal elements of Σ_f are $[120 \quad 200 \quad 200 \quad 200]Hz$, and square roots of the diagonal elements of Σ_s are $[120 \quad 200 \quad 250 \quad 250]Hz$. The transition, state, and observation probability densities are further limited by the following constraints on $\vec{f}_t$:

1. Formant frequencies are drawn from the following frequency ranges: $f_1 \in [400, 1200]$Hz, $f_2 \in [700, 2100]$Hz, $f_3 \in [1600, 3200]$Hz, $f_4 \in [2800, 4600]$Hz.

2. No two formant frequencies are ever less than 200Hz apart, i.e., $f_m \geq f_{m-1} + 200$Hz.

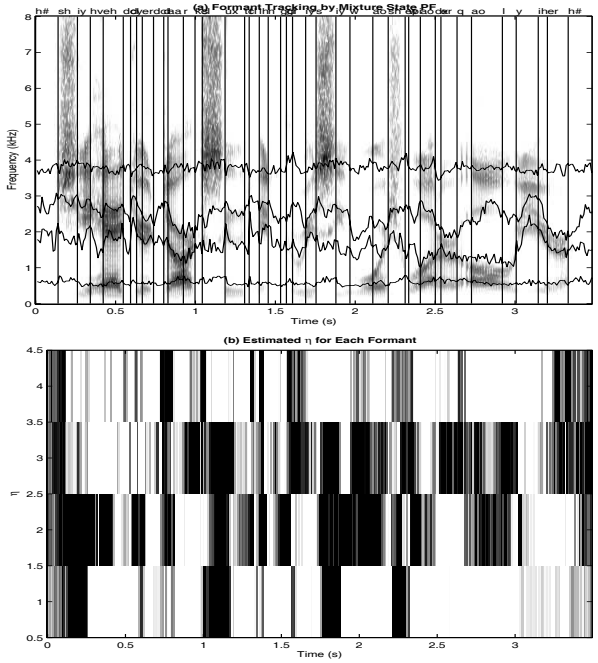

Fig. 2. Dynamic formant tracking results obtained using the mixture particle filtering algorithm. a) Formant tracking result. Vertical line is the boundary between two phonemes and phonemes are labeled at the top.) (b) η for each Formant, where larger η is brighter than smaller one.

3. Ten of the eleven means $\vec{\mu}_s$ in Eq. 13 are set according to the formant frequencies for typical vowels in [9]. And the last state is used as a general term by setting elements of Σ_s to be big.

An example of formant tracking results using mixture particle filter is shown in Figure 2. Figure 2 demonstrates that mixture particle filter is capable to track the dynamics of formant frequency by using a small amount of particles. The advantages of mixture model is obvious. First, it is able to incorporate prior information of the system; second, it offers a better description of complex system than generic particle filter.

4. CONCLUSIONS

Speech synthesis algorithms define the formant frequencies to be the eigenfrequencies of the vocal tract, from larynx to lips, regardless of whether or not the observed acoustic spectrum at any given time contains any evidence for the frequency of a particular formant. Tracking formant frequencies during silences, stops, and fricatives requires the use of an explicit model of formant frequency dynamics. Previous studies have attempted to impose an explicit model of formant frequency dynamics using a discretized formant frequency space with a simplified local observation PDF [5] or an extended Kalman filter with an observation PDF estimated via artificial neural network [4].

The current paper proposes a mixture state particle filter method, incorporating important prior information about the phoneme class, and capable of providing kinematically plausible interpolation of formant frequencies during consonant closure.

5. REFERENCES

[1] G. Larry Bretthorst. *Bayesian Spectrum Analysis and Parameter Estmation*. Springer-Verlag, 1988.

[2] Ronald Christensen. *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer, New York, 2002.

[3] Li Deng, Issam Bazzi, and Alex Acero. Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint. In *Proc. EUROSPEECH*, pages 73–76, 2003.

[4] Li Deng and Jeff Ma. Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics. *J. Acoust. Soc. Am*, 108(6):3036–3048, 2000.

[5] Mark Hasegawa-Johnson. *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification*. PhD thesis, MIT, Cambridge, MA, August 1996.

[6] Lennart Ljung. *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ, 1987.

[7] R.V. Merwe, A. Doucet, N. Freitas, and E. Wan. The unscented particle filter. Technical Report TR380, Cambridge University Engineering Department, 2000.

[8] A.V. Oppenheim, R.W. Schafer, and J. Buck. *Discrete-Time Signal Processing*. Prentice-Hall, New York, 2nd edition, 1999.

[9] Lawrence Rabiner and Bing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[10] Kenneth N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1999.

[11] Yanli Zheng and Mark Hasegawa-Johnson. Particle filtering appoach to bayesian formant tracking. In *IEEE Worksop on Statistical Signal Processing*, 2003.