VOCAL TRACT AREA BASED FORMANT TRACKING USING PARTICLE FILTER

Kaustubh Kalgaonkar & Mark Clements

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332
{kaustubh,clements}@ece.gatech.edu

ABSTRACT

This paper presents a novel method for estimating formant frequencies and bandwidths based on an underlying vocal tract model. A novel statistical model for vocal tract cross-sectional areas is developed which allows computation of full likelihood functions.

Modifications to the basic particle filter algorithm have also been developed to help combat both diversity depletion and convergence problems. The performance of the method is evaluated against hand labeled formant database[1]

Index Terms— Formant Tracking, Particle Filtering, Bayesian Estimation.

1. INTRODUCTION

Formants or vocal tract(VT) resonances play an important role in perception and analysis of speech. Formants are obviously related to the vocal tract shape, but in a one-to-many mapping [2]. Due to the importance of formants, their tracking and estimation has been area of active research for many years. A variety of methods have been proposed, such as those described in references [3, 4, 5]. Many methods are based on linear prediction (LPC), with most of them using either the roots of the LPC polynomial or the smoothed LPC spectrum in some intelligent fashion to estimate peaks and bandwidths. Often, these methods are highly susceptible to merged formants.

Some more modern formant tracking methods [6, 7, 8] construct a state space model for the speech. Due to the presence of non-linearities in such models, where the variables are formants and LPC representations, direct application of Kalman filters and their derivatives becomes difficult. Deng et al. [9] proposed a solution to this problem by linearizing the formant-to-cepstrum transformation, which could then be used to track formants in a Kalman filtering framework.

Particle filters constitute another alternative for tracking formants using the LP model. In references [7, 10] algorithms have been proposed based on particle filtering. In this paper we present a parametric formant tracking method that estimates the state (formant frequency and bandwidth) from noisily observed vocal tract areas.

The paper is organized as follows: Section 2 presents the state-space model. Section 3 discusses the general particle filter and selection of a suitable likelihood, followed by the algorithm and results in Section 4 and Section 5 respectively

2. NON-LINEAR STATE SPACE MODEL

Before presenting the detailed evolution of the state space model, the use of the vocal tract area function as the the observed variable will be discussed and justified.

2.1. Modeling the Vocal Tract Areas

Formants are resonances of the vocal tract, whose parameters are largely determined by the cross-sectional area profile of the vocal tract. Since various configurations of the vocal tract can produce similar resonances, mapping of formants to area functions appear to be be an inappropriate model. This many-to-one mapping between VT areas and formants discourages widespread adoptions of VT areas in speech applications.

Kalgaonkar and Clements in [2] present a method to estimating vocal tract areas. This method of VT area estimation tries to combat the problem of many-to-one mapping by constraining the VT areas to improve their temporal and spacial predictability/smoothness over the ones obtained by traditional method. All the analysis presented in this paper was performed on areas obtained using this method.

We performed a detailed statistical analysis for the VT area functions of various speaker's both male and female. This analysis was performed on the data from both WSJ and TIMIT corpus[11, 12]. Analysis revealed that the VT areas $\mathbf{s} = [s_1, s_2, ..., s_{k+1}]^T$ with $\sum s_i = 1$ can be modeled by a Dirichlet distribution (1) with parameter $\boldsymbol{\alpha}$ which is independent of the speaker and the gender.

$$p(\mathbf{s}; \boldsymbol{\alpha}) = \frac{1}{\mathbf{B}(\boldsymbol{\alpha})} \prod_{i=1}^{k} s_i^{\alpha_i - 1} \tag{1}$$

where $\mathbf{B}(\boldsymbol{\alpha})$ is a multinomial beta function.

A Dirichlet distribution is a way of distributing single unit of a quantity into K pieces where each piece has a mean $\alpha_k \cdot \alpha_0^{-1}$ and variance proportional to the $\alpha_k(\alpha_0 - \alpha_k)$, where $\alpha_0 = \sum_{i=1}^{K} \alpha_i$.

This modeling of vocal tract leads to an interesting insight: to produce a certain set of resonance we draw K tubes from a distribution each tube has its own mean area and variance.

In order to use the VT areas as an observation variable it is extremely necessary to be able to measure and quantify the match/mismatch between two area vectors $\varrho(\mathbf{s}_1, \mathbf{s}_2)$. The mismatch between two area vectors, is the sum of the mismatch in area of each tube. For a human speaker as Dirichlet parameter $\boldsymbol{\alpha}$ is constant both $\boldsymbol{s}$'s are are iid's, so the error in each tube is governed by the variance of each tube.

It is difficult to visualize and explain the difference measure $\varrho(\cdot)$ in M^{th} dimension Dirichlet random variable. We will use the beta random variable to describe the function ϱ since Dirichlet distribution is a multivariate generalization of beta distribution.

Dirichlet random variable with dimension of 2 is beta distributed. Further, since $s_1 + s_2 = 1$ we will only need to consider one dimension (s_1) for the beta case.

Difference of two beta distributed random variables x and y is another random variable z. The distribution of z can be estimated

from the distributions of the beta random variables [13].

$$f_z(z) = f_x(-z) \otimes f_y(z) \tag{2}$$

Although there is no closed-form representation of (2) for the case of beta distributed random variables, an interpretation will be given. Since $0 < (x, y) < 1$, $z \in [-1, 1]$ is obvious. Also the two variables x and y have maximum matching if and only if $z = 0$, and they will have maximum mismatch at the extremes. The distribution of z will be proportional to the variance of the beta distribution. So $\varrho(\cdot)$ must be symmetric and uni-modal with maximum at 0 to measure the level of matching between x and y

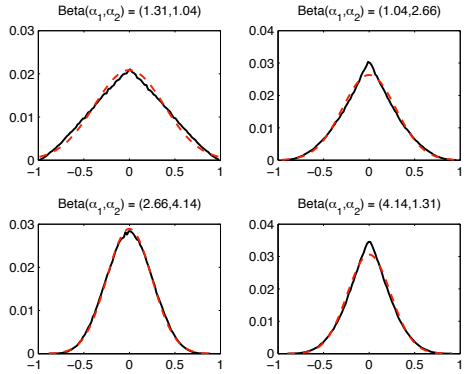

Fig. 1. Matching functions ϱ and Beta parameter

Figure 1 shows the plot of the distributions of z. It can be seen that the shape of the matching function is governed by variance (α) for those dimensions. The red dotted line is a Gaussian fit to the matching function ϱ. The variance of the Gaussian is related to the inverse of the variance of beta random variable.

Extending the above observation to Dirichlet distribution with M dimensions, the matching function $\varrho(\mathbf{s}_1, \mathbf{s}_2)$ for a Dirichlet distributed variable is also uni-modal with maximum at zero which can be modeled by zero mean multivariate Gaussian $\mathcal{N}(\mathbf{s}_1 - \mathbf{s}_2; \underline{0}, R)$. Where R is the diagonal covariance matrix with diagonal entries proportional to variance of the respective Dirichlet dimension.

$$\varrho(\mathbf{s}_1, \mathbf{s}_2) = \frac{1}{(2\pi)^{\frac{N}{2}} |R|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2} \mathbf{x}^T R^{-1} \mathbf{x} \Big) \text{ for } -1 < x < 1 \tag{3}$$

where $\mathbf{x} = \mathbf{s}_1 - \mathbf{s}_2$

2.2. State and Observation Model

Since tracking the formants is the primary interest,the k^{th} resonance of the vocal tract is modeled with a second order digital resonator parameterized by center frequency f_k and bandwidth b_k in Hz. Assuming that M resonators are sufficient to model the spectral envelope at time t, the spectrum can be completely specified by the state vector $\phi_t = [f_1, f_2, ..., f_M, b_1, b_2, ..., b_M]^T$. Given a short frame interval of 10 – 20 ms we can represent the state evolution with equation (4)

$$\phi_{(t+1)} = \phi_t + \mathbf{v}_t \tag{4}$$

where $\mathbf{v}_t \in \mathbb{R}^{2M}$ is additive white Gaussian Noise with $E[\mathbf{v}_i \mathbf{v}_j] = V\delta_{ij}$ and V is the process noise covariance matrix. The process noise perturbs the state from previous time instance to obtain current state.

We observe $\mathbf{s} \in (\mathbb{R}^+)^{M+1\times 1}$ vocal tract area function vector with $\sum s_i = 1$. The area function is obtained form the PARCOR's using the recursion (5)

$$s_m = \frac{1 - r_m}{1 + r_m} s_{m-1} \tag{5}$$

The relationship between the observation vector $\mathbf{s_t}$ and the state vector ϕ_t is nonlinear, as shown in (6)

$$\mathbf{s}_t = k2area\Big(step\ down\Big(\prod_{m=1}^{M}(1 - p_m z^{-1})(1 - p_m^* z^{-1})\Big)\Big) \tag{6}$$

$$p_m = \exp\big(\frac{-b_m \pi}{f_s}\big) \exp\big(\frac{-j2\pi f_m}{f_s}\big) \tag{7}$$

Where p_m is the location of the complex pole and p_m^* is the complex conjugate of p_m and f_s is the sampling frequency. The function $step\ down(\cdot)$ converts the LP polynomial to PARCOR's. $k2area(\cdot)$ converts the PARCOR's to VT area. This whole process of converting formant bandwidth and frequencies to VT area is indicated by function $g : \mathbb{R}^{M\times 1} \rightarrow (\mathbb{R}^+)^{M+1\times 1}$

With this information, the observation model can be specified with Equation (8)

$$\mathbf{s}_t = g(\phi_t) + \mathbf{u}_t \tag{8}$$

Equations (4), (8) provide the framework to implement Bayesian estimation.

3. PARTICLE FILTER FOR FORMANT TRACKING

The observation model in this case is nonlinear and the traditional technique of Kalman filter cannot be straightforwardly applied to track the state. Particle filtering is one technique for implementing a recursive Bayesian estimation filter by Monte Carlo simulation. A key idea in particle filter is to represent the posterior probabilities by a set of random samples and weights, which are then used to compute the estimate.

3.1. General Particle Filter

Some definitions will be presented before going into the details of particle filters. $S_t = [\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3, ..., \mathbf{s}_t]$ represents collection of observations (vocal tract areas) up to time instant t

The estimation and tracking of formants from the Bayesian perspective is that of recursively calculating a conditional estimate.

$$\widehat{\phi_t} = E[\phi_t | S_t] = \int \phi_t p(\phi_t | S_t) d\phi_t \tag{9}$$

To track the state it is necessary to evaluate the posterior pdf $p(\phi_t|S_t)$ which, due to the presence of non-linearity in our case is not available in a closed form.

Particle filters try to approximate the posterior probabilities by using a weighted set of particles $\{(w_t^n, \phi_t^n) : n = 1, 2, \ldots, N\}$, were ϕ_t^n are N particles associated with weights w_t^n . The weights w_t^n approximate the posterior probability time t:

$$p(\phi_t|S_t) \approx \sum_{n=1}^{N} w_t^n \delta(\phi_t - \phi_t^n) \quad s.t. \quad \sum_{n=1}^{N} w_t^n = 1 \tag{10}$$

where

$$w_t^n \propto \frac{p(\phi_t|S_t)}{q(\phi_t|S_t)} = w_{t-1}^n \frac{p(\mathbf{s}_t|\phi_t^n) p(\phi_t^n|\phi_{t-1}^n)}{q(\phi_t^n|\phi_{t-1}^n, \mathbf{s}_t)} \tag{11}$$

The *importance function* $q(\cdot)$ [14] affects the values of weights and consequently the posterior. The performance of the algorithm depends greatly on the choice of the importance function. For the

current experiments we choose to implement the *condensation algorithm* [14], which is a particle filter that uses transition priors as the importance function.

$$q(\phi_t^n|\phi_{t-1}^n, \mathbf{s}_t) = p(\phi_t^n|\phi_{t-1}^n) \tag{12}$$

This particular choice of the importance simplifies the weight calculations (11) to

$$w_t^n \propto w_{t-1}^n p(\mathbf{s}_t|\phi_t^n) \tag{13}$$

Equation (13) can be easily implemented. It is a recursion that uses the past weights and the current likelihood to estimate the weights and the posterior of the particles in current state.

The key requirement in evaluating (13) is the likelihood $p(\mathbf{s}_t|\phi_t)$. The likelihood should available in the form that can be evaluated at a point $(\phi_t^n, \mathbf{s}_t)$

3.2. The Likelihood function

The likelihood $p(\mathbf{s}_t|\phi_t^n)$ in this particular case indicates the *belief* that a given area function $\mathbf{s}_t$ was generated by a particle/state ϕ_t^n. This belief is an indicator of matching between $\mathbf{s}_t$ and $g(\phi_t^n)$. In Section 2.1 we described matching function between two vocal tract areas, and normalized form of this function(3) can be used as a likelihood.

With the last piece in place the Bayesian estimate of the state given by Equation (9) can be evaluated:

$$\widehat{\phi_t} = \int \phi_t \sum_{n=1}^{N} w_t^n \delta(\phi_t - \phi_t^n) d\phi_t = \sum_{n=1}^{N} w_t^n \phi_t^n \tag{14}$$

4. ALGORITHM

The previous section discussed the theory behind estimating of the state which comprise, formants in our case. Particle filtering cannot be directly applied as described in the previous section, unless certain constraints are placed on the state. These constraints which, improve the performance of the filter are listed and explained in this section.

4.1. Particle Generation

A particle filter can be thought of as a statistical search procedure. This algorithm generates random instances of state and tries to match them to the current observation. The goodness of the search depends upon being able to map the state space. The better the map faster the convergence will occur.

The particles or the state ϕ have unknown probability distribution that makes mapping of the state space difficult task. In the current implementation, the state space is divided into manageable blocks. Here, since the state consists of the frequency and bandwidth of the formants, the frequency space $(f_m, 0.5f_s)$ is dividend into M overlapping blocks tracking M formants. Each formant can only occupy a specific range. Further the formant frequencies in ϕ_t^n are also constraint to obey $f_m < f_1 < f_2 < \ldots < f_M < 0.5f_s - \Delta$ where f_s is the sampling frequency and f_m is the lower limit on formant frequency. This constraint will prevent state f_i from accidently tracking another formant f_j. Each formant frequency is uniformly drawn from its section. Formant bandwidths are uniformly drawn from the range (b_l, b_h)

4.2. Algotithm

Draw N state particles ϕ_t^n.
Assign each particles a weight $w_t^n = N^{-1}$

Iterate

1. Use the state propagation model (4) to generate new set of particles from the old set. Generate VT area for current frame of speech($\mathbf{s}_t$).
2. Measure weight of each particle.
$$w_t^n = w_{t-1}^n \varrho(\mathbf{s}_t, \phi_t^n) : n = 1 : N$$
3. Normalize weights.
4. Estimate $\widehat{\phi_t}$ using (14)
5. If ($N_{eff} < N$) Resample particles. Assign weight $w_t^n = N^{-1}$
6. Repair state if particle diversity has depleted.

It is necessary to perform resampling step 5 to prevent degeneracy of the particle filter. Resampling is a process in which particles are replicated/selected with replacement, and the number of times a particle is replicated depends upon its weight. We use the residue resampling technique described in[15].

Step 6 in the algorithm is generally not part of a traditional particle filtering setup. This step maintains state diversity. The difficulty in tracking the formants after rapid changes in their trajectory is due to the depletion of particle diversity. Since formants frequencies are usually slowly varying, all the particles will gather around the true state eventually. However, any sudden change in the format location will not be effectively tracked, as there will be no particles present in that new formant location. During the next few iterations, particles will start moving toward the new formant location, with the speed of convergence depending on the variance of state noise $\mathbf{v}$. To speed up the convergence the variance of the state noise has to be unnaturally large, and number of particle has to be really high. Large number of particles will considerably slow down the filter.

To combat this problem, the state of the particles is "rejuvenated"on such occasions. Diversity of the particles is monitored. If the diversity of the particles is low, newly generated particles are added to recover needed diversity. All the weights are then normalized.

5. EXPERIMENTS AND RESULTS

All the testing was performed using TIMIT copra, WSJ and TIMIT were used for VT area parameter estimation. Hand labeled formant database [1], was used as the 'ground truth' for all error calculations.

5.1. Settings and Empirical parameter Estimation

Parameters for distribution of VT areas were first automatically learned. Approximations of vocal tract cross-sectional areas were extracted from speech data for this estimation. The audio was down-sampled to 8 kHz as only first three formants were to be extracted. Tenth order LPC analysis was performed on 20 ms segments of pre-emphasiszed speech, with a 10 ms frame interval. This model was then used to estimate the VT area function. Parameters for Dirichlet distribution ($\boldsymbol{\alpha}$) were estimated using [16]. The covariance matrix for the Gaussian used for likelihood (3) were derived from $\boldsymbol{\alpha}$

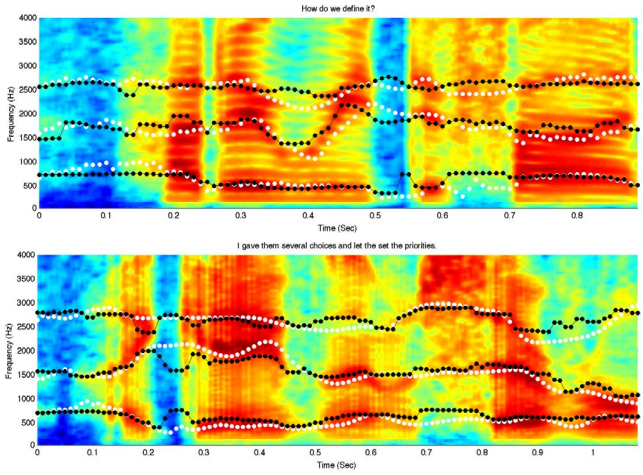

Fig. 2. Formant Tracks white dots - ground truth

The formant frequencies for the state particles are uniformly drawn from there respective sections; $f_1 \in (200, 1000)$Hz, $f_2 \in (800, 2100)$Hz and $f_3 \in (1900, 3500)$Hz. The state noise $\mathbf{v}$ is zero mean white with diagonal covariance. 100 Hz, 200 Hz and 200 Hz were the standard deviations of the state noise for the first, second, and third formants, respectively. All the bandwidths were uniformly distributed in the range $(50, 400)$ Hz and were jittered by white noise with standard deviation 20 Hz

Results are only presented for 538 sentences in TIMIT that have ground truth measurements. 1000 to 1500 particles were used in all the setups.

5.2. Results

Figure 2 shows two formant tracks. The white dots are the ground truth and the black dots are the formant tracks obtained by using the algorithm proposed in this paper. The formant tracks shown here have not been smoothed.

Table 1 shows the root mean square error (RMSE) between the ground truth and the particle filter estimates for all analyzed data. The RMSE is evaluated for two cases with and without voice activity detector (VAD). In the former case the formant error is only evaluated for frames with speech presence.

Table 1. Root Mean Square Error

Formant	No VAD (Hz)	With VAD (Hz)
1	187.19	85.70
2	337.58	197.73
3	395.59	258.58

Although the algorithm does a good job of tracking formants, there are still regions where the estimate does not match the ground truth well. Such behavior usually occurs when two formants are very close. The measurements also show discrepancies with the hand-marked data during the unvoiced sections of the speech, which is largely due to the lack of smoothing and interpolation on formant data on our part. We are currently investigating ways to mitigate this behavior.

6. CONCLUSIONS

In this paper we have presented a new state space model with formants frequencies and bandwidths as the state variables, and vocal tract area function as the observations. Also proposed is a likelihood function that makes the assumption that the vocal tract areas are Dirichlet distributed. Particle filters, with some modification, have been successfully employed to estimate formants from speech data using this model.

To improve the algorithm, current work is focused on methods that dynamically modify formant frequency boundaries, which will improve speed of convergence and reduce the number of particles. Also, effective smoothing algorithms operating on formant trajectories appropriate to this algorithm are being developed.

7. REFERENCES

[1] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *ICASSP*, 2003, pp. 60–3.

[2] K. Kalgaonkar and M. A. Clements, "Vocal tract and area function estimation with both lip and glottal losses," in *Interspeech, Antwerp Belgium*, 2007, p. 550 553.

[3] A. Crowe and M. A. Jack, "Globally optimising formant tracker using generalized centroids," *Electronics Letters*, vol. 23, pp. 1019–1020, 1987.

[4] G. E. Kopec, "Formant tracking using hidden markov models and vector quantization," *Trans. Acoust., Speech, Signal Proc.*, vol. 34, pp. 709–729, 1986.

[5] R. C. Snell and F. Milinazzo, "Formant location from lpc analysis data," *Trans. Speech Audio Proc.*, vol. 1, pp. 129–134, 1993.

[6] G. Rigoll, "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended kalman filter," in *ICASSP*, 1988.

[7] Y. Zheng and M. Hasegawa-Johnson, "Formant tracking by mixture of state particles," in *ICASSP*, 2004, pp. 565–568.

[8] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 6, pp. 36–48, 1998.

[9] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using quantized nonlinear function embedded in a temporal constraint," *Trans. Acoust., Speech, Signal Proc.*, vol. 14, 2006.

[10] Y. Shi and E. Chang, "Sectrogram based formant tracking via particle filters," in *ICASSP*, 2003, pp. 168–171.

[11] D.B. Paul and J.M. Baker, "The design for the wall street journal-based csr corpus," in *ICSLP*, 1992.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," in *Linguistic Data Consortium, Philadelphia*.

[13] A. Papoulis and S. U. Pillai, *Probability, random variables and stochastic processes*, McGraw-Hill, 2002.

[14] M. Isard and A. Blake, "Condensation – conditional density propagation for visual tracking," *Int. J. of Computer Vision*, vol. 29, no. 1, pp. 5–28, 1998.

[15] J. S. Lui and R. Chen, "Sequential monte carlo methods for dynamic systems," *J. American Stat. Assoc*, vol. 93, 1998.

[16] T. Minka, "Estimating a dirichlet distribution," in *www.stat.cmu.edu/~minka/papers/dirichlet*, 2003.