AN EXPECTATION MAXIMIZATION APPROACH FOR FORMANT TRACKING USING A PARAMETER-FREE NON-LINEAR PREDICTOR

Issam Bazzi, Alex Acero, and Li Deng

Microsoft Research
One Microsoft Way
Redmond, WA, USA
{issamb,alexac,deng}@microsoft.com

ABSTRACT

This paper presents a new approach for formant tracking using a parameter-free non-linear predictor that maps formant frequencies and bandwidths into the acoustic feature space. The approach relies on decomposing the speech signal into two components: the first component captures the mapping between formants and acoustic observations, while the second component is intended to capture the residual in the signal. We build the mapping by quantizing the formant space and creating a *predictor codebook*. Formant tracking is achieved by: 1) EM training of the parameters of the residual component, and 2) searching the predictor codebook for the best formant values. We explore both MAP and MMSE methods for performing formant tracking with the proposed approach. Furthermore, we impose first order continuity constraints on formant trajectories, and use Viterbi search to perform formant tracking. We present formant tracking results on data from the Switchboard corpus.

1. INTRODUCTION

The problem of formant tracking has received considerable attention in speech recognition research [1, 2] as formant frequencies are known to be important in determining the phonetic content as well as the articulatory information about the speech signal. For speech recognition, formants can either be used as additional acoustic features [1] or can be utilized as hidden dynamic variables as part of the speech recognition model [3].

While many methods exist for formant tracking from speech waveforms, these methods rely heavily on either: 1) using some type of LPC spectral analysis [2] to compute formant candidates which are then combined with continuity constraints; or 2) matching stored templates of spectral cross sections [1] to the speech signal. In either case, formant tracking is error-prone due to the limited number of candidates from LPC analysis, or the limited number of templates available for comparison.

This paper presents a new method for formant tracking by using a model that decomposes the speech signal into two components. The first component captures the mapping from the formant space into the acoustic measurements (MFCC) space assuming an all-pole model, while the second component captures the residual in the speech signal. We build this mapping by quantizing the formant frequency and bandwidth space and creating a *predictor codebook*. Formant tracking is achieved by searching this codebook for the most suitable set of formant values. We present both MAP and MMSE approaches for formant tracking with and without continuity constraints. Our approach has two key advantages over other approaches: first, the relationship between formant values and their contribution to the acoustic measurement is explicitly represented through the predictor codebook. Second, compared to methods that rely on an LPC analysis or template matching to obtain candidate formants, our approach explores the complete formant space avoiding errors due to premature elimination of formant candidate during the analysis step.

2. THE MODEL

Let o_t be the output observation of the speech signal at time t with dimensionality N. For this paper, o_t represents the standard MFCC coefficients, though other acoustic features would also be possible. We assume that o_t can be decomposed into two additive components:

$$o_t = F(x_t) + r_t \tag{1}$$

where x_t represents the vocal tract resonances (VTR) and their corresponding bandwidths, and $F(x_t)$ is a predictor that maps the VTR space into the acoustic observation space. The following section describes how $F(x)$ is constructed. The second component is r_t which is intended to capture the residual in the speech signal after the VTR contribution is removed with $F(x)$.

2.1. Constructing the Predictor $F(x)$

Let $F(x)$ be represented by an all-pole model as in LPC analysis, and assume that there exists a fixed number I of poles that $F(x)$ can model. For a given I, the VTRs and their corresponding bandwidths are then represented as $x = (F_1, B_1, F_2, B_2, \ldots, F_I, B_I)$. For each pole (F_i, B_i), the corresponding complex root is given by [2]:

$$z_i = e^{-\pi \frac{B_i}{F_s} + j2\pi \frac{F_i}{F_s}} \tag{2}$$

where F_s is the sampling frequency of the input signal. Under the all-pole assumption, the z-transfer function with the I poles, their conjugates, and gain G is:

$$H(z) = G \prod_{i=1}^{I} \frac{1}{(1 - z_i z^{-1})(1 - z_i{}^* z^{-1})} \tag{3}$$

Given the z-transfer function, we can compute the power spectrum and then follow the standard steps for MFCC evaluation as shown in Figure 1: applying the mel-frequency warping, the filter banks, and then the discrete cosine transform to obtain MFCCs.

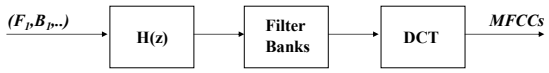

Fig. 1. Generating $F(x)$.

By quantizing x over some range of frequencies and bandwidths, and performing the steps shown for all possible quantized values of x, we can create a table or *codebook* for the mapping $F(x)$. One important property of the obtained mapping is its analytical nature and its independence of any speech data. Depending on the number of formants I and the level of quantization of the formant space, the size of the codebook could be large.

3. EM TRAINING

Let r_t be represented by a single Gaussian with mean $\mu = \mu_1, \mu_2, \ldots, \mu_N$ and covariance $\mathbf{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$. In this section, we describe an EM procedure for training these parameters. Let $\mathbf{o} = (o_1, o_2, \ldots, o_T)$ be the output observation of a speech utterance and $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ be the corresponding formant values, the goal is to estimate the model parameters $\theta = (\mu_1, \mu_2, \ldots, \mu_N, \sigma_1, \sigma_2, \ldots, \sigma_N)$. We assume that the output observations are independent from one frame to another, and that the hidden variable is the set of formant values x. We also assume that the formant values are uniformly distributed over the allowed quantized values of each of the formants and their bandwidths. Given these assumptions, the EM auxiliary function is [4]:

$$Q(\theta, \theta') = E_x[\log\ p(\mathbf{x}, \mathbf{o}|\theta)|\mathbf{o}, \theta'] \tag{4}$$

Under the assumption that the hidden variable x is uniformly distributed, and assuming that x can take any of C quantized values, the auxiliary function reduces to the following:

$$Q(\theta, \theta') \quad = \quad \sum_{t=1}^{T} \sum_{i=1}^{C} p(o_t|x[i], \theta') log\ p(o_t|x[i], \theta) \tag{5}$$

where $p(o_t|x[i], \theta) = N(o_t - F(x[i]); \mu, \Sigma)$. Following the standard steps for solving this EM problem, the auxiliary function leads to the following EM update equations:

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{i=1}^{C} (o_t - F(x[i]))\ N(o_t - F(x[i]); \mu', \Sigma')}{\sum_{i=1}^{C} N(o_t - F(x[i]); \mu', \Sigma')} \tag{6}$$

$$\hat{\Sigma} = \frac{1}{T} \sum_{t=1}^{T} \frac{\sum_{i=1}^{C} (o_t - F(x[i]) - \hat{\mu})^2\ N(o_t - F(x[i]); \mu', \Sigma')}{\sum_{i=1}^{C} N(o_t - F(x[i]); \mu', \Sigma')} \tag{7}$$

4. FORMANT TRACKING

In this section, we present two methods for formant tracking using our speech model and the non-linear predictor $F(x)$. In the first method, formant tracking is done on a frame-by-frame basis with no continuity constraints. In the second method, continuity constraints are added in the form of formant transition probabilities, and tracking is performed using a Viterbi search.

4.1. Frame-by-Frame Formant Tracking

A simple and straight-forward method for formant tracking using our model is to estimate formants for each frame independently. Given an observation o_t, with the uniform assumption over the distribution of x, the MAP estimate reduces to the ML estimate:

$$\hat{x}_{MAP} = \arg\max_i p(o_t|x[i], \theta) \tag{8}$$

Note that estimating $\hat{x}_t$ requires searching $F(x)$ codebook to find the most likely formant estimate. The value of the estimate can only be one of the quantization values used to construct the function. A smoother estimate that allows for continuous formant values is the minimum mean squared error (MMSE) estimate and is given by:

$$\hat{x}_{MMSE} = \frac{\sum_{i=1}^{C} x[i] p(o_t|x[i], \theta)}{\sum_{i=1}^{C} p(o_t|x[i], \theta)} \tag{9}$$

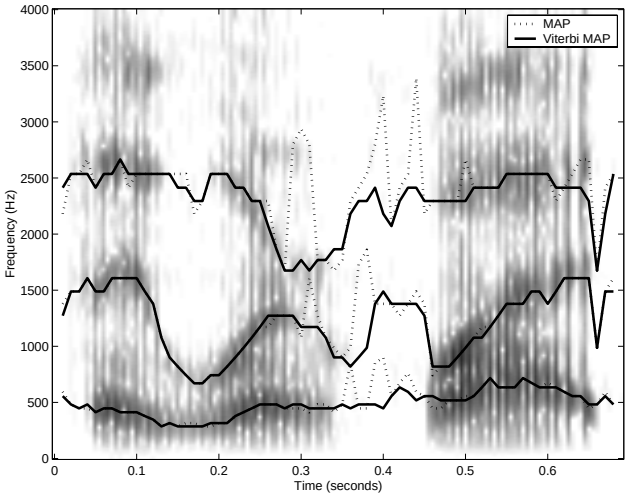

Fig. 2. Formant tracking using the frame-by-frame (dashed) and the Viterbi MAP methods (solid).

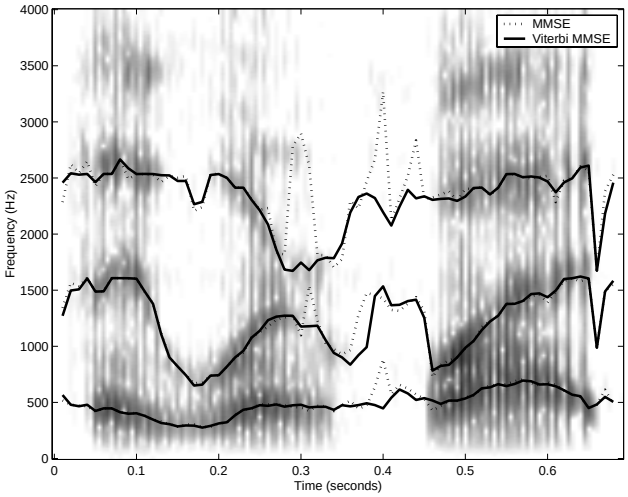

Fig. 3. Formant tracking using the frame-by-frame (dashed) and the Viterbi MMSE methods (solid).

4.2. Formant Tracking with Continuity Constraints

For this method, we use a simple first order state model to impose constraints on how formants can change from one frame to the next [3]:

$$x_t = x_{t-1} + w_t \tag{10}$$

where w_t is modeled as a single Gaussian with zero mean and diagonal covariance Σ_w. With this continuity constraint, formant tracking is performed at the utterance level and the MAP estimate becomes:

$$\hat{\mathbf{x}}_{MAP} = \arg \max_{i_1,\ldots,i_T} \quad \prod_{t=1}^{T} p(o_t|x[i_t],\theta) \; p(x[i_1]) \prod_{t=2}^{T} p(x[i_t]|x[i_{t-1}]) \tag{11}$$

which can be estimated using a standard Viterbi search. For this paper, we do not try to learn the covariance Σ_w of w_t. Instead, we assume that Σ_w is diagonal with variances proportional to the quantization errors of $F(x)$.

The MMSE counter-part of this method involves computing a lattice from the Viterbi search and estimating the Posterior probabilities for various hypotheses in the lattice. For this paper, we explore an approximate MMSE approach where at each node in the lattice, we use a forward-backward Viterbi search to approximate the Posteriors by the maximum probability of going through that node at that particular point in time. These probabilities are then used to obtain an (approximate) MMSE estimate by computing a probability-weighted sum overall all surviving hypotheses in the lattice at every point in time.

5. EXPERIMENTS AND RESULTS

5.1. Constructing $F(x)$

For experiments presented here, we set $F(x)$ to model the first three formants ($I = 3$ in Equation 3). Because we wanted to use $F(x)$ to model all phones, we selected a range of frequencies and bandwidths to cover all phones [5]. Table 1 shows the range for each frequency and bandwidth used and the corresponding number of quantization levels.

	Min(Hz)	**Max(Hz)**	**Num. Quant.**
F_1	200	900	20
F_2	600	2800	20
F_3	1400	3800	20
B_1	40	300	5
B_2	60	300	5
B_3	60	500	5

Table 1. Range of the three formants, bandwidths, and the corresponding quantization levels.

Bandwidths were quantized uniformly while formants were mapped to the mel-frequency scale and then uniformly quantized. The number of quantizations shown in Table 1 yields a total of 1 Million ($20^3 \times 5^3$) entries for $F(x)$, but because of the constraint $F_1 < F_2 < F_3$, the resulting number was 767,500 entries. For evaluating $F(x)$ in the MFCC space, we set the gain $G = 1$ in Equation 3 and we excluded the first MFCC coefficient C_0 from $F(x)$ making the mapping independent of the energy level in the signal. The final dimensionality of $F(x)$ is 12 (remaining MFCC coefficients).

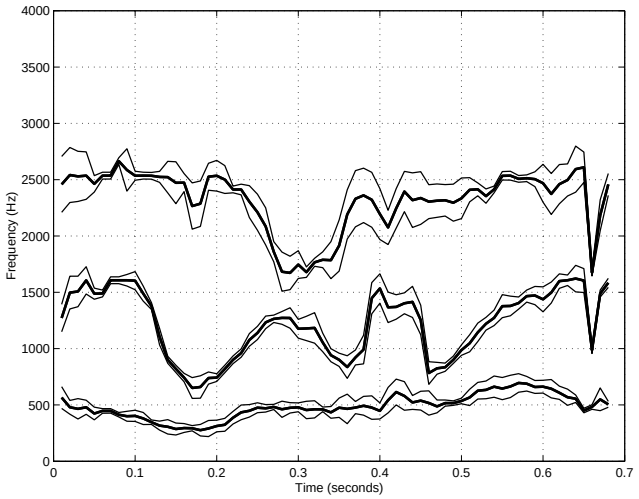

Fig. 4. Bandwidth results with the Viterbi MMSE method. For every formant, we show the formant frequency F (thick curve) and two curves at $F \pm \frac{1}{2}B$.

5.2. Formant Tracking Results

For the EM training, we used 20 utterances from a male speaker in Switchboard. An important property of our approach is that it does not require any data labeling, and EM training can be performed in a fully unsupervised fashion. Figures 2 through 5 shows the results of formant tracking for the Switchboard male speaker uttering the phrase *"they were what"*.

In Figure 2, formant tracking results are super-imposed on the spectrogram. Two formant tracking results are shown: the dotted line shows the frame-by-frame MAP results while the solid line shows the Viterbi MAP results with the continuity constraint. Note that the Viterbi MAP tracking helps most in unvoiced regions where no formants are visible. Figure 3 shows the results with MMSE tracking. The values achieved with this method are continuous and not restricted to the quantized values in the codebook. The MMSE provides for yet smoother formant tracking especially in unvoiced regions. Figure 4 shows the bandwidth results for the same utterance. As expected, the approach uses narrow bandwidths for voiced regions and wide bandwidths for non-voiced regions in the signal. The results shown in these figures are typical of the performance of our approach. Since there is no standard evaluation scheme for formant tracking, we manually examined over 40 utterances from Switchboard and found no *gross errors* in formant tracking, where a gross error is defined as missing a formant or confusing one formant to another. Finally, Figure 5 shows the spectrogram for the residual r_t, where most of the energy due to the first three formants is removed from the signal, an indication of the high accuracy of the approach in estimating both formant frequencies and bandwidths.

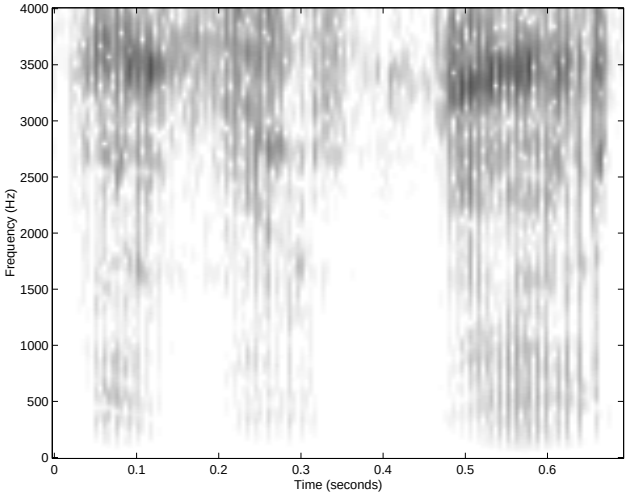

Fig. 5. Spectrogram of the residual r_t.

6. CONCLUSION AND FUTURE WORK

We presented a new approach for automatic formant tracking. By building a parameter-free nonlinear mapping from the formant space to the acoustic measurements space, formant tracking is achieved by searching over all possible quantized values of formants. We showed that formant tracking with first order continuity constraints results in smoother formant tracking. For our future work, we are looking into integrating this approach into our hidden dynamic model approach [3] for speech recognition.

7. REFERENCES

[1] John N. Holmes, Wendy J. Holmes, and Philip N. Garner, "Using formant frequencies in speech recognition," in *Proc. Eurospeech '97*, Rhodes, Greece, Sept. 1997, pp. 2083–2086.

[2] A. Acero, "Formant analysis and synthesis using hidden markov models," in *Proc. of the Eurospeech Conference*, Budapest, September 1999.

[3] L. Deng and Z. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics," *Journal of the Acoustical Society of America*, vol. 108, no. 6, 2000.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[5] J. Allen, M. S. Hunnicutt, and D. Klatt, *From Text toSpeech: The MITalk System*, Cambridge University Press, Cambridge, England, 1987.