A NEW STRATEGY OF FORMANT TRACKING BASED ON DYNAMIC PROGRAMMING

Kun Xia and Carol Espy-Wilson
(xia@vocalpoint.com, espy@bu.edu)

Electrical and Computer Engineering Department, Boston University
8 St. Mary’s St., Boston, MA, USA 02215

ABSTRACT

This paper describes a new method for estimating formant frequencies. It operates in two phases. The first phase, which is similar to a technique developed by Talkin [JASA, vol. 82, S1], finds optimal formant track estimates by imposing frequency continuity constraints using Dynamic Programming (DP). DP is used to select a mapping of candidate frequencies to formant frequencies in oral sonorant regions based on the minimum cost from all possible mappings. The second phase performs a series of postprocessing steps to make formant estimates more robust and accurate and extends the formant estimates into nasal and obstruent regions. Performance statistics comparing the formants obtained with this technique with a set of reference formants using 34 sentences randomly selected from the TIMIT database shows our algorithm gives excellent results when the formants are among the candidate frequencies.

1. INTRODUCTION

The knowledge that the frequencies of the first few formants (i.e. the lowest acoustic resonances of the vocal tract) are primary information carriers in human speech has led many researchers to develop means for estimating formant frequencies from the acoustic signal. An automatic formant estimation algorithm can find applications including: classification of vowel-like phonetic units in a speaker-independent recognition system [1], acquisition of parameters for diphone text-to-speech systems [2] and data reduction for research in acoustic phonetics.

All automatic means for tracking formants solely from a short segment of the acoustic signal sometimes postulate formants which are obviously wrong when viewed in a larger context. The observation that formants are in general slowly varying functions of time has led to attempts to force continuity constraints on the formant selection process. Methods used include non-linear smoothing operations which depend on good estimates in neighbor regions [3], and extension of reliable formant estimates from “anchor frame” found in supposed strong vocalic areas [4,5,6]. Both types of approaches, however, are sensitive to wrong estimates and have a tendency of error-propagation.

Another approach to formant estimation applies HMM type global statistical criteria to find best overall fit of formant envelope to an utterance [7]. Continuity constraints are inherently expressed in the transition probability matrix of a formant model. One advantage of this method is its potential to provide reasonable formant trajectories when formant candidates are missing. The use of ad hoc decision rules and manually adjusted thresholds is also eliminated.

The dynamic programming (DP) approach to speech parameter trajectory estimation [8] offers another dimension of flexibility and tractability. DP has provided excellent performance when applied to fundamental frequency estimation [9]. Talkin also applied DP to formant tracking problem [10]. The formant frequencies are selected from candidates proposed by solving for the roots of the linear predictor polynomial computed periodically from the speech waveform. The local costs of all possible mappings of the complex roots to formant frequencies are computed at each frame based on the frequencies and bandwidths of the component formants for each mapping. The cost of connecting each of these mappings with each of the mappings in the previous frame is then minimized using a modified Viterbi algorithm, where continuity constraints are imposed. This method will usually find reasonable formant estimates when the formant trajectories are apparent, but can make gross errors even in vocalic regions, especially when formant candidates are occasionally missing.

Here we present a new strategy that imports speech segmentation information and integrates a second phase of postprocessing into this DP approach. “Anchor” formant lines are found in oral sonorant regions, which are more reliable than anchor frames and help improve robustness and accuracy of formant estimation.

2. DYNAMIC PROGRAMMING FORMANT TRACK ESTIMATION

Similar to Talkin’s technique [10], the DP algorithm uses the complex roots of the denominator polynomial of the z transform of a linear predictor as the set of formant candidates. Using LPC poles has a number of advantages. The all-pole LP model is a reasonable approximation during vowel-like sounds where formant tracking is most meaningful. By choosing an appropriate predictor order, a usually fixed number of complex poles can be obtained which is a superset of possible formants, thus reducing the problem of merged formants while keeping the number of formant candidates to a minimum.

In our experiments, the speech signal is sampled at 10KHz since we are only interested in finding the first three or four formants in a 5KHz frequency band. A mild first order preemphasis filter of the form $1-0.7Z^{-1}$ is applied to partially compensate for the speech source. An additional non-causal symmetric FIR filter of a unit-minus-Hamming impulse response is applied to strongly attenuate very low frequency components. Autocorrelation LPC
analysis is performed asynchronously at a frame rate of 100Hz using a 49ms duration cos^4 window. The poles of the estimated model are obtained by solving for the zeros of a 12th order linear predictor polynomial. Real poles, which merely contribute to overall spectral slope, are eliminated. The remaining complex poles are then expressed as frequency-bandwidth pairs \([F_i, B_i]\), \(i=1,...,M, M=\frac{P}{2}\), where \(P\) is the LPC order. Thus, for a 10kHz sampling frequency and a 12th-order predictor, five or six candidate formants are typically found at each frame.

To find the best set of trajectories for \(N\) formants through a trellis of candidate frequencies, we minimize the cost of mapping candidate frequencies to formants at each frame over all analysis frames. Figure 1 is a grid form of Viterbi search, where each node represents a mapping of a set of candidate frequencies to formants. The horizontal direction denotes time, and the vertical direction denotes different mappings. The DP cost function is defined as:

\[
C(t,n) = C_{\text{local}}(t,n) + \min_{m} \{ C_{\text{trans}}(t,n),(t-1,m) + C(t-1,m) \} \quad (1)
\]

Where \(C(t,n)\) is the cumulative cost at node \((t,n)\). \(C_{\text{local}}(t,n)\) is the local cost at \((t,n)\), which reflects knowledge about formants without temporal context. Three kinds of knowledge are used in the current algorithm. First, absolute empirical limits on formant frequencies do exist. The first four formants are bounded as follows (in Hz):

100 < \(F_1\) < 1500; 500 < \(F_2\) < 3500; 1000 < \(F_3\) < 4500; 2000 < \(F_4\) < 5000

Second, poles with narrow-bandwidth are more strongly manifested in local spectral energy prominence, and thereby are more likely to be real formants. A quadratic cost function of the bandwidth is assigned to each formant component to penalize large bandwidth poles. Thirdly, formant frequencies tend to fluctuate around some neutral vocal tract values but do change considerably. Thus a mild linear cost is imposed to penalize deviation from neutral values. We give the neutral vocal tract values for the first four formants as follows (in Hz):

\[F_{n1} = 500; \quad F_{n2} = 1500; \quad F_{n3} = 2500; \quad F_{n4} = 3500\]

Therefore, to find the first \(N\) formants, we define \(C_{\text{local}}(t,n)\) as:

\[
C_{\text{local}}(t,n) = \sum_{i=1}^{N} \left[ \alpha_i |B_i| + \beta_i |F_i - F_{n_i}| / F_{n_i} \right] \quad (2)
\]

where \([F_i, B_i]\) is the frequency-bandwidth pair of the \(i\)th component of the mapping at node \((t,n)\).

\(C_{\text{trans}}(t,n),(t-1,m)\) is the cost of transition from node \((t-1,m)\) to \((t,n)\). Since formants vary slowly within phonetic segments, a quadratic cost function of inter-frame frequency change is used to penalize large frequency discontinuities:

\[
C_{\text{trans}}(t,n),(t-1,m) = \sum_{i=1}^{N} \gamma_i (F_i(t) - F_i(t-1))^2 \quad i=1,2,...,N \quad (3)
\]

where \(F_i(t)\) and \(F_i(t-1)\) are the frequencies of the \(i\)th component of the mapping at node \((t-1,m)\) and \((t,n)\) respectively.

The constants \(\alpha_i, \beta_i, \gamma_i\) control the relative weighting of different cost functions for each formant component. They were determined empirically by multiple supervised experiments on 20 sentences from the TIMIT database. More refined values of these constants will be obtained as a larger database is developed. As a general rule, higher order formants have a smaller weight as they have less important effect on overall cost. We chose \(\gamma_i\) to be relatively small since \(F_2\) may change more rapidly than the other formants. For each mapping at each frame, an index to the lowest cost “transition” in the previous frame is maintained. Therefore, at the end of the signal, these indices can be backtraced to obtain the globally lowest cost mappings of candidate frequencies to formants over all frames.

We experimented with this algorithm on some TIMIT sentences and found it showed good performance in speech segments well represented by an LPC all-pole model and when good candidate pole frequencies are available, such as in oral sonorant regions. In nasal and non-sonorant regions, where good candidate frequencies are missing or do not exist at all, formant estimation results can be very unpredictable and may produce gross errors.

Grossly erroneous formant estimates are more disastrous than no estimate when used in speech recognition, and should be avoided even at a high cost. The reason is, in places where formant information is missing, other information (such as the spectral center of gravity in different frequency bands) can be used in speech recognition, whereas wrong formant estimates lead to unrecoverable recognition errors. Based on this philosophy, we proposed some post processing to keep correct estimates and discard errors.

The backtracing in the Viterbi search starts at the end of the speech signal and is carried out over the entire speech signal in [10]. However, formant tracking is most meaningful in vowel-like regions, where tracking results are also most reliable. Formant estimates in vowel-like regions are also helpful in finding out neighboring nasal formants and leakage formants in obstructed regions, where they exist. These observations lead us to the proposed new strategy of using DP Viterbi search coupled with speech segmentation information.

### 3. NEW STRATEGY OF USING DP AND POSTPROCESSING

Formants are fundamentally different in vowel-like regions, nasal regions and obstructed regions. In vowel-like regions formants are usually prominent and all-pole model is a reasonable representation. In nasal regions, formant may be weak due to nearby zeros and an extra nasal formant may be salient. Also it is
common to see a disjoint F1 trajectory on the boundary of a vowel and a nasal region. In obstruent regions, meaningful formants are usually not available. However, formant transitions between the obstruent and the vowel are sometimes observable in the obstruent region. Such formant movement is very helpful in recognition.

Since formant estimation is most meaningful and most reliable in oral sonorant regions, tracking results in these regions should be useful for tracking in other places where formant trajectories are not prominent. Therefore, instead of applying Viterbi search over the entire speech signal, we first segment the speech into three different categories: vowel-like regions, nasal regions and obstruent regions, using a hand transcription or a broad class phonetic recognizer, for example [11]. Following is a classification of all phonemes into these three categories:

vowel-like: iy ih eh ae ix ah ax-h uw uh ao aa ey ay oy aw ow ux er lax r y w l hv dx q
Nasal: m n ng eng nx en em
Obstruent: p t k b d g s z sh zh ch jh th dh f v hh sil h# #h pau pi pcl tcl kcl dcl gcI bcl

Viterbi search and backtracing was carried out separately in vowel-like and nasal regions only, where continuous formant trajectories are expected. Estimated formant lines may still have disjoint components, however, due to missing pole candidates or outlier formant points within the sonorant regions. To solve this ambiguity, a median smoothing is first performed on raw formant estimates to discard outliers. Then each formant trajectory is further divided into concatenative smooth fragments using some continuity criteria. Currently each inter-frame frequency difference is examined against a threshold to determine discontinuity points. The longest continuous segment was found in this manner and was kept as the “anchor formant”. Then a local beam search was performed in both directions starting from the anchor formant, in an attempt to find missed formants within this region and leakage formants in neighboring obstruent regions. Candidate pole frequencies that were within the beam of the anchor formant, at the cost of computational complexity.

This formant tracking algorithm was tested on 34 sentences, each spoken by a different male speaker (see Fig. 1). The speakers were randomly selected from the TIMIT database to provide a good sampling of American dialect. Reference formant data were established using Entropic’s formant tracker [12] making hand corrections to the tracks when needed. The hand transcriptions were used to aid in segmentation of sentences into the three phonetic categories. A broad class recognizer was not used, since the major interest here was to test the performance of the formant tracking procedure without possible degradation due to broad class segmentation errors. The error rate was computed on a per frame bases. Since hand-corrected formant points can never be exactly the same as the formant tracker output, we say an error occurs when the difference of the hand-corrected formant frequency and the formant tracker result differs by more than 200Hz. In places where the tracker produces no output (premature stop in local beam search), a deletion error occurs. The error rate was computed separately for vowel-like and nasal regions. We did not consider obstruent regions since there of often no reliable reference formant data available. Table 1 summarizes the formant tracker’s performance.

<table>
<thead>
<tr>
<th></th>
<th>Sonorant, 5293 frames</th>
<th>Nasal, 520 frames</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>substitution</td>
<td>deletion</td>
</tr>
<tr>
<td>F1(Hz)</td>
<td>0.15%</td>
<td>1.02%</td>
</tr>
<tr>
<td>F2(Hz)</td>
<td>1.09%</td>
<td>1.79%</td>
</tr>
<tr>
<td>F3(Hz)</td>
<td>6.89%</td>
<td>4.14%</td>
</tr>
</tbody>
</table>

This formant tracker works reasonably well in oral sonorant regions. The tracking error is low at the cost of deletion errors, which reflects the underlying philosophy: false estimation is associated with higher cost than missing formants. In nasal regions, the error rates for F2 and F3 are high, generally due to the influence of zeros that make good pole candidates unavailable. Another reason is that nasals usually have a short duration, which makes it more difficult to find reliable anchor formant lines.

Most errors occur around at the boundary of a vowel region and either a nasal or obstruent region, where a low order all-pole model is not a good representation and the location of candidate poles is not stable. Although the anchor formant is helpful in selecting the right candidate poles, there were a few cases where erroneous candidate frequencies were picked by the local beam search because the right ones were missing. One way to alleviate this problem is to increase the order of the LPC analysis adaptively, when the transition cost is high and there is a chance of missing pole candidates. It is possible that a spectral estimation procedure which accounts for both poles and zeros may yield a set of formant candidates containing the “true” formants, at the cost of computational complexity.

Most deletions occur when spectral changes are beyond the range of the local beam search and continuity criteria for segmentation within each region. Instead of doing fixed-value hard thresholding in the local beam search and segmentation, adaptive thresholding modulated by speech spectral stationarity (a weighted cepstral derivative magnitude, for example) can be employed to improve on deletion rates. A more complicated decision structure to handle segmented formant trajectories should also improve accuracy over our present simple use of the longest continuous formant line.

The constants in the cost function can also be modulated by a spectral distance factor to yield improved performance. In places where the spectrum changes rapidly (like around the boundary of a vowel and an obstruent region), formant discontinuity should be less penalized, whereas in stable spectral regions (like in a
stable vowel region) formant continuity should be more strictly enforced.

5. CONCLUSION AND FUTURE WORK

A new strategy of applying Dynamic Programming to formant tracking problem is proposed. A segmentation of the speech signal into vowel-like, nasal and obstruent regions is first performed. DP search is then carried out in vowel-like and nasal regions where continuous formant trajectories are expected. This DP search minimizes the cost of mapping a set of candidate frequencies to formants, so that smooth formant trajectories that correspond to spectral prominences are selected. To improve its robustness against missing or outlier candidate frequencies, a further processing of estimated formants is performed. After a median smoothing, the entire formant line in its region is divided into concatenative continuous fragments and the longest smooth one is used as an anchor formant. A local beam search from the anchor formant is carried out to detect neighboring formant points.

Comparison of the automatically derived formant tracks with hand-corrected formant lines obtained from a commercial recognizer indicates the low error rate of this estimation algorithm for the first three formants in vowel-like regions. The error rate is higher in nasal regions due to zeros. Evaluation of this approach on a larger database is planned.

Several directions for algorithm improvement have been identified, including: adaptive thresholding in local beam search and segmentation post-processing, more complicated handling of formant line fragments, spectrum-adapted cost constants and tuning of mapping cost function in Viterbi search.

6. ACKNOWLEDGEMENTS

This work was supported by NSF grant SBR-9729688

7. REFERENCES

[12] Xwaves 5.3.1, Entropic Research Laboratory Inc.