

Tracking Vocal Tract Resonances Using an Analytical Nonlinear Predictor and a Target-Guided Temporal Constraint

Li Deng, Issam Bazzi, and Alex Acero

Microsoft Research, One Microsoft Way, Redmond WA 98052, USA

Abstract

A technique for high-accuracy tracking of formants or vocal tract resonances is presented in this paper using a novel nonlinear predictor and using a target-directed temporal constraint. The nonlinear predictor is constructed from a parameter-free, discrete mapping function from the formant (frequencies and bandwidths) space to the LPC-cepstral space, with trainable residuals. We examine in this study the key role of vocal tract resonance targets in the tracking accuracy. Experimental results show that due to the use of the targets, the tracked formants in the consonantal regions (including closures and short pauses) of the speech utterance exhibit the same dynamic properties as for the vocalic regions, and reflect the underlying vocal tract resonances. The results also demonstrate the effectiveness of training the prediction-residual parameters and of incorporating the target-based constraint in obtaining high-accuracy formant estimates, especially for non-sonorant portions of speech.

1. Introduction

One fundamental difficulty for machine recognition of casual, conversational speech is the reduction of phonetic information, where the underlying dynamic properties of articulation undergo systematic modifications. The seemingly non-systematic changes in the surface acoustics induced by the articulatory modifications often make speech sound classes highly confusable when only the acoustic information is used by speech recognizers (e.g., those constructed from conventional HMMs). To reduce such confusability, we have developed an approach that takes into account aspects of the underlying dynamic properties of speech production and their causal relationship to the observed speech acoustics. In this paper, we describe such an approach and one of its specific implementations. In particular, we use the vocal tract resonance (VTR) properties as the representation of the underlying dynamics of speech production. The VTRs include formant frequencies and bandwidths for all regions of speech utterances. Importantly, VTRs may not coincide with the spectral prominences during non-sonorant portions of speech¹ but reflect the underlying vocal tract resonance properties even when the mouth is close (partial or full). This is desirable since the formant transitions in sonorant speech into and out of vocal tract closures can be predicted based on the directions of their resonance targets or loci, but not necessarily by the related spectral prominences [6].

In this paper, we will demonstrate the feasibility of the proposed approach to modeling internal speech dynamics by exploiting a target-guided dynamic system model to track hidden

¹The spectral prominences may result from complex pole and zero interactions in non-sonorant sounds, and thus the frequencies corresponding to such spectral prominences may differ from the pole or resonance frequencies.

VTRs using measurable speech acoustics. The organization of the paper is as follows. In Section 2, we present a forward, approximate mapping function from the VTR variables to the speech acoustics represented by LPC-Cepstra. Inversion of this function gives a crude VTR estimate. In Section 3, we introduce trainable residuals and noise to account for errors due to the functional approximation. We further introduce the target-guided dynamic constraint as the prior knowledge for the VTR's temporal behavior. A combination of the constraint and the mapping function with trainable residuals constitutes a dynamic system model of speech. We then present the algorithms for residual's parameter training and for VTR tracking in Sections 4 and 5, respectively. In Section 6, experimental results are presented to demonstrate the effectiveness of the training and of incorporating the target-based constraint in obtaining high-accuracy VTR estimates especially for non-sonorant portions of speech.

2. Nonlinear Mapping Function from VTR to LPC-Cepstrum

As a basis for formant estimation, we in this section present an approximate nonlinear mapping function, $C(x)$, from the VTR variables (x) to observed speech acoustics (C). Inversion of this function provides a straightforward but crude VTR estimate.

Depending on the type of the acoustic measurements as the output, closed-form computation for $C(x)$ may be impossible, or its in-line computation may be too expensive. To overcome these difficulties, we quantize each dimension of x over a range of frequencies or bandwidths, and then compute $C(x)$ for every quantized value of x . In [2], we detailed a procedure for constructing $C(x)$ when the output acoustic measurements are the MFCC features. In this paper, we present the case when the output is LPC-Cepstra, which has a significant advantage of computation efficiency due to the decomposition property which we describe below.

Consider an all-pole model, with each of its poles represented as a frequency-bandwidth pair (f_k, b_k) . Then the corresponding complex root is given by [1]:

$$z_k = e^{-\pi \frac{b_k}{f_s} + j2\pi \frac{f_k}{f_s}}, \quad (1)$$

where f_s is the sampling frequency. The transfer function with K poles and a gain of G is:

$$H(z) = G \prod_{k=1}^K \frac{1}{(1 - z_k z^{-1})(1 - z_k^* z^{-1})}. \quad (2)$$

Taking logarithm on both sides of Eq. 2 and then using

$\log(1-x) = -\sum_{n=1}^{\infty} x^n/n$, we obtain:

$$\begin{aligned} \log H(z) &= \log G - \sum_{k=1}^K \log(1 - z_k z^{-1}) - \sum_{k=1}^K \log(1 - z_k^* z^{-1}) \\ &= \log G + \sum_{k=1}^K \sum_{n=1}^{\infty} \frac{z_k^n z^{-n}}{n} + \sum_{k=1}^K \sum_{n=1}^{\infty} \frac{z_k^{*n} z^{-n}}{n} \\ &= \log G + \sum_{n=1}^{\infty} \left[\sum_{k=1}^K \frac{z_k^n + z_k^{*n}}{n} \right] z^{-n} \\ &= \log G + \sum_{n=1}^{\infty} \left[\sum_{k=1}^K \frac{2}{n} e^{-\pi n \frac{b_k}{f_s}} \cos(2\pi n \frac{f_k}{f_s}) \right] z^{-n}. \end{aligned}$$

The inverse z -transform of the above gives the n -th order LPC-Cepstrum (using the one-sided z -transform definition):

$$C_n = \sum_{k=1}^K \frac{2}{n} e^{-\pi n \frac{b_k}{f_s}} \cos(2\pi n \frac{f_k}{f_s}), \quad n > 0 \quad (3)$$

and $C_0 = \log G$.

Eq. 3 gives the decomposition property of the LPC-Cepstra: each of the LPC-Cepstral coefficients is a sum of the contributions from separate VTRs. This contrasts the MFCC feature of [2], which is a function of all VTRs but is not in a simple additive form such as Eq. 3. The key advantage of the decomposition property is that it makes the optimization procedure highly efficient for inverting the nonlinear function from the acoustic feature to the VTR.

3. Trainable Residuals and Target-Guided Temporal Constraints

In practical implementation, computation of LPC-Cepstra from VTRs according to Eq. 3 can include the sum of only a finite number of poles; in our experiments, we chose $K = 4$, or using an eight-dimensional vector $x = (f_1, f_2, f_3, f_4, b_1, b_2, b_3, b_4)$ as the input to the nonlinear function $C(x)$ with 30 orders of LPC-Cepstra as the output.² The remaining (higher-order) poles and possible zeros (and their interactions with poles) are known to affect speech acoustics and create approximation errors. Further, using LPC as the basis for computing the acoustic feature may cause an observation error. One way to improve the mapping function is to introduce the learnable prediction residuals in order to compensate for all sources of errors.

Denoting the residual by $v(s)$, which is assumed to be a Gaussian random variable with mean vector h_s and (diagonal) precision matrix D_s , which may be dependent on discrete state s (e.g., phone): $v(s) \sim N(v; h_s, D_s)$.

After accounting for the approximation error by the IID residual $v_t(s)$ for each time frame t , the exact relationship between the VTR vector x_t and the LPC-Cepstral vector o_t becomes:

$$o_t = C(x_t) + v_t(s). \quad (4)$$

This forms the *observation equation* of a dynamic system model with the state-space formulation.

We can further improve the prediction from the VTR *sequence* to the LPC-Cepstral *sequence* by exploiting the prior

²In our earlier work in [2], the six-dimensional input vector of $(f_1, f_2, f_3, b_1, b_2, b_3)$ and the 12-dimensional output vector (MFCCs) were used. Use of the fourth formant improves the overall formant tracking accuracy, and use of the LPC-Cepstra improves computational efficiency.

knowledge about the VTR's temporal behavior. This gives the target-guided dynamic constraint expressed by the following first-order *state equation* of the dynamic system model:

$$x_t = r_s x_{t-1} + (1 - r_s) T_s + w_t(s), \quad (5)$$

where the (continuous) state noise at frame t is assumed to be IID, zero-mean Gaussian: $w_t(s) \sim N(w_t; 0, B_s)$, with a discrete state(s)-dependent (diagonal) precision matrix B_s . This state equation has the desirable property that x_t would asymptotically approach the (phone-dependent) target T_s as time $t \rightarrow \infty$ (with the rate controlled by the parameter r_s).

Due to the discrete nature³ in the construction of the nonlinear predictor $C(x_t)$, we quantize the continuous state of x_t in the entire model Eqs. 4 and 5. We denote the value of x_t at the i -th level of quantization as $x_t[i]$ or simply $i_t = i$.

4. Algorithm for Parameter Estimation

Following the EM algorithm, we have derived re-estimation formula for all the parameters in the state-space model consisting of Eqs. 4 and 5. In particular, the parameter of the mean in the prediction residual of Eq. 4 is re-estimated in each EM iteration by:

$$\hat{h}_s = \frac{\sum_{t=1}^N \sum_{i=1}^I \gamma_t(s, i) \{o_t - C(x_t[i])\}}{\sum_{t=1}^N \gamma_t(s)}, \quad (6)$$

where N is the total number of frames in the observation data, I is the total number of quantization levels for the VTRs, and the posteriors

$$\gamma_t(s, i) \equiv p(s_t = s, x_t[i] | o_1^N) \quad \text{and} \quad \gamma_t(s) \equiv p(s_t = s | o_1^N)$$

are computed efficiently using a generalized forward-backward algorithm.

Re-estimation for each diagonal element d_s^{-1} of the residual variance D_s^{-1} is

$$\hat{d}_s^{-1} = \frac{\sum_{t=1}^N \sum_{i=1}^I \gamma_t(s, i) [o_t - C(x_t[i]) - \hat{h}_s]^2}{\sum_{t=1}^N \gamma_t(s)}. \quad (7)$$

5. Algorithm for Formant Tracking

After the state-space model's parameters are trained, the model can be used for simultaneous speech recognition (e.g., decoding of the phone sequence s) and VTR tracking. In this study, we simplify our approach and focus only on the problem of VTR tracking (formant frequencies and bandwidths).

The Viterbi decoding algorithm described here is aimed to find the best single quantized VTR sequence $i_1^N = (i_1, i_2, \dots, i_N)$ for a given observation sequence o_1^N . Let's define the optimal partial score of

$$\begin{aligned} \delta_t(i) &= \max_{i_1^{t-1}} p(o_1^t, i_1^{t-1}, x_t[i]) = \max_{i_1^{t-1}} p(o_1^t, i_1^{t-1}, i_t = i) \\ &= \max_{i_1^{t-1}} \prod_{\tau=1}^{t-1} p(o_\tau | i_\tau) p(o_t | i_t = i) \\ &\quad \times p(i_1) \prod_{\tau=2}^{t-1} p(i_\tau | i_{\tau-1}) p(i_t = i | i_{t-1}). \quad (8) \end{aligned}$$

Each $\delta_t(i)$ defined in Eq. 8 is associated with a node in the trellis diagram. Each increment of time corresponds to reaching a new

³In this work, we use a tabulated form as in [2] to represent the nonlinear function of Eq. 3.

stage in dynamic programming. At the final stage $t = N$, we have the objective function of $\delta_N(i)$, which is accomplished via all the previous stages of computation for $t \leq N - 1$. Based on the optimality principle, the optimal partial likelihood at the processing stage of $t + 1$ can be computed using the following Viterbi recursion:

$$\delta_{t+1}(i) = \max_{i'} \delta_t(i') p(i_{t+1} = i | i_t = i') p(o_{t+1} | i_{t+1} = i) \quad (9)$$

In the above, the “transition probability” $p(i_{t+1} = i | i_t = i')$ is computed based on the state equation Eq. 5 (dependency on the discrete phone state s omitted):

$$p(i_{t+1} = i | i_t = i') = N(x_{t+1}[i]; rx_t[i'] + (1-r)T, B), \quad (10)$$

and the “observation probability” $p(o_{t+1} | i_{t+1} = i)$ is computed from the observation equation Eq. 4:

$$p(o_{t+1} | i_{t+1} = i) = N(o_{t+1}; C(x_{t+1}[i]) + h, D). \quad (11)$$

Back tracing of the optimal VTR quantization index i' in Eq. 9 gives the estimated VTR sequence.

6. Experiments

As an initial step towards implementing a speech recognizer capable of recognizing conversational speech with a casual style, we evaluated the dynamic system model consisting of Eqs. 4 and 5 in the task of formant or VTR tracking. In this section, we first describe some simplification of the algorithms presented earlier for the formant tracking task. We then report the results of formant tracking, and examine specifically the roles of EM training and of the use of targets in Eq. 5 in formant tracking accuracy.

6.1. Implementing parameter estimation

For formant tracking applications, the information about phone identity as the discrete state s in Eq. 4 and 5 is neither available nor needed. Hence, in simplifying the general algorithm for parameter estimation described in Section 4, we tie all states s in Eq. 6 by summing both the numerator and denominator over s to obtain :

$$\hat{h} = \frac{\sum_{t=1}^N \sum_{i=1}^I \gamma_t(i) \{o_t - C(x_t[i])\}}{N}, \quad (12)$$

where the posterior is simplified to

$$\begin{aligned} \gamma_t(i) &= \sum_{s=1}^S \gamma_t(s, i) = p(x_t[i] | o_t^N) \\ &\approx p(x_t[i] | o_t) = \frac{N(o_t; C(x_t[i]) + \hat{h}_0, \hat{D}_0)}{\sum_{i=1}^I N(o_t; C(x_t[i]) + \hat{h}_0, \hat{D}_0)}, \end{aligned}$$

and no forward-backward recursion is required under the above approximation.

Similarly, re-estimation for each diagonal element of the tied observation noise matrix is simplified from Eq. 7 to

$$\hat{d}^{-1} = \frac{\sum_{t=1}^N \sum_{i=1}^I \gamma_t(i) [o_t - C(x_t[i]) - \hat{h}]^2}{N}. \quad (13)$$

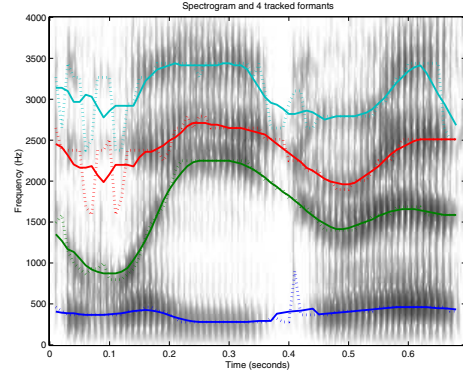


Figure 1: Formant tracking with three EM iterations (no targets).

6.2. Implementing formant tracker

One difficulty for the rigorous Viterbi algorithm described in Section 5 is a large computational cost, due mainly to the very large number of quantization levels (over an eight-dimensional VTR space) that need to be explored theoretically for each frame. In practice, we can use pruning in beam search to significantly cut down the computational cost. In this work, we use an extreme form of beam search — keeping the beam size to be one for all frames. Then, the Viterbi algorithm described in Section 5 is effectively simplified to

$$\hat{i}_t = \arg \max_{i_t} p(o_t | i_t) p(i_t | \hat{i}_{t-1}), \quad (14)$$

for each frame t , and no back tracing is needed.⁴ Further, in our experiments, the probability of $p(o_t | x_t[i])$ in Eq. 14 is efficiently computed using the decomposition property of $C(x)$ for the LPC-Cepstrum output based on Eq. 3. As a result, the formant tracker as implemented in Matlab runs close to real time on a P-III machine.

6.3. Results on the role of EM training

The experimental results presented in this section are obtained using the simplified re-estimation formula Eqs. 12 and 13. Figs. 1 and 2 show the formant tracking (f_1, f_2, f_3, f_4) results, superimposed on spectrograms, with EM iterations of three and five,⁵ respectively. The dashed lines are the results from the tracking algorithm Eq. 14, and the solid lines are their seven-point moving averages. The results are from a Switchboard utterance of “the way you dress” by a male speaker. Due to the elimination of phone label s in Eqs. 12 and 13, the training does not require any data labeling and is thus fully unsupervised. In obtaining Figs. 1 and 2, the role of the VTR-target parameter in Eq. 10 is eliminated by setting $r = 1$. Moving from Figs. 1 to 2, we clearly see significant improvement of the formant tracking accuracy, especially in the regions where the speech energy is relatively low (near time 0.1 and 0.4 sec).

6.4. Results on the role of incorporating targets

Figs. 3 and 4 are obtained in the same way as Figs. 1 and 2, except that we set $r = 0.7$ in Eq. 10 so that the VTR target parameter T plays a role in formant tracking (using the algorithm

⁴This is the same simplification technique used for incorporating the delta-parameter prior for speech feature enhancement developed in [3].

⁵EM convergence is found at about the fifth iteration for this utterance.

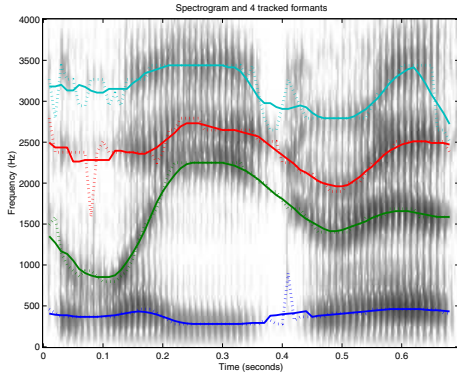


Figure 2: Formant tracking with five EM iterations (no targets).

Eq. 14). Since the VTR targets are dependent on the phone identity s , we use a phone segmentation of the utterance to assign the relevant target values of T_s to each frame in the utterance. The target values of f_1, f_2, f_3 are from the work of [4, 5], and the target value for f_4 is fixed at 3600 Hz for all frames.

Significant improvement of formant tracking accuracy can be seen in Figs. 3 and 4 over that which does not incorporate targets. All the major errors and formant discontinuity (especially around time 0.1 and 0.4 sec) in Figs. 1 and 2 have been eliminated. The slight discontinuity for f_3 and f_4 around time 0.45 sec in Fig. 3 (three EM iterations) is eventually eliminated after two more EM iterations as shown in Fig. 4 (five EM iterations). Further, most of the detailed formant transitions from one sound to another, including those involving vocal tract closures such as /d/ around time 0.4 sec, have been accurately tracked in Figs. 3 and 4. Finally, the incorrectly tracked f_4 towards the end of the utterance in Figs. 1 and 2 has been corrected due to the use of the f_4 target for the sound /s/.

7. Discussions and Conclusions

Previous techniques for formant tracking typically used LPC or other kinds of spectral analysis to compute formant candidates, which were then selected with generic continuity constraints (e.g., [1]). Our recent work [2] improved this by exploring a complete formant space based on a nonlinear predictor. The work described in this paper extends the work of [2] further in three key aspects. First, a target-based temporal constraint is used to replace the generic continuity constraint, enabling accurate tracking of VTRs in non-sonorant speech regions. In addi-

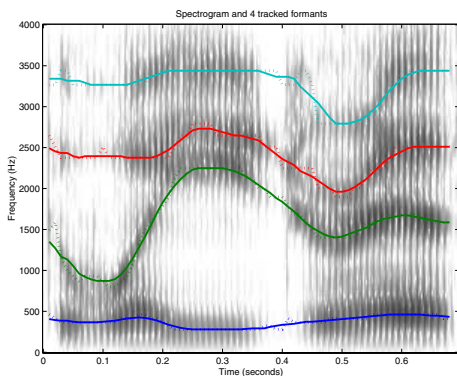


Figure 3: Formant tracking with three EM iterations (with targets).

tion, formant transitions near the CV or VC boundaries can also be tracked accurately using the targets of adjacent sounds. Second, the output of the nonlinear predictor is changed from the previous MFCC to the current LPC-Cepstrum, which provides significant computational advantage due to its decomposition property. Third, an approximate Viterbi algorithm is developed, motivated by the work of [3] in a separate application, which combines the static nonlinear predictor with the target-based dynamic constraint to efficiently carry out formant tracking.

The results of this work provide preliminary evidence that the target-directed state-space formalism consisting of Eqs. 4 and 5 is effective in modeling the hidden dynamics of speech (as represented by VTRs) and its causal relationship to measurable speech acoustics (as represented by LPC-Cepstra). Our future work is to expand our current implementation of the model so that it includes discrete phonological states. This will enable construction of a speech recognizer capable of decoding conversational speech, which is characterized by a high degree of phonetic reduction. This reduction property is already embedded in the target-directed temporal constraint as a critical part of the state-space model presented in this paper.

8. References

- [1] A. Acero, "Formant analysis and synthesis using hidden Markov models," in *Proc. of the Eurospeech Conference*, Budapest, 1999.
- [2] I. Bazzi, A. Acero, and L. Deng. "An expectation-maximization approach for formant tracking using a parameter-free non-linear predictor," *Proc. ICASSP*, Hong Kong, April 2003.
- [3] L. Deng, J. Droppo, and A. Acero. "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior," *Proc. ICASSP*, Orlando, Florida, 2002.
- [4] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics," *J. Acoust. Soc. Am.*, Vol. 108, 2000.
- [5] F. Seide, J.L. Zhou, and L. Deng. "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM — MAP decoding and evaluation," *Proc. ICASSP*, Hong Kong, April 2003.
- [6] K. Stevens, *Acoustic Phonetics*, The MIT Press, Cambridge, MA, 1998.

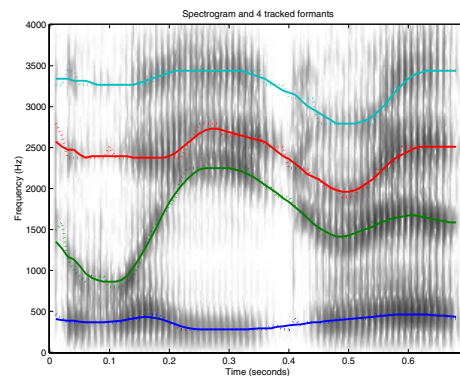


Figure 4: Formant tracking with five EM iterations (with targets).