Conditionally Linear Gaussian Models for Estimating Vocal Tract Resonances

Daniel Rudoy, Daniel N. Spendley, and Patrick J. Wolfe

School of Engineering and Applied Sciences
Department of Statistics, Harvard University
Harvard-MIT Division of Health Sciences & Technology
Oxford Street, Cambridge, MA 02138 USA
{rudoy,spendley,patrick}@seas.harvard.edu

Abstract

Vocal tract resonances play a central role in the perception and analysis of speech. Here we consider the canonical task of estimating such resonances from an observed acoustic waveform, and formulate it as a statistical model-based tracking problem. In this vein, Deng and colleagues recently showed that a robust linearization of the formant-to-cepstrum map enables the effective use of a Kalman filtering framework. We extend this model both to account for the uncertainty of speech presence by way of a censored likelihood formulation, as well as to explicitly model formant cross-correlation via a vector autoregression, and in doing so retain a conditionally linear and Gaussian framework amenable to efficient estimation schemes. We provide evaluations using a recently introduced public database of formant trajectories, for which results indicate improvements from twenty to over 30% per formant in terms of root mean square error, relative to a contemporary benchmark formant analysis tool.
Index Terms: formant tracking, speech analysis, Kalman filtering, vocal tract resonances, system identification

1. Introduction

Vocal tract resonances—often termed formants—play a central role in the perception and analysis of speech sounds. This paper considers the canonical task of estimating such resonances from an observed acoustic waveform, formulated as a statistical model-based tracking problem. In particular, one strand of literature [1, 2, 3] concentrates on continuous-valued formant processes rather than the discretized models [4, 5, 6, 7] often implemented in practical systems such as WaveSurfer [8], which employs an approach based on dynamic programming that imposes continuity constraints [5]. Such an approach has the advantage of readily lending itself to an interpretation in terms of a *generative probabilistic model* for vocal tract resonance (VTR) trajectories. Moreover, the assumption (when justified) of a conditionally linear and Gaussian model can greatly simplify analysis and implementation.

In this vein, Deng and colleagues [9] recently showed that a robust linearization of the formant-to-cepstrum map enables the use of a Kalman filtering framework. As the canonical example of a linear, Gaussian generative model, this corresponds to the assumption of linear dynamics and Gaussian processes. Here, however, we interpret the Kalman filter as an algorithm for the propagation of first- and second-order statistics of the underlying processes. As noted in [9], this interpretation is perhaps most appropriate for the problem as formulated here, which involves an unobserved state consisting of formant locations and bandwidths, with "noisy" data corresponding to the linear predictive cepstral coefficients of the observed acoustic waveform.

The parametric formant-tracking approaches described above aim to estimate the unobserved VTR frequencies and bandwidths directly from observed speech waveform data. However, most of these methods incorporate additional algorithms (such as root finding or peak-picking [10]) as subroutines in order to compute the observations as a function of the hidden variables (that is, to formulate a pseudo-likelihood equation), and hence do not readily yield analytic formulations. On the other hand, an explicit, nonlinear relationship amongst the formant frequencies and bandwidths and the linear predictive coding (LPC) cepstrum of speech is derived in [9], along with an effective linearization procedure that admits a conditional linear and Gaussian form.

Here we concentrate on this same class of models for estimating vocal tract resonances. In Section 2 we state the precise model to be considered, and then extend it both to account for the absence of waveform energy during pauses in utterances, by way of a censored likelihood formulation, as well as to explicitly model correlation structure across formants, using a vector autoregressive model. We show how to realize these extensions while retaining a conditionally linear and Gaussian framework, and provide evaluations using a recently released public database [11] of formant trajectories in Section 3. Finally, we conclude with a discussion of these results in Section 4.

2. Models and Methods

We begin by describing a statistical model for the evolution of vocal tract resonances in speech, elements of which are by now standard in the literature. In a state-space formulation, each VTR is modeled by a second-order digital resonator and is parameterized by a frequency f_k and bandwidth b_k (both in units of Hertz). Assuming that the spectral envelope of speech is well characterized by K resonators, then at frame time index t, the spectral envelope can be parameterized by a vector $\boldsymbol{x}_t \in \mathbb{R}_+^{2K}$, where $\boldsymbol{x}_t = (f_1, \ldots, f_K, b_1, \ldots, b_K)^T$. The VTR evolution is modeled according to a discrete-time Gauss-Markov process:

$$\boldsymbol{x}_{t+1} = \boldsymbol{F}\boldsymbol{x}_t + \boldsymbol{w}_t,$$

where $\boldsymbol{F} \in \mathbb{R}^{2K \times 2K}$ is the state transition matrix and $\boldsymbol{w}_t \in \mathbb{R}^{2K}$ is a white sequence satisfying $\mathbb{E}(\boldsymbol{w}_i \boldsymbol{w}_j^T) = \boldsymbol{Q}\delta_{ij}$, with $\boldsymbol{Q}$ denoting the process noise covariance and δ_{ij} the Kronecker delta. At each time frame t, the first N coefficients of the LPC cepstrum $\boldsymbol{y}_t = (y_t[1], \ldots, y_t[N])^T$ are observed in additive white Gaussian noise.[1] The relationship between the n^{th} (scalar)

[1] Equivalently, and particularly in the case at hand, this "noise" term may be interpreted as an ℓ^2 goodness-of-fit criterion.

cepstral coefficient $y_t[n]$ and the state vector $\boldsymbol{x}_t$ is given by the following well-known expression [12, 9]:

$$y_t[n] = \frac{1}{n}\sum_{k=1}^{K} \exp\left(\frac{\pi n}{f_s} x_t[k+K]\right) \cos\left(\frac{2\pi n}{f_s} x_t[k]\right)$$

where f_s is the sampling frequency of the acoustic waveform. Let $h : \mathbb{R}_+^{2K} \rightarrow \mathbb{R}^N$ be a vector-valued nonlinear mapping from the vector $\boldsymbol{x}_t$, representing resonator frequencies and bandwidths, to the vector $\boldsymbol{y}_t$ of LPC cepstral coefficients, defined coordinate-wise by (1). The model specification is completed by the likelihood equation

$$\boldsymbol{y}_t = h(\boldsymbol{x}_t) + \boldsymbol{v}_t, \tag{2}$$

where $\boldsymbol{v}_t \in \mathbb{R}^N$ is a white sequence satisfying $\mathbb{E}(\boldsymbol{v}_i \boldsymbol{v}_j^T) = \boldsymbol{R}\delta_{ij}$, with $\boldsymbol{R}$ denoting the process noise covariance. Finally, we assume that the process and observation noise sequences are uncorrelated and satisfy $\mathbb{E}(\boldsymbol{v}_i \boldsymbol{w}_j^T) = \boldsymbol{0}$ for all i and j.

Given the state-space formulation of the model, we are interested in computing the distribution $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t})$ of the VTR parameters at frame t, conditioned upon the waveform data observed up to frame t. We then take the mean of this distribution as a point estimate of the VTR parameters. Additionally, note that this statistical formulation allows for uncertainty quantification via the variance of $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t})$—in contrast to the dynamic programming approaches previously cited, where no such method of error quantification is readily available.

Attempts have been made to fit the above model to data, both by Monte Carlo simulation methods [2] as well as analytical approximations [9]. Though the former approach is mathematically straightforward, it has been reported to perform poorly relative to the extended Kalman filter (EKF) of the latter, in which a careful coordinate-wise linearization of the formant-to-cepstrum mapping h is carried out by approximating each term appearing in the summation of (1). Before describing our own fitting procedure in detail, we discuss two important extensions to this model that will be shown to yield significant improvement in VTR estimation performance.

2.1. Extension I: Censoring the Likelihood

The model described above does not explicitly take into account the uncertainty of speech presence. Hence, approaches based on this model as it stands (along with previously cited others) are liable to suffer significant performance degradations in practice—-not only during pauses in utterances, but also whenever formant peaks cease to be observable in the waveform (i.e., are censored in the likelihood of (2)). This effect is strongly pronounced in the trajectories of the VTR database [11], in which much of the hand-labeling effort was devoted to correcting formant tracking errors made during non-speech regions. To address this problem, we censor the likelihood of (2) by augmenting the state vector $\boldsymbol{x}_t$ with a binary indicator variable for each formant. We model these indicators as statistically independent from frame to frame, and assume that in each frame they are estimated by a voice activity detector.[2]

In this augmented state-space model, the optimal state estimate may be obtained in closed form by the classical Schmidt-Kalman filter [13]—thereby avoiding the ambiguity inherent in estimating VTR parameters in non-speech regions of the time-frequency plane (a consideration that is absent from many existing algorithms). Propagating the VTR trajectories in this manner (which is akin to "coasting" the state estimate) provides a coherent and extensible means by which to incorporate the uncertainty of speech presence into VTR estimation. Moreover, the estimator variance grows during silent regions (as it should), reflecting increased uncertainty regarding trajectory locations.

2.2. Extension II: Modeling Formant Cross-Correlation

A separate observation is that formants do not evolve independently of one another. Specifically, the temporal trajectories of formants are correlated in frequency; for example, in synthesis of front vowels, it is common practice to employ a linear regression of formant F3 onto F1 and F2 (see, e.g., [14]). As a preliminary step in our analysis, we empirically estimated the correlation structure amongst all three hand-corrected formant trajectories in the VTR database (F1, F2, and F3), and have observed significant cross-correlation of formant trajectories. Moreover, we observed this empirical cross-correlation function to decay slowly, implying that a set of formant values at frame t may be helpful in predicting values of all formants at frame $t+1$.

This observation suggests another extension to the state space model outlined above: off-diagonal terms of the state transition matrix $\boldsymbol{F}$ may be used to incorporate the structure of the formant cross-correlation function at a lag of one frame; this allows for potential improvements in VTR parameter estimation.[3] Conditioned on utterance data, the elements of $\boldsymbol{F}$ can be straightforwardly estimated using a linear-least-squares estimate [15]. Note that this is equivalent to fitting an order-one vector autoregressive process—a so-called VAR(1) model—to VTR trajectories. By fixing these elements prior to algorithm execution (i.e., by conditioning on some estimate $\hat{\boldsymbol{F}}$ of $\boldsymbol{F}$), the modeling framework remains linear and Gaussian.

3. Experimental Results

In this section we describe a sequence of experiments designed to evaluate the extensions outlined in Sections 2.1 and 2.2. These experiments were carried out using the recently introduced VTR database of [11], which contains a representative subset of the TIMIT speech corpus [16] consisting of diverse, phonetically balanced utterances collated across a range of gender, individual speakers, dialects, and phonetic contexts. The VTR database contains state information consisting of four formants and their bandwidths for each frame of analyzed speech.

3.1. Analysis Settings and Empirical Parameter Estimation

Our experimental setup can be described in three parts as follows. First, we extracted all 516 VTR utterances[4] and corresponding TIMIT waveform data, after which we performed careful pre-processing to evaluate the performance of the WaveSurfer formant tracking algorithm [8] against the VTR "ground truth." All analysis parameters were matched to those used to generate the VTR database: 20 ms Hamming windows with an overlap of 50% were employed, left-aligned with the first sample of each TIMIT utterance, and an LPC model order of 12 was used, in conjuction with a pre-emphasis coefficient

[2] This approach has the advantage of being easily extensible to incorporate frame-to-frame dependence, in order to model more difficult cases of formant disappearance (e.g., pole-zero cancellations).

[3] Cross-correlation at lags of greater than one frame may be incorporated by appropriately augmenting the state vector.

[4] Reference [11] lists 538 entries, but 22 of these are text-file annotations of corrections, leaving a total of 516 TIMIT utterances.

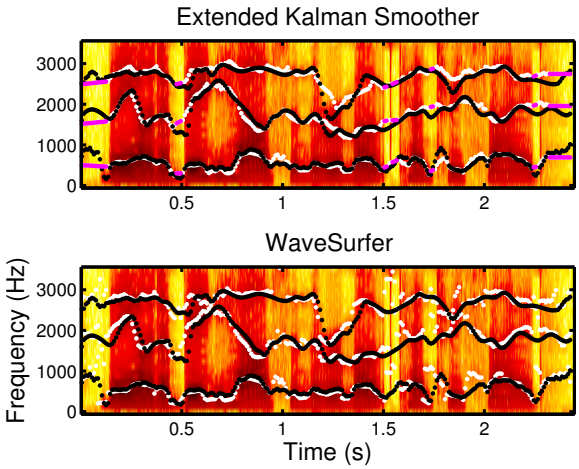

Figure 1: *Comparison of tracking performance for TIMIT utterance, "They own a big house in the remote countryside." The VTR trajectories are shown in black, with the extended Kalman smoother (top, average RMSE 107.5 Hz) and WaveSurfer output (bottom, average RMSE 214.3 Hz), both shown in white.*

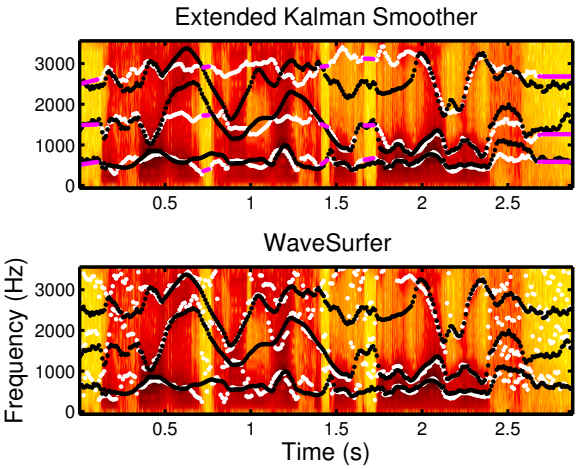

Figure 2: *Comparison of tracking performance for TIMIT utterance, "We saw eight tiny icicles below our roof." The VTR trajectories are shown in black, with the extended Kalman smoother (top, average RMSE 284.6 Hz) and WaveSurfer output (bottom, average RMSE 398.4 Hz), both shown in white.*

of 0.7. Second, we processed the corresponding TIMIT waveform data by first resampling from 16 kHz to 7 kHz (in order to track the first three formants, as these constitute the only hand-corrected entries in the VTR database), and then computing the first 15 LPC cepstral coefficients via the standard recursion using the analysis parameters described above. For each utterance, an energy-based detector was used to identify non-speech frames as those whose energy fell within the lower 15^{th} percentile with respect to the entire utterance.

Finally, we describe our method of fitting the model parameters detailed in Section 2. As our focus here is on model elicitation rather than model fitting per se, we elected to use the output of WaveSurfer as a means of empirically estimating all model parameters.[5] For each utterance, the corresponding state transition matrix $\boldsymbol{F}$ was estimated from WaveSurfer formant tracks using the methods of Section 2.2. The process noise covariance $\boldsymbol{Q}$ was likewise estimated empirically from WaveSurfer output via maximum likelihood, conditioned upon speech presence. We linearized the measurement equation of (2) following the method of [9], to obtain an observation matrix $\boldsymbol{H}$. Finally, we fixed the observation noise covariance matrix $\boldsymbol{R}$ for all utterances to be diagonal, with terms given by $R_{nn} = 1/n$ for $n \in \{1, 2, \ldots, 15\}$. Empirically, we observed this to be in reasonable agreement with the variance of typical LPC cepstrum residual vectors, derived separately from WaveSurfer-generated VTR parameters and TIMIT acoustic waveform data. Owing to the fact that no hand-corrected VTR bandwidths exist in the database, we set the bandwidth state values to be equal to the average of the values reported by WaveSurfer for the duration of each utterance. For computing formant trajectories, we used a well-known generalization of the extended Kalman filter termed the Schmidt-Kalman filter, to properly coast the tracks during intervals of speech absence. We also implemented a Rauch-Tung-Striebel smoother.

3.2. Formant Tracking Results

For each TIMIT utterance in the VTR database, we obtained an associated WaveSurfer point estimate of formant trajectories as well as ones stemming from the filtering and smoothing algorithms described above. We begin by examining in detail two specific cases; subsequently, we summarize results for the Schmidt-Kalman filter via root mean square error (RMSE) in Hertz per formant, averaged over all VTR database utterances.

Figures 1 and 2 show two utterances, selected after paging through approximately 5% of the VTR database entries according to a random sequence. For each utterance we report RMSE, averaged over the first three formants and conditioned on speech presence; regions of speech absence are denoted by darkened segments of the white estimated trajectories, according to the Rauch-Tung-Striebel smoother. Figure 1 indicates a marked improvement in tracking capability relative to the performance of Wavesurfer, both graphically and in terms of absolute RMSE. In contrast, Figure 2 shows an example that, while superior to WaveSurfer's output in many regions of the utterance, does not match the entire VTR trajectories as well as in the previous example, particularly with respect to large motions of F2 and F3.

Table 1 overleaf summarizes the average RMSE reduction per formant relative to WaveSurfer when both modeling extensions described in Section 2 are employed to analyze the entire VTR database. Table 2 details the specific effect of modeling formant cross-correlation as described in Section 2.2, conditioned upon speech presence. Two clear patterns are visible in this table, which specifies all possible ways of capturing cross-correlation through a VAR(1) model. First, it is clear that modeling cross-correlation amongst all formants yields significant improvements relative to assuming mutually independent evolutions (cf. $\emptyset$ and $\{1, 2, 3\}$ in Table 2). The effect is easily seen for all formants when comparing the first and lasts columns of Table 2, and is most pronounced in the tracking of F3.

Second, whenever F3 is treated independently of F1 or F2 (i.e., $\emptyset$ and $\{1, 2\}$), we observe a performance degradation. On the other hand, estimation results for F3 improve when the cross-correlation between it and the other formants is incorporated into the VAR(1) model (i.e., $\{2, 3\}$, $\{1, 3\}$, and $\{1, 2, 3\}$). The fact that F3 depends strongly on F1 and F2 for certain phonemes (front vowels in particular) provides a possible explanation for the performance differential in these cases.

[5] In the design of a practical tracker, a number of methods may readily be brought to bear on the underlying system identification problem.

	Root Mean Square Error Reduction			
Formant	Entire Utterance		Speech Presence	
Number	(Hz)	(%)	(Hz)	(%)
1	90.93	39.32%	52.31	22.62%
2	103.0	30.55%	59.18	17.56%
3	147.3	32.74%	99.79	22.18%

Table 1: *Observed reduction in average root mean square error relative to WaveSurfer [8], taking the VTR database [11] as ground truth. Errors are computed both for the entire utterance (left) and conditioned on speech presence (right).*

	Percentage Root Mean Square Error Reduction				
Formant	Formants included in VAR model				
Number	$\emptyset$	{1,2}	{2,3}	{1,3}	{1,2,3}
1	18.12	22.59	19.29	22.99	22.62
2	7.100	15.48	19.13	13.59	17.56
3	-.0786	2.273	22.68	22.40	22.18

Table 2: *The effect of modeling formant cross-correlation on the percentage reduction of root mean square error relative to WaveSurfer, computed over regions of speech presence. Note the correspondence between the last columns of each table.*

4. Discussion

Here we have considered the problem of estimating vocal tract resonances from an observed acoustic waveform, in the context of a conditionally linear and Gaussian framework. Following the extended Kalman filter proposed by Deng [9], we have presented two important extensions shown to reduce the mean square error in formant estimation relative to WaveSurfer, a contemporary benchmark speech analysis tool. First, by augmenting the state and introducing a censored likelihood as described in Section 2.1, we obtained a model that explicitly accounts for the uncertainty of speech presence. Second, as described in Section 2.2, we showed how to exploit cross-correlation among formant trajectories in order to improve tracker performance.

While our overall results look promising, the values reported represent the first stage of an ongoing investigation into parameter estimation methods for conditional linear Gaussian models in the context of formant tracking. One topic warranting attention is the quantification of the effects of parameter estimation procedures on overall algorithm performance, relative to the improvements due to the censored likelihood and cross-correlation model extensions we have proposed.

Extending this class of models via a more robust voice activity detector, coupled with a greater understanding of the nature of cross-correlation among formants, is another topic of immediate interest. Longer-term goals include employing proper system identification and parameter fitting methods to estimate model parameters online, as well as quantifying robustness of the overall estimation procedure to additive noise and signal degradation. Finally, to enable the reproducibility of these results as well as further experimentation, our software will be made available for download at the corresponding author's home page.

Acknowledgments: *The authors would like to thank Abeer Alwan of the University of California, Los Angeles, Melanie Rudoy of the Massachusetts Institute of Technology, and Geoffrey Morrison of Boston University, for many insightful discussions and much helpful feedback. D. Rudoy is supported by a National Defense Science and Engineering Fellowship.*

5. References

[1] G. Rigoll, "A new algorithm for estimation of formant trajectories directly from the speech signal based on an extended Kalman filter," *Proc. IEEE Int'l. Conf. Acoust. Speech Signal Proces. (ICASSP)*, pp. 1229–1232, 1988.

[2] Y. Zheng and M. Hasegawa-Johnson, "Formant tracking by mixture state particle filter," *Proc. IEEE Int'l. Conf. Acoust. Speech Signal Proces. (ICASSP)*, vol. 1, pp. 565–568, 2004.

[3] L. Deng, L. J. Lee, H. Attias, and A. Acero, "Adaptive Kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Trans. Audio Speech Language Proces.*, vol. 15, no. 1, pp. 13–23, 2007.

[4] G. Kopec, "Formant tracking using hidden Markov models and vector quantization," *IEEE Trans. Acoust. Speech Signal Proces.*, vol. 34, no. 4, pp. 709–729, 1986.

[5] D. Talkin, "Speech formant trajectory estimation using dynamic programming with modulated transition costs," *J. Acoust. Soc. Am.*, vol. 82, no. S1, pp. S55, 1987, abstract Y6.

[6] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 36–48, 1998.

[7] S. Manocha and C. Y. Espy-Wilson, "Knowledge-based formant tracking with confidence measure using dynamic programming," *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1930, 2005, abstract 2pSC5.

[8] K. Sjölander and J. Beskow, "WaveSurfer 1.8.5 for Windows," http://www.speech.kth.se/wavesurfer/wavesurfer-185-win.zip, version 1.8.5 of 01 November 2005.

[9] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," *IEEE Trans. Audio Speech Language Proces.*, vol. 14, no. 2, pp. 425–434, 2006.

[10] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust. Speech Signal Proces.*, vol. 22, no. 2, pp. 134–141, 1974.

[11] 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," *Proc. IEEE Int'l. Conf. Acoust. Speech Signal Proces. (ICASSP)*, vol. 1, pp. 369–372, 2006.

[12] A. V. Oppenheim, R. W. Schafer, and T. G. Stockham, "Nonlinear filtering of multiplied and convolved signals," *Proc. IEEE*, vol. 56, no. 8, pp. 1264–1291, 1968.

[13] S. F. Schmidt, "Application of state space methods to navigation problems," in *Advances in Control Systems*, C. T. Leondes, Ed., New York, 1966, pp. 293–340, Academic Press.

[14] T. M. Nearey, "Static, dynamic, and relational properties in vowel perception," *J. Acoust. Soc. Am.*, vol. 85, pp. 2088–2113, 1989.

[15] J. D. Hamilton, *Time Series Analysis*, Princeton University Press, Princeton, NJ, 1994.

[16] J. S. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, PA, 1993.