Use of Neural Network Mapping and Extended Kalman Filter to Recover Vocal Tract Resonances from the MFCC Parameters of Speech

Roberto Togneri [*],
Li Deng [**]

[*] University of Western Australia, Crawley, WA 6009, Australia
[**] Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA
roberto@ee.uwa.edu.au, deng@microsoft.com

Abstract

In this paper, we present a state-space formulation of a neural-network-based hidden dynamic model of speech whose parameters are trained using an approximate EM algorithm. The training makes use of the results of an off-the-shelf formant tracker (during the vowel segments) to simplify the complex sufficient statistics that would be required in the exact EM algorithm. The trained model, consisting of the state equation for the target-directed vocal tract resonance (VTR) dynamics on all classes of speech sounds (including consonant closure) and the observation equation for mapping from the VTR to acoustic measurement, is then used to recover the unobserved VTR based on Extended Kalman Filter. The results demonstrate accurate estimation of the VTRs, especially those during rapic consonant-vowel or vowel-consonant transitions and during consonant closure when the acoustic measurement alone provides weak or no information to infer the VTR values.

1. Introduction

Resonance frequencies in the human vocal tract, sometimes called formant frequencies for the speech sounds with no substantial supraglottal constriction, are the eigenfrequencies of the air path in the tract from glottis to lips. We refer to them as the vocal tract resonances (VTRs) and regard them as the characteristics of a physical system. Even if no measurable acoustic signals are emitted during some portions of speech (such as closing of mouth during the closure phase of sound /p/), the VTR still exists at some frequency values. In fact, such hidden resonances during consonant closure or contribution have been used in well-known formant synthesizers to provide the "target" positions in the consonant for deriving the transitional formant values during the adjacent consonant-vowel or vowel-consonant junctures [1, 2, 3]. Further, VTR frequencies are constrained to change continuously over time since they are a smooth function of the vocal tract's airway shape, which is in turn determined by the continuously moving articulators. Therefore, VTRs do not disappear, split, or merge during any portion of a speech utterance.

However, many existing formant trackers simply used the acoustic information about the spectral prominences as the basis to determine the formant values as equivalent to the physical resonances. Therefore, when the peaks in the acoustic spectrum become ambiguous, typically during consonant closure, the estimated formants often disappear, split, or merge [4, 5]. Moreover, when the degree of the ambiguity varies, the estimated formants often change arbitrarily. Also, high-order resonances often can be mis-identified as low-order ones if the low-frequency energies are weak or missing, despite physically plausible resonances in the low-frequency region.

In this paper, we report a new approach to estimating VTRs in fluent speech utterances, which overcomes the above problems in many current techniques. We note that estimation of VTRs during consonantal closure has not been adequately dealt with in the existing literature on formant tracking. Our approach is to use the state equation in the hidden dynamic model of speech, which was applied to speech recognition in the past [6, 7], as the constraint to incorporate the prior knowledge of the target-directed dynamic behavior of VTR in absence of measured acoustic data. When the targets associated with consonant closure and constriction are appropriately initialized, according to the physical plausible resonance values, the prior constraint is able to balance the negative evidence in acoustics of lacking spectral prominances. As we will show in the results section, this balance will be automatically provided by the Kalman gain in the Extended Kalman Filter (EKF) algorithm.

The organization of this paper is as follows. In Section 2, we will outline the hidden dynamic model used in this work. The approximate EM algorithm developed in this work is presented in Section 3 that utilizes an off-the-shelf formant tracker to simplify the intricate and rigorous sufficient statistics that would be required in the exact EM algorithm. The VTR recovery algorithm based on EKF is described and analyzed in Section 4, and experimental results presented in Section 5.

2. VTR dynamics and neural-net mapping

A causal and linear first-order "state" equation is used to describe the VTR dynamics (f_1, f_2, and f_3) according to

$$\mathbf{z}(k+1) = \mathbf{\Phi}^j \mathbf{z}(k) + (\mathbf{I} - \mathbf{\Phi}^j)\mathbf{t}^j + \mathbf{w}(k), \quad j = 1, 2, ..., J_P \quad (1)$$

where $\mathbf{z}(k) = [f_1(k) \quad f_2(k) \quad f_3(k)]'$ is the three-dimentioanl "state" vector at discrete time step k, $\mathbf{\Phi}^j$ and $\mathbf{t}^j$ are the system matrix and target vector associated with phone j. Both $\mathbf{\Phi}^j$ and $\mathbf{t}^j$ are a function of time k via their dependence on j. $\mathbf{w}(k)$ is the discrete-time state noise, modeled by an IID, zero-mean, Gaussian process with covariance matrix $\mathbf{Q}$.

The observation equation in the model is nonlinear, noisy, and static, and is described by

$$\mathbf{o}(k) = h^{(r)}[\mathbf{z}(k)] + \mathbf{v}(k), \quad (2)$$

where the acoustic observation $\mathbf{o}(k)$ consists of MFCC (Mel-Frequency Cepstral Coefficients) measurements, $\mathbf{v}(k)$ is the additive observation noise modeled by an IID, zero-mean, Gaussian

process with covariance matrix $\mathbf{R}$. The multivariate nonlinear mapping, $h^{(r)}[\mathbf{z}(k)]$, is implemented by multiple switching MLP (Multi-Layer Perceptron) neural networks, with each MLP associated with a distinct manner (r) of articulation of a phone.

A three-layer feedforward MLP is implemented with a linear activation function on the output layer and the antisymmetric hyperbolic tangent function on the hidden layer:

$$h_i(\mathbf{z}) = \sum_j W_{ij} \cdot g_j(\sum_l w_{jl} \cdot \mathbf{z}_l), \tag{3}$$

where i, j, and l are the indices of output, hidden, and input units, respectively, and $g(\cdot)$ is the hyperbolic tangent function.

3. Approximate EM algorithm for learning

A version of the (generalized) EM algorithm involving the observed MFCC data as the input to the algorithm was developed and analyzed in [6, 8]. The E-step requires a rather complex process of computing sufficient statistics via the EKF, and the degree of accuracy cannot be easily assessed. One main contribution of this work is to simplify the computation of the sufficient statistics via the use of the ESPS/Wavesurfer formant tracker, a public, off-the-shelf tool [9]. In this section, we will first show the more rigorous version of the algorithm, and then show how the use of the formant tracker drastically simplifies the solution.

3.1. Auxiliary Q function

Denote the acoustic observation sequence by $\mathbf{o} = [\mathbf{o}(1), \mathbf{o}(2), \ldots, \mathbf{o}(N)]$ and hidden VTR sequence by $\mathbf{z} = [\mathbf{z}(1), \mathbf{z}(2), \ldots, \mathbf{z}(N)]$. The E step involves evaluating the following conditional expectation:

$$\begin{aligned} Q(\mathbf{z}, \mathbf{o}, \boldsymbol{\theta}) \quad = \quad & \underbrace{-\frac{N}{2} \log\{|\frac{1}{N} \sum_{k=0}^{N-1} E[\mathbf{e}_{k1} \mathbf{e}'_{k1} | \mathbf{o}, \boldsymbol{\theta}]|\}}_{Q_1(\mathbf{z}, \mathbf{o}, \boldsymbol{\Phi}, \mathbf{t})} \\ & \underbrace{-\frac{N}{2} \log\{|\frac{1}{N} \sum_{k=1}^{N} E[\mathbf{e}_{k2} \mathbf{e}'_{k2} | \mathbf{o}, \boldsymbol{\theta}]|\}}_{Q_2(\mathbf{z}, \mathbf{o}, W_{ij}, w_{jl})} + \mathbf{C}. \end{aligned} \tag{4}$$

where $\mathbf{e}_{k1} = \mathbf{z}(k+1) - \boldsymbol{\Phi}\mathbf{z}(k) - (\mathbf{I} - \boldsymbol{\Phi})\mathbf{t}$ and $\mathbf{e}_{k2} = \mathbf{o}(k) - h(\mathbf{z}(k))$.

3.2. Model parameter updates

The M step of the EM algorithm aims at optimizing the Q function in (4) with respect to model parameters $\boldsymbol{\theta}$=$\{\mathbf{t}, \boldsymbol{\Phi}, W_{ij}, w_{jl}\}$. This yields the following stationary optimization points for parameters of $\mathbf{t}$, $\boldsymbol{\Phi}$:

$$N\boldsymbol{\Phi}\mathbf{t}\mathbf{t}' - \boldsymbol{\Phi}\mathbf{t}\mathbf{A}' - \boldsymbol{\Phi}\mathbf{A}\mathbf{t}' - N\mathbf{t}\mathbf{t}' + \mathbf{t}\mathbf{A}' + \mathbf{B}\mathbf{t}' + \boldsymbol{\Phi}\mathbf{C} - \mathbf{D} = 0,$$

and

$$N\boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{t} - \boldsymbol{\Phi}'\boldsymbol{\Phi}\mathbf{A} - N\boldsymbol{\Phi}'\mathbf{t} - N\boldsymbol{\Phi}\mathbf{t} + \boldsymbol{\Phi}'\mathbf{B} + \boldsymbol{\Phi}\mathbf{A} + N\mathbf{t} - \mathbf{B} = 0.$$

where

$$\mathbf{A} = \sum_{k=0}^{N-1} E[\mathbf{z}(k) | \mathbf{o}, \boldsymbol{\theta}], \qquad \mathbf{B} = \sum_{k=0}^{N-1} E[\mathbf{z}(k+1) | \mathbf{o}, \boldsymbol{\theta}], \tag{5}$$

$$\mathbf{C} = \sum_{k=0}^{N-1} E[\mathbf{z}(k)\mathbf{z}(k)' | \mathbf{o}, \boldsymbol{\theta}], \ \mathbf{D} = \sum_{k=0}^{N-1} E[\mathbf{z}(k+1)\mathbf{z}(k)' | \mathbf{o}, \boldsymbol{\theta}].$$

The coefficients $\mathbf{A}, \mathbf{B}, \mathbf{C}$, and $\mathbf{D}$ above constitute the sufficient statistics for parameter optimization, where all the conditional expectations may be computed via EKF. In this work, we simplify such computation by approximating the conditional expectations above by the results of ESPS formant tracker, denoted by $\bar{\mathbf{z}}(k)$. This gives

$$\mathbf{A} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k), \qquad \mathbf{B} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k+1),$$

$$\mathbf{C} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k)\bar{\mathbf{z}}(k)', \ \mathbf{D} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k+1)\bar{\mathbf{z}}(k)' \tag{6}$$

when solving for $\mathbf{t}$ and $\boldsymbol{\Phi}$. In this study the target vector values were initialized and then fixed.

Using the same approximation above, we approximate the gradients of the Q function with respect to the MLP weights as

$$\frac{\partial Q_2}{\partial W_{ij}} \quad \propto \quad \sum_{k=1}^{N} [\mathbf{o}(k) - h(\bar{\mathbf{z}}(k))]' \frac{\partial h(\bar{\mathbf{z}}(k))}{\partial W_{ij}} \tag{7}$$

$$\frac{\partial Q_2}{\partial w_{jl}} \quad \propto \quad \sum_{k=1}^{N} [\mathbf{o}(k) - h(\bar{\mathbf{z}}(k))]' \frac{\partial h(\bar{\mathbf{z}}(k))}{\partial w_{jl}}. \tag{8}$$

This approximation amounts to treating $\bar{\mathbf{z}}(k)$ as the input to the MLP and the observation $\mathbf{o}(k)$ as the output of the MLP. And the gradients expressed in (7) and (8) can be shown to be exactly the same as those in the backpropagation algorithm.

The state and observation covariances, $\mathbf{Q}$ and $\mathbf{R}$, are important model parameters that need to estimated directly from:

$$\mathbf{Q} = \frac{1}{N} \sum_{k=0}^{N-1} E[\mathbf{e}_{k1} \mathbf{e}'_{k1} | \mathbf{o}, \boldsymbol{\theta}], \ \ \mathbf{R} = \frac{1}{N} \sum_{k=0}^{N-1} E[\mathbf{e}_{k2} \mathbf{e}'_{k2} | \mathbf{o}, \boldsymbol{\theta}]$$

where

$$\begin{aligned} E[\mathbf{e}_{k1} \mathbf{e}'_{k1} | \mathbf{o}, \boldsymbol{\theta}] \quad = \quad & \mathbf{G} - \mathbf{D}\boldsymbol{\Phi}' - \mathbf{B}\mathbf{t}' - \boldsymbol{\Phi}\mathbf{D}' + \boldsymbol{\Phi}\mathbf{C}\boldsymbol{\Phi}' \\ & + \boldsymbol{\Phi}\mathbf{A}\mathbf{t}' - \mathbf{t}\mathbf{B}' + \mathbf{t}\mathbf{A}'\boldsymbol{\Phi}' + N\mathbf{t}\mathbf{t}' \end{aligned} \tag{9}$$

$$E[\mathbf{e}_{k2} \mathbf{e}'_{k2} | \mathbf{o}, \boldsymbol{\theta}] \quad = \quad [\mathbf{o}(k) - h(\bar{z}(k)][\mathbf{o}(k) - h(\bar{z}(k)]' \tag{10}$$

and

$$\mathbf{G} \approx \sum_{k=0}^{N-1} \bar{\mathbf{z}}(k+1)\bar{\mathbf{z}}(k+1)'.$$

4. Kalman filtering for VTR recovery

Given the trained model parameters just described from a set of training data, for any new utterance we can apply the EKF algorithm to the model for state estimation — to recover the VTR state sequence. In this section, we outline the computation steps and give insight into how the algorithm strikes an optimal balance between the exploitation of the state equation (as the prior knowledge on the VTR dynamics with no observation data) and of the observation equation making direct use of the MFCC acoustic observation data.

Denote by $\hat{\mathbf{z}}(k|k)$ the EKF state estimate and by $\hat{\mathbf{z}}(k+1|k)$ the one-step EKF state prediction. Then the prediction step of the EKF is

$$\hat{\mathbf{z}}_{k+1|k} = \boldsymbol{\Phi}\hat{\mathbf{z}}_{k|k} + (\mathbf{I} - \boldsymbol{\Phi})\mathbf{t}. \tag{11}$$

where the parameters $\mathbf{t}$ and $\boldsymbol{\Phi}$ are known after the training given the phone sequence and phone boundaries. (Dependency of the

parameters on the phone segment is omitted here for notational simplicity.) The intuition behind (11) is that the state predictor for time $k+1$ based on the current EKF state estimate at time k will move the VTR towards the target vector $\mathbf{t}$. Such desirable dynamics come directly from state equation (1), and it is, in fact, in exactly the same form as the noise-free version of the model's state equation.

Denote by $\mathbf{H}_z(\hat{\mathbf{z}}_{k+1|k})$ the Jacobian matrix for the nonlinear MLP neural network evaluated at the point of $\hat{\mathbf{z}}_{k+1|k}$. Also denote by $h(\hat{\mathbf{z}}_{k+1|k})$ the MLP output for the input $\hat{\mathbf{z}}_{k+1|k}$. Then, we have the EKF corrector of

$$\hat{\mathbf{z}}_{k+1|k+1} = \hat{\mathbf{z}}_{k+1|k} + \mathbf{K}_{k+1}\Big\{\mathbf{o}(k+1) - h(\hat{\mathbf{z}}_{k+1|k})\Big\}, \quad (12)$$

where $\mathbf{K}_{k+1}$ is the (time-varying) Kalman gain. Kalman gain is a function of Jacobian matrix $\mathbf{H}_z$ and of all other model parameters. In particular, the greater the variance in the observation equation $\mathbf{R}$ is, the smaller Kalman gain $\mathbf{K}$ becomes according to the following relationship:

$$\mathbf{K} = \mathbf{F}_1\Big\{\mathbf{F}_2 + \mathbf{R}\Big\}^{-1}, \quad (13)$$

where $\mathbf{F}_1$ and $\mathbf{F}_2$ are quantities not directly related to $\mathbf{R}$.

We now offer the intuition on how the EKF algorithm optimizes the balanced use of prior knowledge of VTR dynamics in (11) and of the MFCC observation sequence $\mathbf{o}$ in (12). For each time frame, the predictor step in (11) first applies prior knowledge to direct the VTR estimate towards the phone target. The error in the MFCC domain between the observation $\mathbf{o}(k+1)$ and the MLP output $h(\hat{\mathbf{z}}_{k+1|k})$ is used in the second, corrector step in (12) to adjust the VTR estimate. The adjustment is proportional to the error, and the proportionality constant is precisely the Kalman gain. When variance $\mathbf{R}$ is large, the observation and/or the MLP prediction become less reliable; hence, the proportionality constant for the error adjustment in correcting the VTR estimate should be small. The intuitatively appealing relationship described above in qualitative terms is governed precisely by the quantitative dependence of Kalman gain $\mathbf{K}$ on $\mathbf{R}$ in (13).

5. Experimental setup and results

We have implemented the approximate EM algorithm for parameter learning and the EKF algorithm for VTR tracking using the TIMIT corpus. The full training set in TIMIT are used for the training, and the disjoint test set used for the VTR recovery from the MFCC data.

The implmentation of the MLP neural network training in our experiments is as follows. The observation data are 13-dimensional static MFCC vectors. The lowest three formants are computed from the ESPS/WaveSurfer software to form the 3-dimensional state vector. Thus the non-linear mapping function, $h^r(\boldsymbol{z}(k))$, requires a 3-input, 13-output MLP. One unique 3-input, 12-hidden, 13-output MLP is used for each distinct manner of articulation in implementing the mapping function. We used the MATLAB neural network toolbox to train each MLP, with Bayesian regularisation [10] to improve generalizability of the training and preprocessing of the input and output data [11] to improve the training efficiency.

After the training, the test data, with the segment sequence and the boundaries provided from TIMIT, are subject to VTR tracking using the EKF described in Section 4.

In Figs. 1 and 2, we show the VTR tracking results, in comparison with the ESPS/Wavesurfer formant estimates, for a test TIMIT utterance (sx383) from a male and a female speaker, respectively. The utterance reads: *"The carpet cleaners shampooed our oriental rug."* The results are superimposed on the respective spectrograms to facilitate the examination of the quality. Plots (a) in both Figs. 1 and 2 are the ESPS/Wavesurfer formant tracks, which are seen from spectrogram comparison to be quite accurate for all vocalic segments. However, during most of the consonantal segments, the formant estimates vary widely, with little correlate to the underlying resonances. Indeed, these resonances do not manifest themselves directly as spectra peaks in the acoustics, and hence could not be picked up by the tracker based solely on the acoustic information. In contrast, the EKF algorithm described in Section 4 makes use of prior knowledge of VTR dynamics as the state equation, in addition to the use of acoustic information represented as MFCCs.

The recovered VTR sequences from the MFCCs using the EKF algorithm are shown in plots (b) of Figs. 1 and 2, where Kalman gain $\mathbf{K}_k$ is determined optimally according to the algorithm. These results represent the optimal balance between the use of prior VTR dynamic information and of the acoustic observation. They are not only accurate during vocalic segments where the spectra reveal the resonances, but most importantly, the recovered VTRs are also meaningful during consonant closure and constriction where the spectra do not directly reveal the underlying resonances. These physically meaningful VTRs during consonants are particularly useful for accurately directing formant/VTR trajectories in rapic consonant-vowel or vowel-consonant transitions. The knowledge source provided to our algorithm to overcome weak or no information in the acoustic measurement for infering the VTR values is the state equation (1). This is utilized directly in the Kalman predictor step (11) of the EKF algorithm.[1]

To illustrate the nature of this knowledge source, we artificially set Kalman gain $\mathbf{K}_k$ to zero, thereby eliminating the information provided by MFCCs entirely, and show the VTR tracking results in plots (c) of Figs. 1 and 2. (The vertical lines here indicate the segment boundaries for the state equation derived from the TIMIT segmentation data.) We observe that while frequent mismatches occur with the true VTRs, the VTRs predicted from state equation alone are reasonably valued, especially in consonant segments, to constrain the estimated VTRs from undesirably widely varying as shown on plots (a) for the conventional formant tracker.

6. Discussion and conclusion

This paper presents a new approach to tracking VTR in fluent speech, which is shown to be superior to the conventional formant trackers (such as the ESPS/WaveSurfer tool) during consonant closure when only the front cavity's resonances in the vocal tract are activated. The new approach overcomes the common difficulties of tracking formants and resonance frequencies in consonants and in rapid speech sound transitions when the acoustic signal (in the form of spectral prominences) provides weak or no information to infer the underlying resonances in these sounds.

Our approach is based on the prior knowledge of the target-directed VTR dynamics, in conjunction with speech acoustics using the EKF algorithm for optimal combination of these two information sources. While this philosophy is the same as that used in the earlier work of [14], the approaches differ in how the rela-

[1]The specific VTR target values used in producing results of Fig. 1 (male) are taken from Tables 10.1 and 10.2 (page 364) in Chapter 10 of [12]. They are modified from a synthesizer setup [13]. The target values for Fig. 2 (female) are set to be 20% higher than those for Fig. 1.

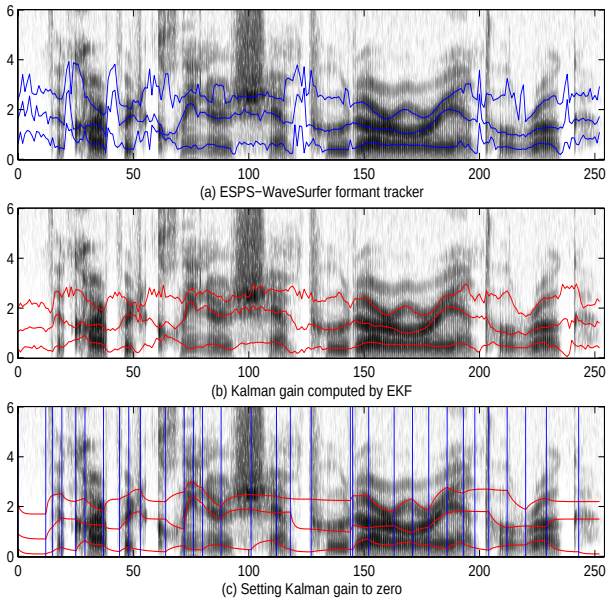

Figure 1: (a) ESPS/Wavesurfer formant estimates, with widely varying values during consonants; (b) Recovered VTRs from MFCCs when Kalman gain is determined optimally according to the EKF algorithm. (c) Recovered VTRs after setting Kalman gain to zero to illustrate the target-directed constraint provided by the state equation. All the results in plots (a), (b), and (c) are from a male speaker in TIMIT database.

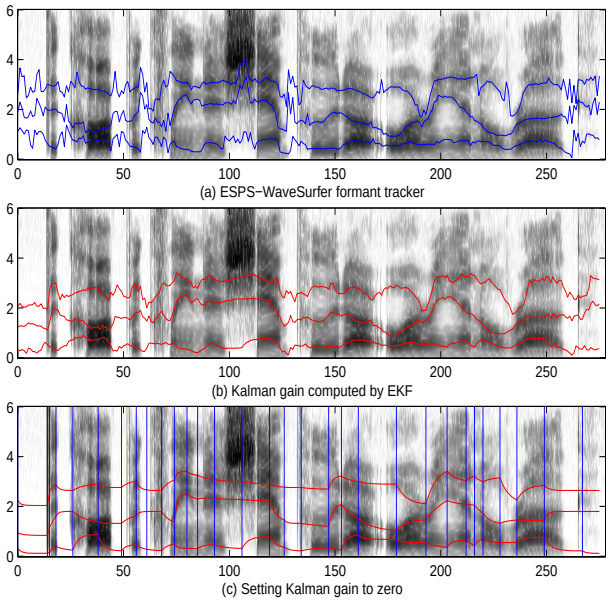

Figure 2: Same as Figure 1 except the results are from a female speaker for the same utterance.

tionship is represented between the hidden VTR domain and the acoustic observation domain and in how the optimal combination rule is. The neural-network mapping approach described in this paper is a highly flexible one and has the potential to incorporate other effective acoustic features beyond the currently used MFCC measures.

One specific contribution of this work is the development of the approximate EM algorithm for learning MLP weights. This has been shown to be effective in representing the forward mapping from VTR to MFCCs as evaluated in the EKF framework. Since the input to the MLP during the training is taken from the ESPS/WaveSurfer tool, which produces many incorrect incorrect values during consonants, we expect that iterating the VTR tracking (by the EKF) and MLP training (with inputs from the tracking results) will significantly improve the quality of the final tracking results.

7. References

[1] J. Allen, M. S. Hunnicutt, and D. Klatt, *FROM TEXT TO SPEECH: The MITalk System*, Cambridge University Press, Cambridge, England, 1987.

[2] D. Klatt. "Software for a cascade/parallel formant synthesizer," *J. Acoust. Soc. Am.*, Vol. 67, 1980, pp. 971-995.

[3] K. Stevens and C. Bickley. "Constraints among parameters simplify control of Klatt formant synthesizer," *J. Phonetics,* Vol. 19, 1991, pp. 161-174.

[4] S. McCandless. "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust. Speech and Signal Proc.*, Vol. 22, 1974, pp. 135-141.

[5] G. Kopec. "Formant tracking using HMMs and vector quantization," *IEEE Trans. Acoust. Speech and Signal Proc.*, Vol. 34, 1986, pp. 709-729.

[6] 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, No. 6, 2000, pp. 3036-3048.

[7] F. Seide, J. Zhou, and L. Deng. "Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM — MAP decoding and evaluation," *Proc. ICASSP*, 2003, pp. 748-751.

[8] R. Togneri and L. Deng. "Joint state and parameter estimation for a target-directed nonlinear dynamic system model," *IEEE Trans. Signal Processing* , Vol. 51, Dec. 2003, pp. 3061-3070.

[9] K. Sjolander. "Recent developments regarding the Wavesurfer speech tool", 2002. http://www.speech.kth.se/qpsr/tmh/2002/02-44-053-056.pdf

[10] F. D. Foresee and M. T. Hagan, "Gauss-Newton approximation to Bayesian regularization," *Proc. Intern. Joint Conf. Neural Networks*, 1997, pp 1930-1935.

[11] S. Haykin, *NEURAL NETWORKS: A Comprehensive Foundation*, Prentice-Hall, 1999.

[12] L. Deng and D. O'Shaughnessy. *SPEECH PROCESSING — A Dynamic and Optimization-Oriented Approach,* Marcel Dekker: New York, NY, 2003.

[13] K. Stevens. Course Notes: "Speech Synthesis with a Formant Synthesizer", Cambridge, MA, July, 1993.

[14] L. Deng, I. Bazzi, and A. Acero. "Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint," *Proc. Eurospeech,* 2003, pp. 73-76.