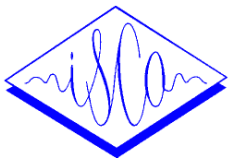

Second ESCA/IEEE Workshop
on Speech Synthesis
Mohonk, New Paltz, NY, USA
September 12-15, 1994

A Dynamical System Model for Generating F_0 for Synthesis

Ken Ross Mari Ostendorf
Boston University, 44 Cummington St., Boston, MA 02215 USA

Abstract

This work develops a new model of fundamental frequency (F_0) generation that incorporates traditional methods of F_0 modeling, but also has parameters that can be automatically estimated from prosodically labeled speech. We generate F_0 with a state-space dynamical system model which assumes that there is an unobserved state vector corresponding to the noisy observation of F_0 and energy. Parameters of the model are specified to capture segment, syllable, and/or phrase level effects. Since there are missing observations corresponding to the state vector and unvoiced segments, we use a non-traditional method for parameter estimation based on an EM algorithm developed for speech recognition applications. In experiments on an independent test set, we obtained a rms error of 33 Hz for F_0.

1. Introduction

This work is directed toward developing a new model for fundamental frequency (F_0) generation that incorporates both the target and the filtering approaches to F_0 modeling and that has parameters that could be automatically estimated from prosodically labeled speech. Additional benefits of the model proposed here are that it can jointly represent energy and F_0 contours and that it can be used for both synthesis and recognition. Here we explore its application to speech synthesis.

2. Dynamical System Model

Our approach to generate F_0 and energy is a state-space dynamical system model. The dynamical system model has been used for many different types of modeling, including speech recognition (Digalakis *et al.*, 1993).

2.1. Model Structure

The dynamical system model represents an unobserved state vector, $\mathbf{x}_k$, and a noisy vector observation of that state, $\mathbf{y}_k$. In this case, the observation vector contains values for the F_0 and energy. The general form for the discrete-time, state-space model is as follows:

$$\mathbf{x}_{k+1} = \mathrm{F}_k\mathbf{x}_k + \mathbf{u}_k + \mathbf{w}_k \quad (1)$$

$$\mathbf{y}_k = \mathrm{H}_k\mathbf{x}_k + \mathbf{b}_k + \mathbf{v}_k \quad (2)$$

where $\mathbf{u}_k$ is a time-varying, additive input to the state equation (Equation 1), $\mathbf{b}_k$ is an additive input to the observation equation (Equation 2), and $\mathbf{w}_k$ and $\mathbf{v}_k$ are zero-mean, uncorrelated Gaussian noises with covariances Q and R, respectively. Furthermore, the initial state is assumed to be a Gaussian random variable that is independent of the observation and state noises.

This approach of using a dynamical system for modeling F_0 and energy can be thought of as a variation of a hybrid target/filter model formed by a network of filters (Hirose *et al.*, 1982, Anderson *et al.*, 1984). The targets are specified generically by $\mathbf{b}_k$ and filtered by

F_k, with pitch range and baseline controls entering in the observation equation rather than in the target specification. The dynamical system model has advantages over other hybrid target/filter approaches in that it provides for a joint model of F_0 and energy, it can be automatically trained to jointly account for the different sources of variability in F_0 and energy contours, and it can be used easily in recognition as well as synthesis. In addition, the observation and state noises are useful for modeling not only system disturbances and noise corruptions (such as F_0 tracking errors), but also the uncertainties inherent in the mathematical model itself.

An important part of development of the dynamical system model of F_0/energy is the proper selection of model parameter dependencies. This model incorporates three levels of phenomena that affect the F_0 contour: (1) the intermediate phrase, (2) the syllable, and (3) the phoneme. The basic model is at the syllable level where a syllable is categorized according to its abstract prosodic label (labels are based upon the ToBI labeling system (Silverman *et al.*, 1992)). The phoneme-level dependency incorporates segmental effects into the model. The phrase level parameters represent the position of the syllable within the phrase. To simplify the estimation problem, higher level range effects are removed by normalizing the F_0 and energy with respect to the peak for the intermediate phrase. Later, F_0 and energy range will be predicted based on paragraph-level and simple discourse information. Ideally all of the model parameters would depend on all three levels of phenomena and the estimation algorithm would find the correct way to incorporate these phenomena into the model. But, because of limited training data and problems with local optima, we limit the dependencies based upon linguistic theory. Equally important to model quality is the selection of initial conditions.

The dynamical system equations with the parameter dependencies explicitly marked were as follows:

$$\mathbf{x}_{k+1} = \mathrm{F}(\alpha_k)\,\mathbf{x}_k + \mathbf{u}(\alpha_k, \gamma_k) + \mathbf{w}_k \tag{3}$$

$$\mathbf{y}_k = \mathrm{H}(\beta_k, \gamma_k)\,\mathbf{x}_k + \mathbf{b}(\beta_k, \alpha_k, \gamma_k) + \mathbf{v}_k \tag{4}$$

where β_k represents the position of the syllable containing segment k within the intermediate phrase, α_k represents the prosody of the syllable containing segment k, and γ_k represents the phoneme class of segment k. Additive effects are incorporated into the model through $\mathbf{u}(\alpha_k, \gamma_k)$ which is an additive input to the state equation (Equation 3) and $\mathbf{b}(\beta_k, \alpha_k, \gamma_k)$ which is an additive input to the observation equation (Equation 4). The noise components of the model come from the terms $\mathbf{w}_k$ and $\mathbf{v}_k$ which are zero-mean, uncorrelated Gaussian noises with covariances:

$$E\{\mathbf{w}_k \mathbf{w}_l^T\} = \mathrm{Q}(\alpha_k)\,\delta_{kl} \tag{5}$$

$$E\{\mathbf{v}_k \mathbf{v}_l^T\} = \mathrm{R}(\beta_k)\,\delta_{kl} \tag{6}$$

where δ_{kl} is the Kronecker delta. Parameters F, u, H, and R are also conditioned on the region within the syllable. Parameter sharing is used for the noise covariances to allow the data to be used more efficiently (across the syllable regions for Q and across phoneme classes for R).

2.2. Model Training and Synthesis

Finding the dynamical system parameters automatically with traditional techniques is difficult, particularly since observations are missing in the unvoiced regions. Instead, we use a non-traditional iterative method for parameter estimation based upon an algorithm developed for speech recognition (Digalakis *et al.*, 1993). The algorithm relies upon the idea that, if the state was observable, the model parameters would be relatively easy to estimate. This approach to parameter estimation uses the two-step, iterative expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) for maximum likelihood estimation of processes with

unobserved components, which in this case are the state vectors and unvoiced data. The EM algorithm uses a multivariate regression to find the maximum likelihood solution for the dynamical system. The required statistics for $\mathbf{x}_k$ are computed with the fixed interval form of the Kalman smoothing filter using the Rauch, Tung, and Striebel algorithm (Rauch *et al.*, 1965). This algorithm uses the standard forward recursions for a Kalman filter (Kalman, 1960) followed by a backward pass.

Once this model has been trained from a labeled database, it can be used for generating F_0 and energy contours for new text. The model is operated in the generation mode by setting the model's random noise components to zero and controlling the model with the appropriate sequence of phoneme labels, durations, and abstract prosodic labels. For the case where the model has been trained on data normalized for pitch (and/or energy) range, the resulting values must be scaled according to the predicted (or original, in this case) range parameters.

3. Experiments

This model for F_0 and energy was trained and tested using speech from speech database that has abstract prosodic labels, phonetic segment boundaries, and actual F_0 and energy contours.

3.1. Corpus

The model presented here can be trained from prosodically labeled data using any speaking style, but in this case, Boston University's radio news corpus was used for training. This corpus is drawn from a collection of recorded FM radio news broadcasts spoken by seven radio announcers associated with WBUR, a public radio station in Boston. The subset of the data used here is annotated and digitized and consists of the speech from one female radio announcer contained in 23 stories. These stories are studio recordings of actual radio broadcasts that were hand labeled for pitch accents, boundary tones, and break indices according to the ToBI labeling system (Silverman *et al.*, 1992). Also, the text was automatically annotated with part-of-speech labels, and lexical stress assignments were available from an on-line dictionary. Phonetic labels and segment boundaries were obtained automatically using the Boston University speech recognition system (Kimball *et al.*, 1992). The 23 radio news stories contain 6009 words, 10028 syllables, and 3384 pitch accents. Excluding silences, this speech has a duration of approximately 34 minutes. For testing the model, a smaller set (approximately 8 minutes) of radio news stories read by the same speaker is available. All of the speech data has F_0 and rms energy contours that were created by the Entropic Waves version 5.0 pitch tracker.

3.2. Results

The fundamental frequency was normalized by subtracting the minimum F_0 value for the sentence and then dividing by the peak F_0 during the intermediate phrase. The energy was normalized by dividing by the energy corresponding to the phrase's peak F_0. For the model parameters, there were five classes of phonemes: stressed vowel, unstressed vowel, sonorant, voiceless obstruent, and voiced obstruent. Phrases were broken up into three regions: beginning, middle, and end. Each syllable was divided up into six regions by a linear mapping. This model used seven categories of pitch accents (none, X*, H*, !H*, L*, L+H*, and L+!H*) and seven categories of phrase tones (none, L-L%, H-L%, L-H%, L-, H-, and !H-). Limited amounts of data forced us to combine rarer prosodic features with more common ones such as H-H%'s with L-H%'s, %H's with H*'s, and only three types of pitch accents (none, X*, or H*) were allowed on syllables that also had phrase tones. After 20 iterations of the EM algorithm, the model operating in generation mode fit the data with

rms errors of 30 Hz for the F_0 contours. Testing this model on the independent test set resulted in rms errors of 33 Hz.

The F_0 model was also tested with a perceptual test using the AT&T text-to-speech synthesizer. For this test, 16 participants were asked to rate 15 sentences from radio news stories that were synthesized using our F_0 model with the actual prosodic labels (and a reduced F_0 range), the synthesizer's F_0 model with the actual prosodic labels, and the synthesizer's default F_0 (all intonational contours were produced using the same word pronunciations and durations). The participants listened to the 3 versions of each sentence as many times as they wanted and were asked to rate the naturalness on a scale of 1 to 5 with 1 representing the most natural. Our F_0 model had a mean rating of 2.6 and the AT&T default contour with actual prosodic labels had a mean rating of 3.0 with an average difference between the two of 0.47 which indicates that our F_0 model is significantly better ($p < 5 \times 10^{-4}$, $t = 4.9$).

4. Conclusion

A model has been presented for the generation of fundamental frequency (F_0) and energy for speech synthesis. This model is automatically trainable from a prosodically labeled and aligned speech database. We have shown that this approach is capable of producing good approximations to observed F_0 contours using lexical information and abstract prosodic labels. Using this method of generating F_0 in conjunction with a model for predicting the abstract prosodic labels from text can produce a text-to-speech synthesizer with a more natural intonation. This model is also able to generate energy contours which will be evaluated for their impact on the naturalness of text-to-speech synthesis in future work.

Acknowledgments

This work was supported by a research grant provided by NYNEX.

References

[1] M. D. Anderson, J. B. Pierrehumbert and M. Y. Liberman. "Synthesis by rule of English intonation patterns," *Proc. of ICASSP*, 1984, pp. 2.8.1–2.8.4.

[2] 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. 37, No. 1, 1977, pp. 1–38.

[3] V. Digalakis, J. R. Rohlicek and M. Ostendorf. "ML Estimation of a Stochastic Linear System with the EM Algorithm and its Application to Speech Recognition," *IEEE Trans. on Speech and Audio Proc.*, October 1993, pp. 431–442.

[4] K. Hirose and H. Fujisaki. "Analysis and Synthesis of Voice Fundamental Frequency Contours of Spoken Sentences," *Proc. of ICASSP*, 1982, pp. 950–953.

[5] R. E. Kalman. "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering (ASME)*, March 1960, pp. 35–45.

[6] O. Kimball, M. Ostendorf and I. Bechwati. "Context Modeling with the Stochastic Segment Model," *IEEE Trans. on Signal Proc.*, June 1992, pp. 1584–1587.

[7] H. E. Rauch, F. Tung and C. T. Striebel. "Maximum Likelihood Estimates of Linear Dynamic Systems," *AIAA Journal*, Vol. 3, No. 8, August 1965, pp. 1445–1450.

[8] K. Silverman *et al.*"ToBI: A Standard for Labeling English Prosody," *Proc. of ISCLP*, October 1992, pp. 867–870.