PARAMETER EXTRACTION OF A QUANTITATIVE INTONATION MODEL WITH WAVELET ANALYSIS AND EVOLUTIONARY OPTIMIZATION

Hans Kruschke, Andreas Koch

Dresden University of Technology
Laboratory of Acoustics and Speech Communication
D-01062 Dresden, Germany

ABSTRACT

State-of-the-art speech technology requires computation of large amount of data. This is only possible with reliable algorithms for automatic data analysis. This contribution therefore deals with automatic extraction of the parameters of a quantitative intonation model developed by Fujisaki and his coworkers. For detection of accent and phrase commands of this model frequency analysis based on Wavelet Transform is proposed. Furthermore an Evolution Strategy is used to optimize the model parameters of obtained first-order approximation. First results show that the quality of extracted parameters is comparable to reference data.

1. INTRODUCTION

Analysis of intonation and intonation generation are important issues in speech research and especially in speech synthesis. Therefore several intonation models have already been proposed. Among them the quantitative model developed by Fujisaki and his coworkers [1]. This model is based on physiological interpretation of the shape of fundamental frequency (F0) contours. The original development was done for common Japanese but the model has shown to be applicable to many other languages as well. If the model parameters are well estimated then F0 contours will be generated which are quite close to measured F0 contours. The problem for a widely-used application of the model is a lack of reliable algorithms for automatic extraction of model parameters for a given F0 contour. Since the parameter extraction can not be done analytically proposed methods follow a multistage procedure of successive approximation. The main steps are (1) Correction of gross errors of the measured F_0 contour, removal of microprosody and interpolation of unvoiced segments, (2) Estimation of a first-order approximation of model parameters and (3) Optimization of the first-order approximation. The steps may be repeated and recombined. Approaches following this scheme are [2] and [3]. This contribution proposes new methods for steps (2) and (3). After an introduction of the quantitative model the proposed methods will be introduced in section 3 and section 4 respectively. Based on these methods an algorithm for automatic extraction of the model parameters was developed. This algorithm will be described in section 5.

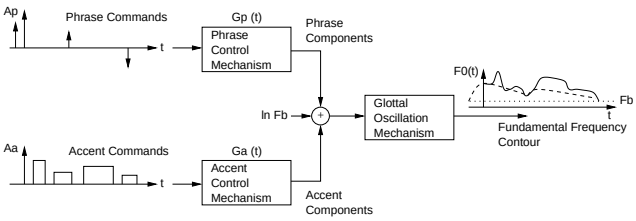

Fig. 1. A command response model for F_0 contour generation of human utterances.

2. THE QUANTITATIVE INTONATION MODEL

Development of the quantitative intonation model is based upon the observation that F0 contours are characterized by slow undulations and by relatively fast rise and fall patterns. The slow undulations roughly correspond to larger phrases and sentences whereas the rise and fall patterns correspond to lexical accents of words. Accordingly the F0 contour is generated by superposition of the slow undulations which are modeled by impulse response of a phrase control mechanism, the rise and fall patterns which are modeled by step response of an accent control mechanism and a base frequency value. The model is sketched in figure 1 and expressed by the following equations:

$$lnF_0(t) = lnFb + \sum_{i=1}^{I} Ap_i Gp(t - T_{0i}) + \sum_{j=1}^{J} Aa_j [Ga(t - T_{1j}) - Ga(t - T_{2j})] \quad (1)$$

$$Gp(t) = \begin{cases} \alpha_i^2 t \, \exp(-\alpha_i t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (2)$$

$$Ga(t) = \begin{cases} \min[1-(1+\beta_i t)\ \exp(-\beta_i t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (3)$$

The symbols indicate:

Fb	:	baseline value of fundamental frequency,
I	:	number of phrase commands,
J	:	number of accent commands,
Ap_i	:	magnitude of the *i*th phrase command,
Aa_j	:	amplitude of the *j*th phrase command,
T_{0i}	:	timing of the *i*th phrase command,
T_{1j}	:	onset of the *j*th accent command,
T_{2j}	:	offset of the *j*th accent command,
α_i	:	natural angular frequency of the *i*th phrase command,
β_j	:	natural angular frequency of the *j*th accent command,
γ	:	relative ceiling level of the accent commands (generally set to $\gamma = 0.9$).

3. WAVELET ANALYSIS OF F0 CONTOURS

Section 2 explained that F0 contours consist of superposition of relative slow and fast undulations. Hence phrase and accent components are considered as signals laying in two separate frequency bands. Therefore the idea of this work is to detect phrases and accents by the help of methods of frequency analysis. For this purpose specific characteristics of F0 contours have to be taken into account: (1) phrase and accent components are of relative low frequency ($f < 10$ Hz), (2) frequency bands of phrase and accent components are relatively close together, (3) each phrase and accent is a unique event, i.e. there are neither stationary nor quasi-stationary signals. Especially the requirement of a fine resolution in frequency range causes problems. Because of the uncertainty theorem there is always a trade-off between resolution in time and frequency domain. In the commonly used Short Time Fourier Transform (STFT) the length of the time window determines a constant resolution in frequency domain. For the discussed problem a satisfying frequency resolution requires a window length which can not provide the desired time resolution. Therefore in this investigation the Wavelet Transform (WT) is used since it provides a variable resolution in frequency range. For fast computation of the WT in this work the algorithm of Mallat is used together with a Daubechies-4-Wavelet [4].

To perform a WT with the Mallat algorithm the signal to transform must have a length of 2^n as for FFT. For the investigated speech signals a length of at least 1024 samples was chosen. If a signal segment does not fit in this length then it will be padded with the value of its last actual sample.

The described WT analysis of F0 contours shows that accents and phrases are associated with maxima in different scales. According to experimental results the 4th scale is searched for accents whereas phrases are detected in the 6th scale. A detailed description of the accent and phrase detection procedure is given in section 5. In figure 4 an example of a F0 contour and its used scales is plotted.

4. EVOLUTIONARY OPTIMIZATION OF MODEL PARAMETERS

After detection of phrase and accent commands the parameters of the this first-order approximation have to be refined. This is done in an Analysis-by-Synthesis (A-b-S) procedure. With the extracted parameters a F0 contour is generated and compared with the input F0 contour by means of the Root Mean Square Error (RMSE). To get the lowest possible RMSE the parameters are altered in a recursive optimization process. For this purpose usually a Hill Climb Search (HCS) is used [2]. The HCS optimizes the parameters in a hierarchic order, i.e. the most important parameter is altered with a preset step size while the RMSE is reduced. Afterwards the search direction is changed and later on the step size is reduced. If the optimal value of the parameter is reached the procedure will continue with the next parameter and so forth.

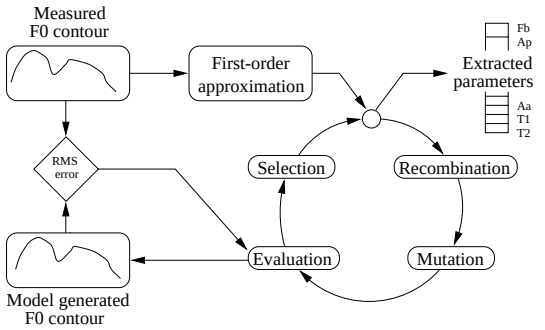

Fig. 2. Optimization of extracted model parameters with an Evolution Strategy.

In this investigation the parameter optimization in the A-b-S procedure is controlled by an Evolution Strategy (ES) (figure 2). The main principle of the used $(\mu, \kappa, \lambda, \rho)$-ES [5] is that a population of μ individuals is set up from the first-order approximation. Each individual is represented by a vector holding a set of model parameters. By this parent generation λ offspring individuals are created combining the parameters of the mating individuals. The parameters of the new individuals are now mutated in a two step process. First the mutation step size σ_i of each parameter is altered:

$$\sigma_i^{'} = \sigma_i * exp(N(0,\tau_0) + N(0,\tau_i)) \quad (4)$$

After that the parameters itself are mutated with the new step sizes:

$$x_i^{'} = x_i + \sigma_i^{'} * N(0,1) \quad (5)$$

The symbols indicate:

σ_i	:	mutation step size of the vector element i
τ_0	:	global mutation variance
τ_i	:	local mutation variance
x_i	:	ith element of the vector x
$N(0,1)$	:	normally distributed random number between $0 \ldots 1$

The RMSE between the input contour and a F0 contour generated with an individual's parameters determines the fitness of the individual. The μ best individuals are selected as parent generation for the next evolution cycle. Each individual is thereby allowed to survive a maximum of κ generations. The evolution process terminates after a specified fitness is achieved or the algorithm has run through a maximal number of generations.

One of the main advantages of the ES over the HCS is that model parameters might be optimized in parallel. It is even possible to optimize the parameters of adjacent phrase commands or a phrase command with its accents at the same time. A consideration of the left to right dependency of the model parameters also improves the used ES over the results in [6].

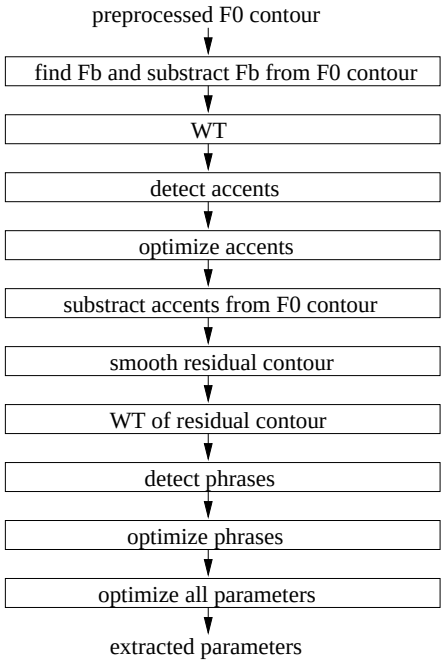

Fig. 3. Flowchart of parameter extraction algorithm.

5. OUTLINE OF THE ALGORITHM

Based on the methods described in section 3 and section 4 an algorithm for automatic extraction of the model parameters was developed. The main stages of the algorithm are presented in figure 3.

Prerequisite of the algorithm is a preprocessing of the measured F0 contour. Aims of the preprocessing are (1) removal of gross errors, (2) removal of microprosodic variations caused by production of several phonemes and (3) interpolation of discontinuations of the F0 contour caused by unvoiced phonemes and pauses. The methods applied for this purposes are described in [6]. In this work the F0 contour is additionally smoothed and stylized by piecewise polynomial approximation similar to the descriptions in [3]. The logarithm of a preprocessed F0 contour is the input to the described algorithm.

The lowest value > 0 of the input F0 contour is assigned as first-order approximation of Fb and subtracted from the entire contour. A WT is performed with the residual signal. To detect accents the 4th scale is searched for maxima. Each of this maxima has two corresponding values in the 3rd scale. The larger one of them is taken as a first refinement. The maximum of the F0 contour closest to this point is assigned as T_{2j}. The minimum of the F0 contour right before T_{2j} is assigned as T_{1j}. The contour segment between T_{1j} and T_{2j} is used to find Aa and β in a simple optimum search, i.e. within specified boundaries Aa and β are incremented with a preset step size and F0 contours are generated with this parameters. The pair of Aa and β values with the lowest RMSE between generated contour and reference contour is assigned as Aa_j and β_j. The F0 contour is searched for a minimum right after T_{2j} which is assumed as the time when the current accent disappears. The F0 segment between this point and T_{1j} serves as reference for the ES-optimization of the parameters of this accent.

With the obtained parameters of all accent commands a F0 contour is generated and subtracted from the contour of the previous stage. The residual contour is smoothed to minimize influence of errors from accent detection. A WT is performed with the smoothed residual signal. To detect phrase commands the 6th scale is searched for maxima. The corresponding F0 values are assigned as T_{max_i}. Absolute minima of the F0 contour between adjoining T_{max_i} are assigned as T_{min_i}. F0 segments between adjoining T_{min_i} are used as reference to find best pairs of Ap_i and α_i in an optimum search as described for Aa_j and β_j. After T_{max_i} is refined on the F0 contour $T_{0i} = T_{max_i} - 1/\alpha_i$ can be calculated. Parameters of the phrase commands are also optimized with an ES but in a special left to right procedure i.e. starting form the first phrase and stepping forward each time a new phrase is added to the optimization the mutation step size σ_i of all parameters of all preceding phrases is initialized with low values to allow them slight variations only. The last stage of the algorithm is to optimize accent and phrase parameters together against the input contour. This is also done in a left to right procedure there each phrase is optimized together with its accents. Figure 4 gives an example of a F0 contour and its different states in the parameter extraction algorithm.

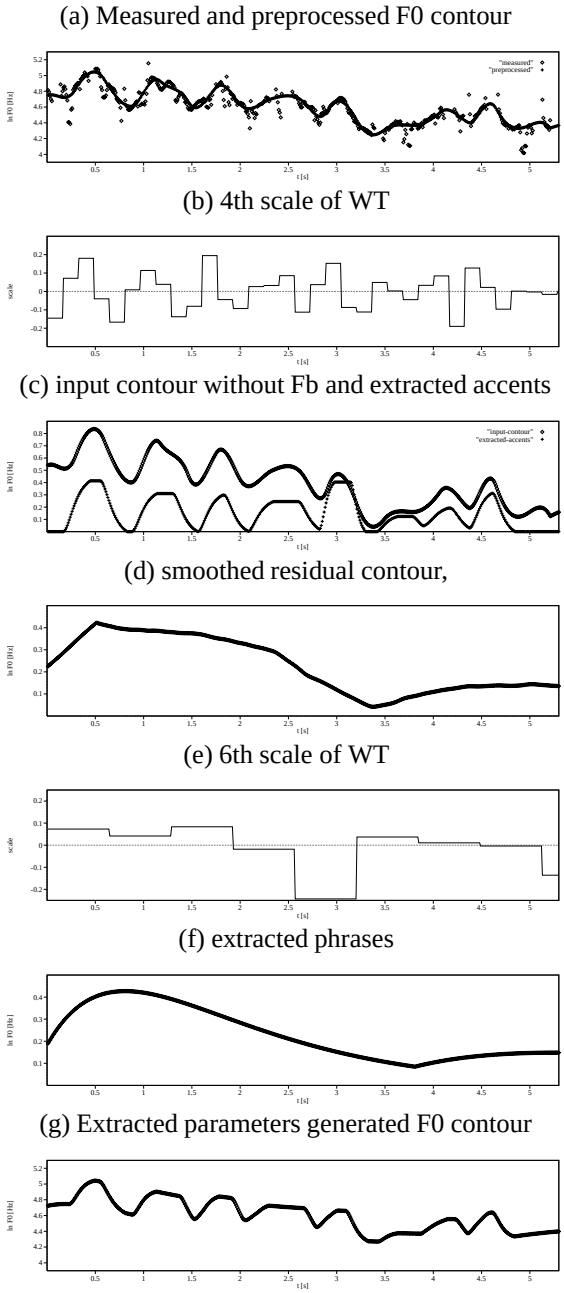

Fig. 4. F0 contour and its different states in the parameter extraction procedure. Sentence: "Israelische Soldaten haben heute an der Grenze zu Jordanien einen arabischen Freischärler erschossen." (Israeli soldiers have shot an arabic franctireur today.)

6. EXPERIMENTAL RESULTS

The proposed algorithm was tested on a German speech corpus compiled by the Institute of Natural Language Processing at the University of Stuttgart [7]. This corpus consists of 48 minutes broadcast news recordings. For this data the parameters of the quantitative intonation model were already extracted with the algorithm described in [2]. After automatic extraction the parameters were manually revised. This parameters served as reference data in the present work. Average RMSE of this parameters, measured between model generated F0 contour and gross error corrected measured F0 contour, is 4.74 Hz. Parameters extracted with the proposed algorithm achieve an average RMSE of 7.21 Hz.

7. CONCLUSION

In this contribution new methods for the automatic extraction of the parameters of a quantitative intonation model are presented. This are in particular a WT based frequency analysis of F0 contours for detection of accents and phrases and an ES to optimize the parameters of the obtained first-order approximation. Preliminary results show that parameter extraction with this methods is possible. The quality of the extracted parameters falls in the range of reference data. Further research will concentrate on improvement of the proposed algorithm.

8. REFERENCES

[1] H. Fujisaki, "Modelling in the study of tonal features of speech with application to multilingual speech synthesis," in *Joint Conference of SNLP and Oriental COCOSDA*, Hua Hin, Prachuapkirikhan, Thailand, 2002.

[2] H. Mixdorff, "A novel approach to the fully automatic extraction of fujisaki model parameters," in *Proc. ICASSP 2000*, Istanbul, 2000.

[3] S. Narusawa, H. Fujisaki, and S. Ohno, "A method for automatic extraction of parameters of the fundamental frequency contour," in *Proc. ICSLP 2000*, Beijing.

[4] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, 1998.

[5] T. Bäck and H.-P. Schwefel, "Evolution strategies I: Variants and their computational implementation," in *Genetic Algorithms in Engineering and Computer Science, Proc. First Short Course EUROGEN-95*, pp. 111–126. Wiley, Chichester, 1995.

[6] H. Kruschke, "Advances in the parameter extraction of a command-response intonation model," in *Proc. ISPACS 2001*, Nashville, 2001.

[7] S. Rapp, *Automatisierte Erstellung von Korpora für die Prosodieforschung*, Ph.D. thesis, University of Stuttgart, 1998.