Estimation of the Parameters of the Quantitative Intonation Model with Continuous Wavelet Analysis

Hans Kruschke, Michael Lenz

Laboratory of Acoustics and Speech Communication
Dresden University of Technology, Germany

Abstract
Intonation generation in state-of-the-art speech synthesis requires the analysis of a large amount of data. Therefore reliable algorithms for the extraction of the parameters of an intonation model from a given F0 contour are required. This contribution proposes improvements concerning the extraction of the parameters of the quantitative intonation model developed by Fujisaki. The improvements are mainly based on the application of the continuous wavelet transform for the detection of accents and phrases in a F0 contour. A detailed explanation of the underlying idea of this approach is given and the implemented algorithm is described. Results prove that with the proposed method a significant improvement in the accuracy of the extracted parameters is achieved. Thereby the structure and the rules of the algorithm are kept relatively simple.

1. Introduction
In speech research and especially in speech synthesis, intonation analysis and intonation generation are important issues. For this purpose Fujisaki and his coworkers developed the quantitative intonation model [1]. One of its advantages is that it is based on a physiological interpretation of the speech production process. Furthermore, it has proved to be applicable for many languages of the world. If the model parameters are well estimated, fundamental frequency (F0) contours will be generated which are close to the measured F0 contours.

In order to achieve a widely used application of the model, the problem of the extraction of its parameters needs to be solved. A direct estimation of the parameters from the F0 contour is not possible due to the setup of the model. Proposed algorithms therefore follow a multistage approach. Examples of developed algorithms are [2], [3] and [4]. The aim of this contribution is to propose improvements to the algorithm described in [4], which are mainly based on the application of the continuous wavelet transform (CWT) for accent and phrase detection. After a short presentation of the quantitative intonation model in section 2, the approach of the parameter estimation is described in section 3. In particular, the reasons for applying the CWT are explained. The algorithm implemented on basis of this method is described in section 4 and its results are given in section 5.

2. The quantitative intonation model
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 the impulse response of a phrase control mechanism, the rise and fall patterns which are modeled by the step response of an accent control mechanism and a base frequency value. The model is sketched out in figure 1 and expressed by the following equations:

\[
\ln F_0(t) = \ln F_b + \sum_{i=1}^{I} A_p_i G_p(t - T_{0i}) + \sum_{j=1}^{J} A_a_j \left[ G_a(t - T_{ij}) - G_a(t - T_{2j}) \right]
\]

\[
G_p(t) = \begin{cases} 
\alpha_t^2 t \exp(-\alpha_t t), & t \geq 0 \\
0, & t < 0
\end{cases}
\]

\[
G_a(t) = \begin{cases} 
\min[1 - (1 + \beta_j \gamma) \exp(-\beta_j \gamma)], & t \geq 0 \\
0, & t < 0
\end{cases}
\]

The symbols indicate:
- \(F_b\) : baseline value of fundamental frequency,
- \(I\) : number of phrase commands,
- \(J\) : number of accent commands,
- \(A_p_i\) : magnitude of the \(i\)th phrase command,
- \(A_a_j\) : amplitude of the \(j\)th accent command,
- \(T_{0i}\) : timing of the \(i\)th phrase command,
- \(T_{ij}\) : onset of the \(j\)th accent command,
- \(T_{2j}\) : offset of the \(j\)th accent command,
- \(\alpha_t\) : natural angular frequency of the \(i\)th phrase command,
- \(\beta_j\) : natural angular frequency of the \(j\)th accent command,
- \(\gamma\) : relative ceiling level of the accent commands (generally set to \(\gamma = 0.9\)).

Figure 1: The command response model for F0 contour generation of human utterances.
3. Proposed method

Since the quantitative intonation model generates the F0 contour by superposition of the output of the phrase and the accent control mechanism and the base frequency value, a direct estimation of the model parameters from a given F0 contour is not possible. The proposed solution is therefore a stepwise decomposition of the F0 contour in its different components. That means, the parameters of one mechanism are estimated and the F0 contour, generated with this parameters, is subtracted from the input contour. The parameters of the next mechanism are then estimated from the residual signal. The parameters estimated with this method constitute a first-order approximation of the actual model parameters. This approximation is improved in an analysis-by-synthesis (A-b-S) procedure, i.e. a F0 contour is generated with the estimated parameters and compared to the measured contour. In an optimization procedure the parameters will then be altered to minimize the error between measured and model generated contour.

The least difficult is the estimation of $Fb$, which is assigned to the lowest value of the F0 contour and subtracted from the contour in logarithm domain. In the residual contour $F0_{res1}(t)$, accents and phrases have to be detected. The relatively fast rise and fall patterns of the accents form the most significant features now. Intuitively, one would like to subtract these bumps from the contour and detect the phrases in the residual. In searching for an analysis method, appropriate for the detection of the accents and phrases, several characteristics have to be taken into account:

- An accent or phrase has an impulse like shape and does not make any oscillation around the amplitude axis.
- The frequency of each of this impulses is very low (<5 Hz) and the frequency bands of the accents and phrases are very close together.
- Each accent and each phrase is an unique event, i.e. there are neither stationary nor quasi-stationary signal components.

The most common method for analysis of non-stationary signals in speech processing is the Short-Time Fourier Transform (STFT). It produces good results for the analysis of quasi-stationary signals like voiced phonemes, however it is unsuitable for the given problem. This can be explained by the fact that the STFT constitutes its family of base functions by varying the oscillation frequency of its prototype, the complex sine function. Since accents and phrases do not make any oscillation around the amplitude axis this method is inadequate.

As explained above, accents are easier to detect than phrases due to their rise and fall characteristics and their comparatively short duration. The main differences between accents consist in their duration and their height. Hence, it seems to be a better approach to analyze the signal with a transformation that has a prototype function similar to the shape of the accents. The family of base functions should then be constituted by stretching and compressing, i.e. scaling, the prototype function. A transformation providing these features is the wavelet transform (WT).

3.1. Wavelet based analysis

The family of base functions of the WT is constituted by scaling and shifting its prototype function (mother wavelet)

$$\psi_{a,t}(\tau) = [|a|^{-1/2}] \psi(\frac{\tau - t}{a})$$

(4)

The mother wavelet must provide a localization in frequency domain, expressed by the admissibility condition

$$C_{\psi} = \int_{-\infty}^{\infty} \left| \frac{\psi(\omega)}{\omega} \right|^2 d\omega < \infty$$

(5)

This imposes that $\psi(t)$ is of zero mean

$$\int_{-\infty}^{\infty} \psi(t) dt = 0$$

(6)

Therefore, the mother wavelet has to perform at least one oscillation around the zero axis. With this prototype function the CWT is given by

$$X(a, t) = (x, \psi_{a,t}) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi^*\left(\frac{\tau - t}{a}\right) d\tau$$

(7)

For computation on a computer its discrete version is used:

$$X_n(a, n \cdot \Delta t) = \frac{1}{\sqrt{a}} \sum_{k=-\infty}^{\infty} x(k \Delta t) \psi^* \left(\frac{k - n}{a} \Delta t\right)$$

(8)

Furthermore, there exist the faster working discrete wavelet transform (DWT), which is based on a special filter bank structure. Its two different types are the à trous and Mallat algorithm, which are both time-variant. Additionally, the Mallat algorithm realizes scaling via down-sampling between succeeding scales and therefore does not provide a scale value for every point in time. Hence, the CWT performs best in feature extraction.

The choice of the mother wavelet has an important influence on the results of the WT. As explained above, it is desired to use a function that is of a shape similar to the accents. This could be, for instance, the Gaussian function $e^{-t^2/2}$, but due to the admissibility condition this is not possible. The proposed solution is to use its second derivative instead (which is normalized so that its $L^2$-norm is 1)

$$\psi(t) = \frac{2}{\sqrt{3}} \pi^{-1/4} (1 - t^2) e^{-t^2/2}$$

(9)

Figure 2: A single accent (first), the Mexican hat wavelet (second), a single scale of wavelet coefficients (third) and a scalogram with coefficients of different scales (fourth).
Because of its shape (see figure 2), this wavelet is called Mexican hat [5]. It is widely used in wavelet analysis.

Under certain conditions, the coefficients of the wavelet transform may be interpreted as correlation coefficients. As an explanation of the proposed method we may therefore imagine that with the given signal and wavelet we calculate the correlation between a function and its second derivative. Since the maxima and minima of the second derivative represent the points of inflection of the original function, we assume that the resulting function of this correlation also shows maxima and minima at the points of inflection of the analyzed function.

The points of inflection of an accent contour are its starting point, which roughly corresponds to \( T_{ij} \), its maximum, which may roughly correspond to \( T_{ij} \), and the point where its influence on the F0 contour vanishes (\( T_{ij} \)).

The proposed method for estimation of accent parameters is therefore to search the coefficients of the wavelet transform with a Mexican hat for local maxima above zero and local minima below zero. The advantage of this method over a direct application of curve sketching is that due to the given size of the wavelet not every uneveness of the F0 contour, e.g. caused by special articulation phenomena of particular phonemes, is taken into account. Figure 2 displays a single accent, the Mexican hat wavelet and the coefficients of their transformation. Once the accents are detected and subtracted from the F0 contour \( F_{0,\text{res,1}}(t) \), the phrases are detected in the residual signal with the same method.

4. Proposed algorithm

With the method proposed in section 3 an algorithm for automatic extraction of the parameters of the quantitative intonation model was developed. Its main stages are presented in figure 3, and in figure 4 an example of a F0 contour at the different stages is given.

Input signal of the algorithm is a preprocessed F0 contour. The preprocessing algorithm that is described in [6] removes gross errors and microprosodic variations caused by production of several phonemes from the extracted F0 contour. Unvoiced speech segments and pauses are interpolated. Furthermore the F0 contour is stylized by piecewise polynomial approximation. The result is a smoothed steady curve.

The lowest value greater than zero is assigned first-order approximation of \( Fb \) and subtracted from the F0 contour. With the residual signal \( F_{0,\text{res,1}}(t) \) a WT is performed. Thereby the coefficients of only one scale are calculated. With the proposed Mexican hat wavelet a scaling parameter \( \alpha = 0.09 \) is used. Good results are also obtained using the real part of the Morlet Wavelet [5] with \( \omega_0 = \omega_0 = 2 \). Its shape is very similar to the Mexican hat, but in this case a scaling parameter \( \alpha = 0.13 \) is used. During the transformation the F0 contour is also mirrored at the beginning and end of the contour to avoid boundary effects. In the calculated scale, maxima greater than a very small threshold above zero and local minima lower than a very small threshold below zero are marked. The thresholds are introduced due to rounding errors caused by computation. From left to right the first marked maximum is searched and assigned the maximum of a detected accent. The preceding marked minimum is assigned first-order approximation of \( T_{ij} \) and the succeeding marked minimum is assigned \( T_{ij} \). The contour of the isolated accent is formed with all F0 values within these boundaries subtracted from a line between \( T_{ij} \) and \( T_{ij} \). The first-order approximation of the parameters \( \alpha_{ij}, \beta_{ij} \) and \( T_{ij} \) is obtained in a pattern comparison, i.e. within specified boundaries \( \alpha_{ij}, \beta_{ij} \) and \( T_{ij} \) are successively incremented with a preset step size. F0 contours generated with these parameters are compared to the input accent contour. The parameter set with the lowest least square error between generated contour and input contour is taken as first-order approximation of the parameters \( \alpha_{ij}, \beta_{ij} \) and \( T_{ij} \). The accent detection procedure continues by searching the next marked maximum after \( T_{ij} \).

After the detection of all accents, the parameters of the obtained first-order approximation are optimized in an A-b-S procedure, which is controlled by an evolution strategy as explained in [4]. With the optimized parameters of all accent commands a F0 contour is generated and subtracted from the \( F_{0,\text{res,1}}(t) \) contour. The residual contour \( F_{0,\text{res,2}}(t) \) is smoothed to minimize the influence of possible errors in the accent detection.

Again a WT is calculated with the Mexican hat wavelet. The scaling parameter is set to \( \alpha = 0.5 \). As explained above, all local extrema of the calculated scale are marked. Each marked maxima is assigned to a phrase. The point in time 200ms before a maximum at the beginning of the F0 contour is assigned first-order approximation of \( T_{ij} \). For the other phrases \( T_{ij} \) is searched backwards starting from its marked maximum. The search is stopped when a minimum or a maximum is reached. The point in time with the lowest F0 value within this search interval is assigned first-order approximation of \( T_{ij} \). With this procedure the (theoretical) infinite length of a phrase command and the resulting left to right dependency of phrases is taken into account. From the obtained residual signal the reference segment for the pattern comparison and pa-
parameter optimization of the succeeding phrase is taken. The algorithm continues this way until the parameters of the last phrase are estimated.

The obtained parameters are further refined by means of the evolution strategy stepping phrase by phrase from left to right. But at this time alteration of the parameters of preceding phrases is also allowed and hence no subtraction of preceding phrase contours is performed.

At the last stage, the parameters of all phrase and accent commands are optimized together with the evolution strategy. The procedure steps from left to right as explained for the last optimization of the phrase command parameters.

5. Experimental results

For the evaluation of the developed algorithm, a German speech corpus compiled by the Institute of Natural Language Processing at the University of Stuttgart [7] was used. The corpus consists of 48 minutes broadcast news recorded from the radio station Deutschlandfunk. According to the topics, the recordings are split up into 73 paragraphs. The paragraphs are used as input segments for the extraction of the parameters of the quantitative intonation model with the proposed algorithm. The model parameters of this data were also extracted with the algorithms described in [2] and [4].

The performance of each algorithm is measured by the RMSE between the input F0 contour and the model generated F0 contour. The algorithm described in [2] achieves an average $RMSE = 4.74 \text{Hz}$ and the algorithm of [4] achieves an average $RMSE = 7.21 \text{Hz}$. With the algorithm proposed in this contribution, an average $RMSE = 2.84 \text{Hz}$ is achieved.

6. Conclusions

In this contribution improvements for the automatic extraction of the parameters of the quantitative intonation model are proposed. The improvements are mainly based on the application of the CWT for the detection of accents and phrases in a F0 contour. This replaces the DWT based on the Mallat algorithm used in [4]. The advantage is a higher accuracy in the parameter extraction. Results show that with the proposed method a significant improvement is achieved. This is attained with a larger amount of computation, but since the algorithm works off-line, computation time is not crucial.

The rules used in the implemented algorithm are relatively simple. Therefore, the algorithm may easily be re-implemented by others. Further research may test the application of the algorithm for languages with positive and negative local commands like Chinese and Swedish. Analysis of the relationship between extracted model parameters and underlying linguistic information is another future task.

7. References