Estimation of Pitch Targets from Speech Signals by Joint Regularized Optimization

Peter Birkholz, Patrick Schmager
Institute of Acoustics and Speech Communication
TU Dresden, Germany
Email: peter.birkholz@tu-dresden.de

Yi Xu
Department of Speech, Hearing and Phonetic Sciences
University College London, UK

***Abstract*—This paper presents a novel method to estimate the pitch target parameters of the target approximation model (TAM). The TAM allows the compact representation of natural pitch contours on a solid theoretical basis and can be used as an intonation model for text-to-speech synthesis. In contrast to previous approaches, the method proposed here estimates the parameters of all targets jointly, uses 5th-order (instead of 3rd-order) linear systems to model the target approximation process, and uses regularization to avoid unnatural pitch targets. The effect of these features on the modeling error and the target parameter distributions are shown. The proposed method has been made available as the open-source software tool TargetOptimizer.**

I. Introduction

The perceived naturalness of synthetic speech depends a great deal on the generated intonation, i.e., the pitch contour [1]. Most systems for text-to-speech synthesis use some sort of intonation model that encodes the sequence of samples of the actual pitch contour in terms of a reduced set of model parameters. Prominent examples for intonation models are the Fujisaki model [2], the tilt intonation model [3], or the target approximation model [4], [5]. For speech synthesis, the parameters of the used model are first predicted from the intonational form of the utterance (e.g., pitch accent types and positions) by means of a machine-learning technique, and the final pitch contour is then deterministically calculated (decoded) from these model parameters. In order to train the prediction algorithm, the pitch contours of natural utterances need to be encoded in terms of intonation model parameters, i.e., the model parameters need to be estimated from real pitch contours.

In this study, we investigated the estimation of the pitch targets of the target approximation model (TAM) [4], [5]. The TAM has been previously shown to be well suited for encoding intonation for multiple communicative functions in parallel [6], and has the potential to generate highly natural intonation for text-to-speech synthesis on a solid theoretical basis. The basic principle of the TAM is shown in Fig. 1 for the German word "versuchen". Here, the f_0 contour measured for the spoken utterance is shown by the connected gray dots, and the f_0 contour re-synthesized by the TAM is given by the black dots on the smooth black curve. The TAM assumes that the surface f_0 contour results from the sequential approximation of pitch targets, which are shown as (oblique) dashed lines

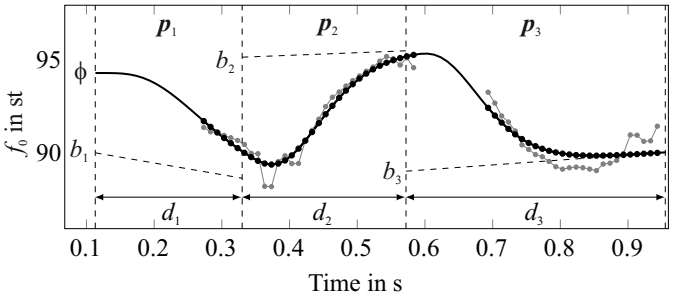

Fig. 1. Illustration of the target approximation model for the German word "versuchen" [fɐ.zˈuː.xn] with three syllables. The f_0 contour of the natural utterance is shown as gray dots and the modeled f_0 is shown as black dots. Syllable boundaries are indicated as vertical dashed lines, and pitch targets as oblique dashed lines.

in Fig. 1. These targets represent the goals of the (log) f_0 movement in terms of linear time functions. It is assumed that there is one target per syllable. The target approximation is modeled by a 3rd-order critically damped linear system. Accordingly, the f_0 contour can be considered as the output of a low-pass filter applied to the sequence of syllable-wise linear target functions.

To determine the target parameters that best reproduce the f_0 contour of a given natural utterance, Prom-On et al. developed the tool PENTAtrainer1 [5]. This tool first interpolates the f_0 contour of the natural utterance in the voiceless parts, and then performs an exhaustive search of the target parameters over a grid of integral numbers to find the parameter values that minimize the root mean square error (RMSE) between the modeled and (interpolated) natural f_0 contours. The target parameters are optimized sequentially for one syllable after the other. An alternative method to estimate pitch target parameters was implemented in the tool PENTAtrainer2 [6]. However, instead of estimating the TAM parameters that best reproduce the f_0 of *one specific utterance*, this tool estimates the best TAM parameters for specific combinations of "communicative functions" (according to the Parallel Encoding and Target Approximation framework [7]) that have been assigned to the syllables of a whole corpus of speech, using the optimization method of simulated annealing.

In the present study we propose an alternative to the previous methods and show how it improves the modeling of

ISBN 978-90-827970-1-5

the f_0 of natural utterances. The main features of the proposed method are the following:

- The f_0 contour in the voiceless parts of the original utterance is *not* interpolated before the target estimation, because it is not guaranteed that the interpolated sections support the optimal reproduction of the f_0 in the truly voiced sections.
- The targets for all syllables of the utterance are *jointly* optimized, instead of for one syllable after the other.
- We argue that different combinations of TAM parameters can generate almost identical f_0 contours. Therefore we introduce regularization in order to prefer natural or plausible pitch targets.
- The TAM uses 5th-order systems instead of 3rd-order systems, as they allow a more accurate modeling of pitch contours.

In addition, we tested whether shifting the syllables boundaries compared to the conventional syllable segmentation would improve f_0 modeling, as hypothesized by Xu and Liu [8] (which was not the case).

II. Method

A. Target approximation model

The TAM assumes one pitch target for each syllable of an utterance. Within the interval of a syllable, the target $x(t)$ is defined as the linear function

$$x(t) = mt + b, \tag{1}$$

where m (in st/s) and b (in st) denote the the slope and height of the target, respectively. The time t is defined relative to the onset of the syllable for the interval $[0, d]$, where d is the syllable duration. The f_0 (in st) within the syllable is the response of a critically-damped low-pass filter of the order N (i.e., the concatenation of N identical first-order low-pass filters) with the time constant τ, i.e.,

$$f_0(t) = (c_0 + c_1 t + \ldots + c_{N-1} t^{N-1}) e^{-t/\tau} + (mt + b), \tag{2}$$

where the constants $c_0 \ldots c_{N-1}$ depend on the initial conditions. These constants are calculated such that f_0 and its derivatives $f_0^{(n)} = d^n f_0(t)/dt^n$ at the beginning of a new syllable equal the f_0 and its derivatives at the end of the previous syllable, so that the dynamic state of the system is transferred from one syllable to the next:

$$\begin{aligned} c_0 &= f_0(0) - b \qquad (3)\\ c_n &= (f_0^{(n)}(0) - \frac{d^n}{dt^n}(mt+b) - \\ &\quad \sum_{i=0}^{n-1} c_i (-1/\tau)^{n-i} \binom{n}{i} i!)/n!, \quad n = 1 \ldots N-1 \end{aligned}$$

The derivatives of f_0 at the end of a syllable are given by

$$\begin{aligned} f_0^{(n)}(t) &= \frac{d^n}{dt^n}(mt+b) + e^{-t/\tau} \sum_{i=0}^{N-1} t^i \cdot \qquad (4)\\ &\quad \left(\sum_{j=0}^{\min\{N-1-i,n\}} (-1/\tau)^{(n-j)} \binom{n}{j} c_{i+j} \frac{(i+j)!}{i!} \right) \end{aligned}$$

for $t = d$. At the beginning of the first syllable of an utterance, f_0 is given as an onset value ϕ and the derivatives of f_0 are set to zero. As a formal difference to the original quantitative TAM [5], we used the time constant τ to characterize the linear system instead of the parameter $\lambda = 1/\tau$ to avoid confusion with the regularization parameter λ (see below).

B. Estimation of pitch targets

For the estimation of the pitch targets, we assume that the pitch contour to be reproduced by the TAM is given in terms of samples $f_0(k\Delta t)$ in the voiced parts of the corresponding utterance, where Δt is the sampling interval (typically 10 ms) and k is the sampling index. We also assume that the syllable boundaries are given, and hence the target durations. The unknown parameters of the TAM are the offset b_s, the slope m_s, and time constant τ_s of every syllable (i.e., target) s. These parameters can be summarized in vectors $\mathbf{p_s} = (m_s, b_s, \tau_s)^T$, where $s = 1, 2, \ldots, S$ and S is the number of syllables. The initial f_0 value ϕ of the TAM is set to the first f_0 sample of the utterance.

In contrast to the previous approaches implemented in PENTAtrainer1 and 2, we propose the *joint* estimation of all TAM parameters by *regularized* optimization. Compared to the syllable-wise successive estimation of target parameters in PENTAtrainer1, the joint estimation has a better chance to find an optimal solution for the whole utterance. In addition, regularization helps to obtain solutions that are not only optimal in a mathematical sense but also physiologically most plausible. The proposed objective function to be minimized is

$$\begin{aligned} g(\mathbf{p}_1 \ldots \mathbf{p}_S) &= ||f_0(k\Delta t) - \hat{f}_0(k\Delta t, \mathbf{p}_1 \ldots \mathbf{p}_S)||_2^2 \qquad (5)\\ &+ \lambda \sum_{s=1}^{S} (\mathbf{p}_s - \overline{\mathbf{p}})^T W (\mathbf{p}_s - \overline{\mathbf{p}}) \end{aligned}$$

subjected to the linear constraints (acting as search bounds)

$$\begin{pmatrix} -50 \text{ st/s} \\ 75 \text{ st} \\ 12.5 \text{ ms} \end{pmatrix} \le \begin{pmatrix} m_s \\ b_s \\ \tau_s \end{pmatrix} \le \begin{pmatrix} 50 \text{ st/s} \\ 115 \text{ st} \\ 1 \text{ s} \end{pmatrix} \quad s = 1 \ldots S. \tag{6}$$

The first term on the right-hand side of Eq. (5) is the squared Euclidian distance between the original f_0 samples in the voiced parts of the utterance and the corresponding values $\hat{f}_0$ generated by the TAM. The second term is the regularization term that penalizes parameter values based on their deviation from the *preferred* values $\overline{\mathbf{p}} = (\overline{m}, \overline{b}, \overline{\tau})^T$. The degrees of penalization for the different TAM parameters are adjusted by the elements of the weight matrix $W = \text{diag}(w_m, w_b, w_\tau)$, and λ determines the overall degree of regularization. For the present study, the weights were empirically adjusted to $w_m = 1$ s^2/st^2, $w_b = 0.6$ st^{-2}, and $w_\tau = 0.2$ s^2. We furthermore set $\overline{m} = 0$ (static targets are preferred) and $\overline{\tau} = 12.5$ ms (this is considered the "typical" time constant). $\overline{b}$ was set to the average pitch of the corresponding natural utterance. With regard to the constraints (6), the bounds for τ

ISBN 978-90-827970-1-5

were adopted from [5], and the bounds for b were adjusted to cover a frequency range of 76-767 Hz ($f_0[Hz] = 2^{f_0[st]/12}$).

In principle, the optimization problem above can be solved using any method for bound-constrained optimization that requires only the availability of an objective function but no derivative information (see [9] for a recent review of this class of methods). Here we used the algorithm BOBYQA (Bounded Optimization by Quadratic Approximation) [10], which is available as a C++ implementation in the modern open-source software toolkit dlib [11], and generally performs better than the widely used Nelder-Mead simplex algorithm [12], according to [9]. Like most optimization methods for non-convex problems, BOBYQA cannot guarantee to find the global minimum of the objective function. Accordingly, the method should be run multiple times with different random initial parameter values (within the respective bounds), and the best solution should be selected. Here we used $5S + 10$ random initializations per optimization, as the complexity of the problem increases with the number of syllables S.

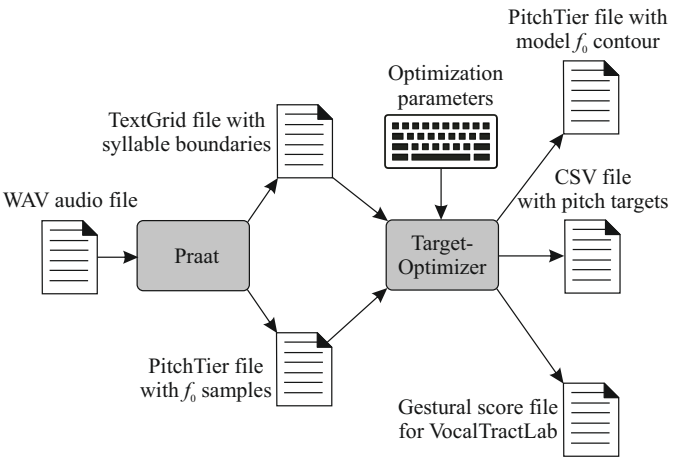

Fig. 2. Information flow diagram for the TargetOptimizer.

The target optimization described above has been implemented as the GUI-based open-source C++ software "Target-Optimizer", which is available at http://www.vocaltractlab.de/index.php?page=targetoptimizer-download. Fig. 2 illustrates the information flow for the tool. The input data for the TargetOptimizer are a TextGrid file with syllable boundaries and a PitchTier file with the f_0 samples to be reproduced by the TAM. Both files can be created with the software Praat [13] from the audio file of the original utterance. The optimization parameters (i.e., the regularization parameters λ and W, and the bounds of the search space) can be set in the GUI or as command line parameters. The results of the optimization can be saved as a PitchTier file with the f_0 samples of the synthesized pitch contour, as a table with pitch target parameters, or as a gestural score for the articulatory speech synthesizer VocalTractLab 2.2 ([14], www.vocaltractlab.de).

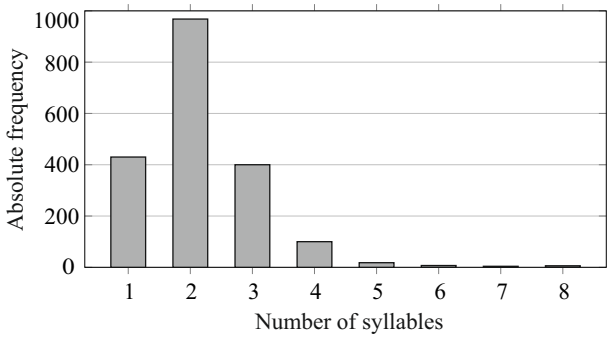

Fig. 3. Histogram of syllables in the words of the evaluation corpus.

III. Evaluation

A. Speech corpus

For the evaluation of the estimation method we used a corpus of German words spoken by a professional female speaker. The words are audio samples that supplement a German pronunciation dictionary [15] and are available as WAV files from https://www.degruyter.com/view/product/19839. In total, we used 1934 words with 4175 syllables, with the syllable histogram given in Fig. 3. The reason for using individual words instead of longer utterances was that this study was embedded in a project aiming to predict the standard intonation for single German words. For each spoken word, the syllable boundaries were manually segmented according to conventional acoustic landmarks using the software Praat and saved as TextGrid files. In addition, the f_0 contour was extracted for each word and saved as a PitchTier file. Each automatically determined pitch contour was carefully checked for inaccurate pitch samples and, where necessary, manually corrected.

B. Comparison with PENTAtrainer 1

To evaluate the effect of the *joint* estimation of the pitch targets as opposed to the previously proposed sequential estimation, we compared the performance of the proposed method (here without regularization, i.e., $\lambda = 0$) with that of PENTAtrainer1 for all words of our corpus. The performance of both methods was quantified using (a) the RMSE and (b) Pearson correlation coefficient ρ between the pitch samples of the natural utterances and the corresponding modelled pitch values. With the proposed method, we got $\mathrm{RMSE} = 0.557$ st and $\rho = 0.946$, and with PENTAtrainer1 we got $\mathrm{RMSE} = 1.028$ st and $\rho = 0.883$. Hence, the joint estimation is a clear improvement over the sequential estimation.

C. Effect of model order

The previous methods PENTAtrainer1 and PENTAtrainer2 are based on a 3rd-order TAM, i.e., $N = 3$ in Eq. (2). Here we examined whether a different system order is possibly more suitable to reproduce the natural pitch contours in our corpus. To this end, the pitch contour of each word was estimated for all $N \in \{2, 3, \ldots, 10\}$, using no regularization again ($\lambda = 0$). The average RMSE as a function of the model order (Fig. 4)

ISBN 978-90-827970-1-5 © EURASIP 2018

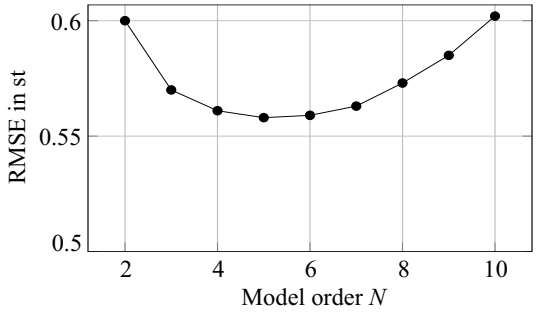

Fig. 4. Root mean square error between the measured and modeled f_0 samples for different orders of the target approximation model.

was lowest for $N = 5$ instead of $N = 3$. This happens to conform with the optimal model order for the reproduction of articulatory trajectories for lip and jaw movements that was found previously [16]. Hence, a 5th-order TAM is superior to a 3rd-order model for both pitch and supraglottal articulatory movements.

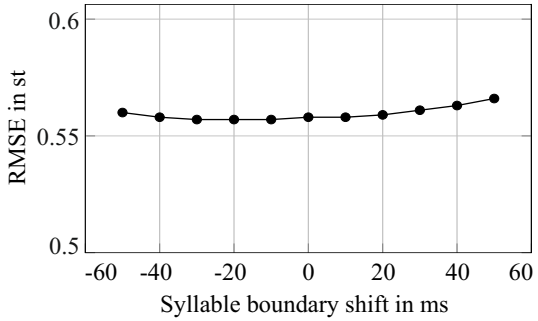

Fig. 5. Root mean square error between the measured and modeled f_0 samples for different shift values for the syllable boundaries.

D. Effect of syllable boundary shift

As hypothesized by Xu and Liu [8], it could be beneficial to move the syllable boundaries (and hence the temporal domains of the targets) obtained by the conventional segmentation rules towards the left by about 20-30 ms. Here we systematically shifted the syllable boundaries (all boundaries simultaneously) from their conventional positions by -50 ms to +50 ms (in steps of 10 ms), and tested the model performance for all words in the corpus (for $\lambda = 0$ and $N = 5$). Although the optimal target parameters of individual syllables varied quite strongly depending of the temporal shift, there was hardly any overall effect on the RMSE, as shown in Fig. 5. Hence, the conventional way of segmenting syllables is suitable for the optimal reproduction of pitch contours.

E. Effect of regularization

The visual inspection of the estimated pitch targets using the optimization without regularization, i.e., $\lambda = 0$, revealed that the estimated targets were often not plausible in the sense of the TAM. As an example, Fig. 6 (top) shows the pitch contour of a three-syllabic word and the corresponding

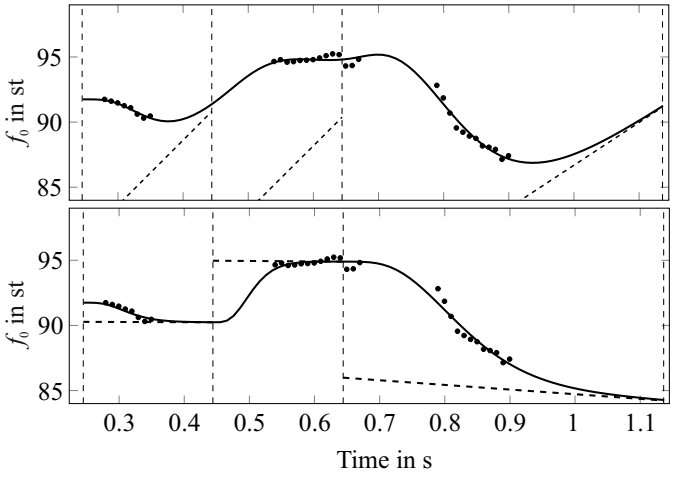

Fig. 6. Comparison of measured f_0 samples (dots) and modelled f_0 contours (curves) for the word "Ästhetik" [ɛs.t'eː.tik] without regularization ($\lambda = 0$ and RMSE = 0.303 st, top) and with regularization ($\lambda = 0.75$ and RMSE = 0.348, bottom).

estimated targets for $\lambda = 0$. Obviously, the targets exhibit strong slopes although the pitch contour is rather constant in each of the first two syllables. Hence, an optimal reproduction of the original pitch contour has been achieved at the expense of rather unnatural targets. This effect is explained by the distributions of pitch target parameters obtained by un-regularized optimization, as shown in Fig. 7 (top). Here, parameter values at the bounds of the search space are often preferred. This effect is counteracted by the regularization term in Eq. (5), which penalizes extreme values and prefers slope values around zero, offset values around the mean pitch of the utterance, and time constants around 12.5 ms. Fig. 7 (bottom) shows the distributions for a regularization parameter of $\lambda = 0.015$, where slope and offset values are now almost normally distributed. Fig. 6 (bottom) illustrates that regularization yields pitch targets that are far more plausible than without regularization, while the RMSE gets only marginally worse (from $\mathrm{RMSE} = 0.56$ st with $\lambda = 0$ to $\mathrm{RMSE} = 0.62$ st with $\lambda = 0.015$ for the whole corpus).

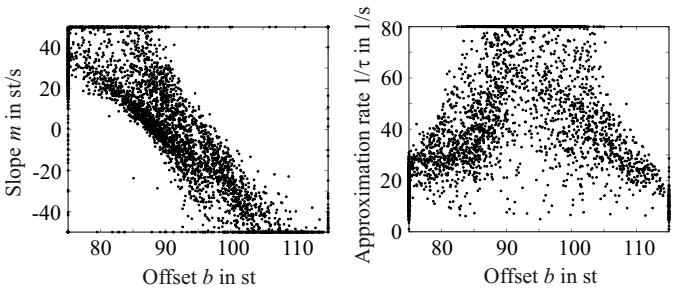

Fig. 8. Scatter plots of the estimated parameters of the target approximation model for $\lambda = 0$ (no regularization).

The observations above indicate that the TAM is possibly overdetermined from the viewpoint of optimization, i.e., different pitch target parameter combinations can lead to (nearly)

ISBN 978-90-827970-1-5 © EURASIP 2018

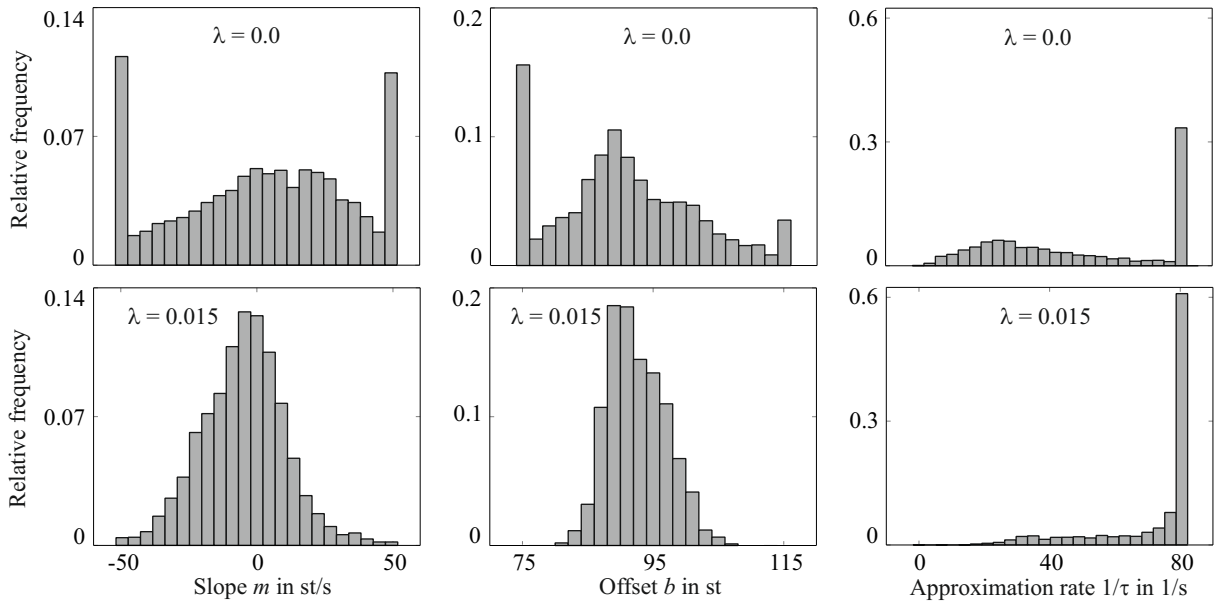

Fig. 7. Influence of the regularization parameter λ on the distributions of the estimated parameters of the target approximation model.

identical pitch contours. This assumption is also supported by the scatter plots in Fig. 8, which show that target parameter values obtained without regularization are strongly correlated. Hence, regularization is proposed as an essential feature to estimate natural and plausible TAM parameters.

IV. Conclusion

We have demonstrated that an estimation of pitch targets based on joined regularized optimization using a 5th-order TAM allows not only the re-synthesis of natural f_0 contours with a smaller error than the previous estimation method of PENTAtrainer1 but also yields more "natural" pitch targets in the sense of the model. This makes the TAM a highly interesting intonation model for future text-to-speech synthesis systems. In this study, the proposed method has only been tested with German utterances. In future work it would be interesting to apply the method to tone languages like Mandarin Chinese. In this case, the estimated target parameters should reflect the type of tone associated with the individual syllables, e.g., the method should ideally yield targets with positive slopes for raising tones, and with negative slopes for falling tones. Furthermore, future work is needed to determine the relation between targets and syllables with multiple morae, and the robustness of the TAM to ambiguities of syllabic segmentation in running speech. For example, for Japanese, Lee [17] found evidence that two consecutive morae may carry a single pitch target, which is synchronized with the syllable.

References

[1] P. Taylor, *Text-to-speech synthesis*. Cambridge University Press, 2009.

[2] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233–241, 1984.

[3] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[4] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Communication*, vol. 33, pp. 319–337, 2001.

[5] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 405–424, 2009.

[6] Y. Xu and S. Prom-On, "Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning," *Speech Communication*, vol. 57, pp. 181–208, 2014.

[7] Y. Xu, "Speech melody as articulatory implemented communicative functions," *Speech Communication*, vol. 46, pp. 220–251, 2005.

[8] Y. Xu and F. Liu, "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," *Italian Journal of Linguistics*, vol. 18, pp. 125–159, 2006.

[9] L. M. Rios and N. V. Sahinidis, "Derivative-free optimization: a review of algorithms and comparison of software implementations," *Journal of Global Optimization*, vol. 56, no. 3, pp. 1247–1293, 2013.

[10] M. J. Powell, "The BOBYQA algorithm for bound constrained optimization without derivatives," *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 2009.

[11] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[12] J. A. Nelder and R. A. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308–313, 1965.

[13] P. Boersma and D. Weenik, "Praat: doing phonetics by computer [software]," 2014. [Online]. Available: http://www.praat.org/

[14] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLoS ONE*, vol. 8, no. 4, p. e60603, 2013.

[15] E.-M. Krech, E. Stock, U. Hirschfeld, and L.-C. Anders, *Deutsches Aussprachewörterbuch*. Walter de Gruyter, 2009.

[16] P. Birkholz and P. Hoole, "Intrinsic velocity differences of lip and jaw movements: preliminary results," in *Interspeech 2012*, Portland, Oregon, USA, 2012, pp. 2017–2020.

[17] A. Lee, "The dynamics of Japanese prosody," PhD Dissertation, University College London, 2015.

ISBN 978-90-827970-1-5 © EURASIP 2018