ACCOUNTING FOR MICROPROSODY IN MODELING INTONATION

Peter Birkholz, Xinyu Zhang

Institute of Acoustics and Speech Communication, TU Dresden, Germany

ABSTRACT

Intonation models are often used for the generation of fundamental frequency (f_0) contours in speech synthesis. Current intonation models only represent the intentional f_0 components that are related to the phonological structure of the utterance. However, natural speech also contains non-intentional microvariations of f_0, which are usually not accounted for. Here, we derived models for two forms of microvariations: the drop in f_0 during voiced obstruents, and the increased f_0 at the onset of vowels following voiceless obstruents. These models were applied to remove the microvariations of f_0 in a database of natural speech before the f_0 contours were reproduced with the Target Approximation Model. The previously removed microvariations were then superimposed on the modeled f_0 contours. The resulting model f_0 contours were significantly more similar to the original (natural) f_0 contours than model contours that did not account for the microvariations. This approach might improve f_0 modeling in future parametric speech synthesizers.

Index Terms— pitch modeling, intrinsic f0 variation, cointrinsic f0 variation, Target Approximation Model

1. INTRODUCTION

Most text-to-speech synthesizers use some kind of intonation model that represents the f_0 contour by a compact set of parameters, which can be predicted from the phonological structure of the intended utterance. Popular intonation models are the Fujisaki model [1], the tilt intonation model [2], and the Target Approximation Model [3, 4]. These models usually represent the f_0 contour at the phrase or syllable levels. For example, the Fujisaki model represents the f_0 as a superposition of phrase and accent components [1], and in the Target Approximation Model the f_0 contour is generated by the successive approximation of linear pitch targets with one target per syllable [4]. The model contours can be interpreted to represent the (linguistically) intended intonation.

However, real speech also contains small systematic pitch variations on the order of 1-2 semitones at the level of individual speech sounds (also called microvariations, or microprosody) that are *not* actively controlled by the speaker and

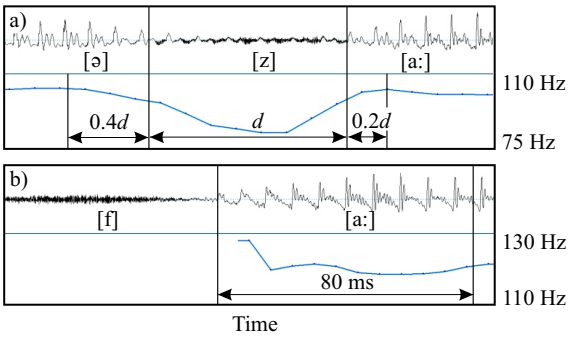

Fig. 1. a) Audio and f_0 waveforms of the section [əzaː] from sentence #3 from the database showing the IF0 effect of the [z]. b) Audio and f_0 waveforms of the section [faː] from the same sentence, showing the CF0 effect of the [f] on the [aː].

universal across languages [5]. Studies often distinguish *intrinsic* effects of phones on f_0 (IF0), and *co-intrinsic* effects (CF0) [6]. IF0 effects are for example observed for vowels, with high vowels tending to have higher f_0 values than low vowels [5]. CF0 effects occur when an obstruent is followed by a vowel. For voiceless obstruents, the f_0 at the beginning of the following vowel tends to be higher than in the middle of the vowel (Fig. 1b), and for voiced obstruents, the f_0 at the beginning of the following vowel tends to be lower [6]. The latter is caused by an intrinsic drop of the f_0 during the voiced obstruent (also an IF0 effect, see Fig. 1a), which is partly carried over into the following vowel [7, 8]. The mechanisms behind these effects are still not fully understood, but different explanations can be found, e.g., in [9, 8, 10].

With regard to speech synthesis, considering microprosody can potentially make the synthesis more intelligible, because it helps to perceptually discriminate voiced from unvoiced obstruents [11, 12, 13]. This would especially benefit synthesis methods that model the f_0 contour explicitly, e.g., articulatory synthesis [14] or HMM-based synthesis [15]. Whether modeling microprosody can improve the *naturalness* of synthetic speech has often been conjectured (e.g. [10, 16]), but has not been conclusively demonstrated. The only study that actually tested this found that the presence or absence of the f_0 drop during voiced obstruents has *no* significant effect on the naturalness [17]. However, the study was based on a limited set of CVC stimuli and analyzed just

This research was partly funded by the German Federal Ministry for Economic Affairs and Energy (BMWi/ZIM), grant number ZF4443004BZ8

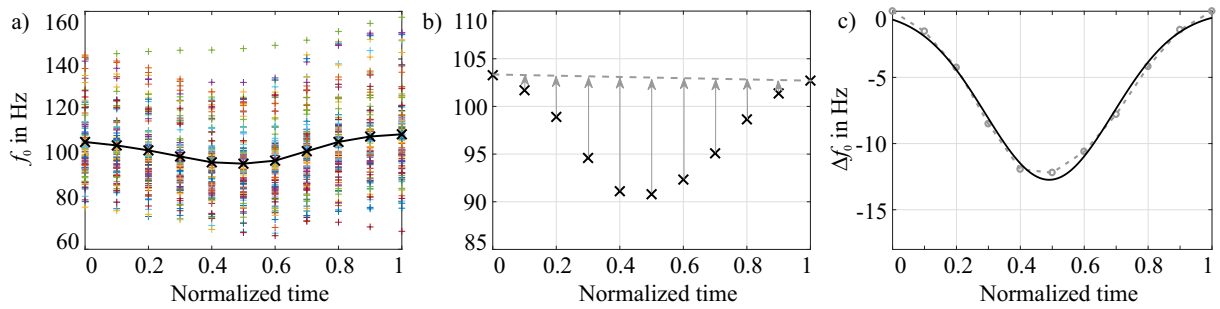

Fig. 2. Steps towards a model of intrinsic f_0 effects of voiced obstruents.

one type of microprosodic effect. Hence, a beneficial effect may become evident on the basis of a more extensive study.

The present study makes the following contributions.

1. We derived prototypical time functions for the microprosodic f_0 deflections from the intentional f_0 contour and analyzed how they are affected by phone identity and duration. While most previous studies on microprosody used "stereotyped speech" or lab speech, e.g., a set of carefully selected target words in a carrier sentence [18], we analyzed a big corpus of read speech, where the phones of interest occur in all kinds of different contexts.

2. We explored whether the f_0 contour of natural speech can be represented more accurately by the Target Approximation Model [4] (using automatic optimization) when microprosody is removed using the prototypical functions derived in 1). Put differently, to what extent does the reproduction of natural f_0 contours by the intonation model improve when it accounts for microprosodic effects [19]? This is important because the reproduction of natural f_0 contours by an intonation model is the foundation to train machine learning methods to predict the model parameters for speech synthesis.

2. CORPUS

Our analysis and modeling was based on the "BITS unit selection corpus" [20]. It contains audio recordings of 1683 German sentences, which were each read fluently with normal intonation by four speakers. The sentences were selected to cover every possible German diphone combination in as many contexts as possible. All sentences were segmented and transcribed at the phone level. Syllable boundaries, which were not contained in the corpus, were manually added for this study (they were needed for the Target Approximation Model in Sec. 4). The present study used the 1683 sentences of speaker 3, a 40 year old male radio announcer. For each sentence, the f_0 contour was determined using Praat [21] with an f_0 search range from 50 to 200 Hz.

3. PROTOTYPICAL FUNCTIONS FOR IF0 AND CF0

3.1. Modeling intrinsic f_0 effects

To model the intrinsic f_0 drop during voiced obstruents, we selected all instances of /b, d, g, v, z, ʒ, j, ʁ/ that were embedded between two vowels (to exclude f_0 effects in consonant clusters) and had a length between 30 and 150 ms (4238 items in total). To analyze potential effects of duration on the IF0 deflection patterns, the selected instances for each voiced obstruent were further partitioned into four duration groups. The duration limits for the four groups were selected such that each group contained about the same number of items. For example, the 587 instances of /z/ were split into 148 items with durations from 30-60 ms, 150 items from 60-70 ms, 143 items from 70-80 ms, and 146 items from 80-150 ms.

Within each duration group, an average IF0 deflection pattern was then determined as follows. First, a segment of the f_0 contour was extracted for each sound in the group. For voiced fricatives of duration d, the extracted f_0 segment started $0.4d$ before the beginning of the sound, and ended $0.2d$ after the end of the sound (see Fig. 1a). For voiced plosives, slightly different intervals were extracted, starting $0.5d$ before the beginning of the sound, and ending $0.2d$ after the end of the sound. These boundaries were selected such that the f_0 deflection pattern determined in the following steps was as symmetric as possible. All f_0 segments of the same group were time-normalized to the interval [0, 1] and averaged at 11 equidistant points, as shown in Fig. 2a, yielding a (negative) bell-shaped waveform. A potential tilt of the bell shape was removed by subtracting the values on a straight line connecting the first and the last sample points, as shown in Fig. 2b. Finally, the curve was fitted with a Gaussian of the form

$$\Delta f_0(t') = a \cdot \exp(-(t'-b)^2/c^2) \qquad (1)$$

in the least-squares sense, where t' is the normalized time. Fig. 2c shows the Δf_0 samples obtained with the steps above (circles connected with dashed lines) and the fitted Gaussian (black curve) for the /z/ instances in the (80 ms,150 ms] duration group.

The ranges of the parameters of Eq. (1) obtained for the examined obstruents (across the duration groups) are summa-

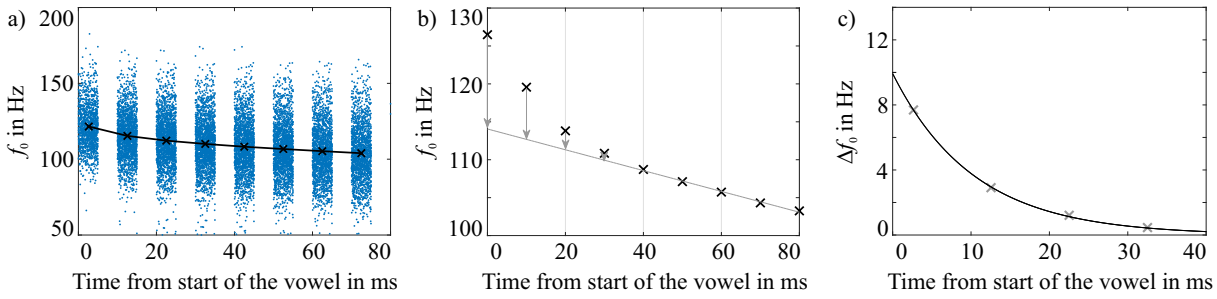

Fig. 3. Steps towards a model of co-intrinsic f_0 effects of voiceless obstruents on the following vowel.

Table 1. Estimated parameters for the IF0 function given by Eq. (1). The unit of a is Hz, while b and c are dimensionless.

	count	$[\lvert a_{\min}\rvert, \lvert a_{\max}\rvert]$	$[b_{\min}, b_{\max}]$	$[c_{\min}, c_{\max}]$
/b/	562	[10.8, 16.4]	[0.46, 0.51]	[0.21, 0.27]
/d/	1036	[7.4, 10.8]	[0.49, 0.51]	[0.21, 0.24]
/g/	573	[8.5, 10.9]	[0.45, 0.48]	[0.25, 0.26]
/v/	421	[11.9, 16.9]	[0.47, 0.52]	[0.26, 0.27]
/z/	587	[11.1, 12.7]	[0.48, 0.49]	[0.27, 0.28]
/ʒ/	50	[9.9, 13.3]	[0.49, 0.51]	[0.25, 0.28]
/j/	99	[0.9, 8.9]	[0.33, 0.58]	[0.18, 0.28]
/ʁ/	910	[9.0, 10.3]	[0.5, 0.51]	[0.26, 0.28]

rized in Table 1. For most phonemes, the parameter ranges were rather limited, indicating that they do not strongly depend on the duration (group). Furthermore, the parameter values were similar across the phonemes. Hence, it seems adequate to model IF0 deflections with the same parameters for all voiced obstruents and durations. Taking the median of the respective parameter values, we get

$$\Delta f_0(t') = -10.75 \text{ Hz} \cdot \exp(-(t' - 0.49)^2/0.26^2). \quad (2)$$

3.2. Modeling co-intrinsic f_0 effects

To model the CF0 effect of voiceless obstruents on the following vowel, we selected all vowels that followed one of the phonemes /p, t, k, f, s, ʃ, ç, pf, ts, tʃ/ from the dataset. For each of these vowels, the f_0 contour during the first 80 ms of the vowel was extracted, as illustrated in Fig. 1b. These extracted f_0 contour segments were then averaged across all vowels, separately for each preceding obstruent. As an example, Fig. 3a shows the f_0 contour points (dots) of all the vowels that followed /t/, and the mean f_0 contour (black line) in terms of eight equidistant f_0 samples. The f_0 drop caused by the preceding obstruent occurs mainly within the first 40 ms of the vowel and then approaches a constant f_0 declination after about 40 ms. To separate the initial drop caused by the obstruent from the underlying declination, the averaged f_0 samples were subtracted from a regression line fitted through the f_0 samples between 40 to 80 ms, as shown in Fig 3b. The average f_0 contours obtained in this way were then fitted with an exponential of the form

$$\Delta f_0(t) = g \cdot \exp(-ht), \quad (3)$$

separately for each preceding obstruent. The values of the parameters g and h obtained for the different preceding obstruents have been summarized in Table 2. Apart from the very high h value for /p/, the parameter values are generally similar across the preceding consonants and warrant a common time function to model the f_0 decay in the following vowel. Taking the median values for g and h, this function becomes

$$\Delta f_0(t) = 11.1 \text{ Hz} \cdot \exp(-95.9 \text{ s}^{-1} \cdot t). \quad (4)$$

Table 2. Estimated parameters for the CF0 function. The units for g and h are Hz and s^{-1}, respectively.

	count	g	h		count	g	h
/p/	561	11.5	955.0	/ʃ/	388	11.1	96.2
/t/	2154	9.9	96.4	/ç/	199	17.7	71.2
/k/	931	7.6	137.5	/pf/	63	12.0	108.3
/f/	1153	11.2	83.8	/ts/	734	10.8	93.8
/s/	563	10.7	95.7	/tʃ/	106	11.9	58.0

4. REPRODUCING F_0 CONTOURS WITH THE TARGET APPROXIMATION MODEL

4.1. Method

Based on the prototypical IF0 and CF0 deflections given by Eqs. (2) and (4), four f_0 contours were created for each of the sentences in the dataset:

1. The *original f_0 contour* was extracted from the audio files using Praat [21] without further manipulation.

2. The IF0 effects of voiced obstruents were removed from the original f_0 by subtracting Δf_0 according to Eq. (2) for all voiced obstruents embedded between two vowels.

3. The CF0 effects of voiceless obstruents were removed from the original f_0 by subtracting Δf_0 according to Eq. (4) in all vowel segments following voiceless obstruents.

4. Both IF0 and CF0 effects were removed from the original f_0 contour.

The f_0 contours in all four conditions were then fitted with the Target Approximation Model (TAM), based on the optimization method presented in [22]. According to the TAM, the f_0 contour results from the sequential approximation of pitch targets, with one pitch target per syllable [3, 4]. Besides the syllable duration, each pitch target has three parameters, namely the offset and the slope of the target, and the speed of target approximation. For each sentence, the parameters of all targets were jointly estimated by minimizing the difference between the input and the model f_0 contours using the software "TargetOptimizer" [22]. For 332 of the 1683 sentences, the estimation algorithm did not achieve the convergence criterion within the preset maximal number of iterations (mainly sentences with a very high number of syllables), and the corresponding sentences were removed from the analysis. For each of the four conditions, the microprosodic f_0 deflections that were removed before the fitting of the TAM model were then re-applied (added) to the f_0 contour synthesized from the fitted TAM to obtain the *synthesized f_0 contour*. Finally, for each condition, the root-mean-square error (RMSE) and the Pearson's correlation coefficient r between the original f_0 contours and the corresponding synthesized f_0 contours were calculated across all sentences.

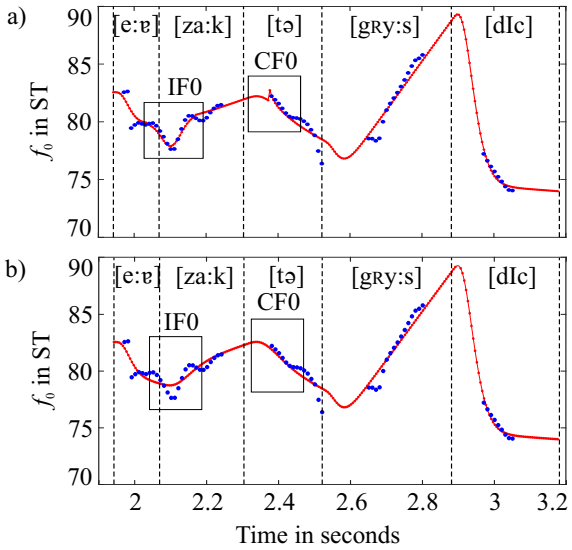

Fig. 4. a) Original f_0 contour (blue dots) for the sentence "Er sagte grüß dich" [ʔeːɐ.zaːk.tə.gʀyːs.dɪç] compared to the synthesized f_0 contour (in red) in condition 4 (consideration of both IF0 and CF0). b) Same as a) but with the synthesized f_0 contour of condition 1 (no consideration of IF0 and CF0).

Table 3. Root-mean-square errors (RMSE) in semitones (ST) and correlation coefficients (r) between the original f_0 contours and the synthesized f_0 contours for the four conditions. Significant deviations of the RMSE from condition 1 ($p < 0.005$) are indicated by *.

condition	accounting for	RMSE in ST	r
1	-	0.740	0.960
2	IF0	0.710*	0.963
3	CF0	0.732	0.961
4	IF0 & CF0	0.700*	0.964

4.2. Results

As an example, Fig. 4 shows the original f_0 contour of the sentence [ʔeːɐ.zaːk.tə.gʀyːs.dɪç] (blue dots) compared to the synthesized f_0 contour in condition 4 (fully accounting for IF0 and CF0) in a), and in condition 1 (IF0 and CF0 are not accounted for) in b) (red lines). While the intrinsic f_0 drop during [z] in the syllable [zaːk] is well reproduced in condition 4, the corresponding part of the f_0 contour is smoothed out by the TAM in condition 1. A similar effect (but less obvious) can be seen for the initial CF0 drop in [ə] after the voiceless obstruent [t] in the syllable [tə].

The RMSE and the correlation coefficients between the original and synthesized f_0 contours are reported in Table 3. Here, both considering IF0 and CF0 do improve the modeling of intonation, while the improvement due to the consideration of IF0 is stronger than the improvement due to CF0. The greatest reduction of the RMSE is achieved when both IF0 and CF0 are accounted for. The mean differences of the RMSE between conditions 1 and 2, and between conditions 1 and 4 are significant with $p = 0.0028$ and $p = 0.0001$, respectively, based on paired two-tailed t-tests.

5. DISCUSSION

The reduction of the RMSE due to the consideration of IF0 and CF0 in Table 3 seems relatively small, because the errors were calculated over the *entire* sentences and not just the signal parts that actually contain IF0 and CF0 effects. However, during the time intervals of the affected phones, the improved f_0 modeling due to microprosody can be quite obvious, as Fig. 4 shows. In summary, the present study showed that IF0 and CF0 effects can be effectively considered in an intonation model using *prototypical* time functions (with fixed parameters) that describe the microprosodic deflection of f_0 from the intended f_0 contour. However, the present study analyzed these effects for only one specific speaker and can therefore not be generalized across speakers yet. Furthermore, the perceptual impact of modeling microprosodic effects should be explored in detail in future research.

6. REFERENCES

[1] Hiroya 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.

[2] Paul 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.

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

[4] Santhitam Prom-on, Yi Xu, and Bundit 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.

[5] Douglas H Whalen and Andrea G Levitt, "The universality of intrinsic F0 of vowels," *Journal of Phonetics*, vol. 23, no. 3, pp. 349–366, 1995.

[6] A. Di Cristo and D. J. Hirst, "Modelling French micromelody: Analysis and synthesis," *Phonetica*, vol. 43, pp. 11–30, 1986.

[7] Vinay Kumar Mittal, B Yegnanarayana, and Peri Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1932–1941, 2014.

[8] James P Kirby and D Robert Ladd, "Effects of obstruent voicing on vowel F0: Evidence from "true voicing" languages," *The Journal of the Acoustical Society of America*, vol. 140, no. 4, pp. 2400–2411, 2016.

[9] Eli Fischer-Jørgensen, "Intrinsic f0 in tense and lax vowels with special reference to German," *Phonetica*, vol. 47, no. 3-4, pp. 99–140, 1990.

[10] Helen M Hanson, "Effects of obstruent consonants on fundamental frequency at vowel onset in English," *The Journal of the Acoustical Society of America*, vol. 125, no. 1, pp. 425–441, 2009.

[11] Osamu Fujimura, "Remarks on stop consonants: Synthesis experiments and acoustic cues," *Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen. Odense: Akademisk Forlag, Copenhagen*, pp. 221–232, 1971.

[12] Dominic W Massaro and Michael M Cohen, "The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction," *The Journal of the Acoustical Society of America*, vol. 60, no. 3, pp. 704–717, 1976.

[13] Klaus J Kohler, "F0 in the perception of lenis and fortis plosives," *The Journal of the Acoustical Society of America*, vol. 78, no. 1, pp. 21–32, 1985.

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

[15] Keiichi Tokuda, Heiga Zen, and Alan W Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.

[16] Kim Silverman, "F0 segmental cues depend on intonation: The case of the rise after voiced stops," *Phonetica*, vol. 43, pp. 76–91, 1986.

[17] Achuth Rao M V, Shiny Victory J, and Prasanta Kumar Ghosh, "Effect of source filter interaction on isolated vowel-consonant-vowel perception," *The Journal of the Acoustical Society of America*, vol. 144, no. 2, pp. EL95–EL99, 2018.

[18] Yi Xu, "In defense of lab speech," *Journal of Phonetics*, vol. 38, no. 3, pp. 329–336, 2010.

[19] Uwe D Reichel and Raphael Winkelmann, "Removing micromelody from fundamental frequency contours," in *Proc. of the 5th International Speech Prosody Conference (SP-2010)*, Chicago, USA, 2010.

[20] Tania Ellbogen, Florian Schiel, and Alexander Steffen, "The BITS speech synthesis corpus for German," in *Proc. of LREC 2004*, Lisbon, Portugal, 2004.

[21] P. Boersma and D. Weenik, "Praat: doing phonetics by computer [software]," 2017.

[22] P. Birkholz, P. Schmager, and Y. Xu, "Estimation of pitch targets from speech signals by joint regularized optimization," in *26th European Signal Processing Conference (EUSIPCO 2018)*, Rome, Italy, 2018, pp. 2075–2079.