Applying the quantitative target approximation model (qTA) to German and Brazilian Portuguese

Plínio A. Barbosa[1], *Hansjörg Mixdorff*[2], *and Sandra Madureira*[3]

[1]Speech Prosody Studies Group, Dep. of Linguistics, State Univ. of Campinas, Brazil
[2]Dep. of Computer Science and Media, Beuth University of Applied Sciences, Berlin, Germany
[3]LIACC, Dep. of Linguistics, Catholic Univ. of São Paulo, Brazil
pabarbosa.unicampbr@gmail.com, mixdorff@tfh-berlin.de, madusali@pucsp.br

Abstract

This work is an attempt to explore a different prosodic domain for the quantitative target approximation model (qTA) model than the syllable. This is done by studying the model's ability to synthesise the melodic contours of two different languages, German and Brazilian Portuguese, in two distinct speaking styles, reading and storytelling. The connected utterances studied here present more complex material than hitherto studied using the qTA model. However, the modelling accuracy on these data is similar to that of the Fujisaki model. The results show that the word can be the domain for both prominence marking and phrase boundary type (terminal and non-terminal). By restricting the qTA parameter search space for the two mentioned functions, it is possible to develop an encoding scheme for them.

Index Terms: intonation modelling, speaking styles, cross-linguistic prosody

1. Introduction

This work extends the domain of the PENTA model from the syllable to the word for two prosodic functions. Differing from the melodic-contour-based approaches such as the tilt [11], and the SFC model [1], the Fujisaki model, as well the PENTA model assume that articulatory-related mechanisms guide the generation of the F_0 contours. For the Fujisaki model, the accent and the phrase commands are related to commands generated by separate parts of the cricothyroid muscle in the larynx. For the PENTA model, limits on the speed of target approximation within the syllable domain are imposed by articulatory constraints. Both models are presented in the next two sections, supplemented by details of the procedure of analysis-by-synthesis in each case. The two models interpolate the voiceless portions of the raw F_0 contour, thus providing a value of F_0 for every point in time (cf the Momel model in [4]).

This paper attempts to help develop the PENTA model: (1) by extending its domain to the word in the case of prominence and boundary-type functions; (2) by testing the model in two different languages and two speaking styles; (3) by giving a first sketch of an encoding scheme for two prosodic functions. Due to its longer period of testing, the Fujisaki model is given as a reference for comparison. The investigation reported here is part of a joint work to ultimately compare the two approaches.

2. The Fujisaki model

The Fujisaki model [3] is capable of producing close approximations of natural F_0 contours from two kinds of input commands: phrase commands, implemented by impulse signals, and accent commands, implemented by rectangular pulses. By using these two distinct commands, the F_0 contour is decomposed into three components which are additive in the logarithmic scale: the base frequency component (a constant), the phrase response component, and the accent response component. Whereas phrase commands typically occur at deep prosodic boundaries, accent commands occur in the vicinity of accented syllables, usually to signal the accentedness of a word.

In recent years, the Fujisaki model was applied to many different languages in domains such as the analysis of speaking styles [8] and dialects [5]. The work shown here applies it to two speaking styles in the two aforementionned languages.

2.1. Analysis-by-synthesis in the Fujisaki model

Mixdorff [7] published an automatic approach to the extraction of Fujisaki model parameters which is available as a software tool. The parameters are estimated from an ESPS-waves-based F_0 contour in a multi-step procedure, consisting of a quadratic spline stylisation, a component separation by filtering, followed by command initialization, and a hill-climb search to reduce the overall mean-square-error in the log frequency domain. The process of analysis-by-synthesis is entirely contour-guided, not depending on prespecified domains or prosodic functions.

3. The PENTA and the qTA models

The Parallel ENcoding and Target Approximation (PENTA) model was proposed by Xu [14, 12] to associate tone and intonation into a single generative approach. Its main principle relies on the assumption that the communicative functions control F_0 contours via specific (and parallel) encoding schemes. These encoding schemes specify the values of the melodic primitives, which include (local) pitch target, pitch range, articulatory strength and duration. The values of the melodic primitives can be specified both symbolically (high, low for F_0, long, short for duration, and weak, strong for strength) and numerically. One of the PENTA model's strength is to define F_0 excursions in the vicinity of a communicative function's implementation as a part of the encoding scheme of that function. [15] showed that the deaccenting in the post-focal region of an utterance is necessarily taken as part of the encoding scheme of the focal function, at least for English and Mandarin. In a second example, [12] showed that the rising of both F_0 and intensity at the start of a new topic, both related to the topical function, is part of its encoding scheme. Since the encoding schemes are hypothesised to be language-specific, one of the goals of this paper is to propose encoding schemes for word prominence and the terminal/non-terminal discourse functions in both BP and German. Another strong assumption of the model is the articulatory synchronisation of syllable boundaries and $\mathrm{F_0}$ excursion for tone (or intonation). Even if in Mandarin tone refers to syllable margins as its domain of realisation, this does not need to be the case for every function and language. In English, pitch target and range refer to word edges. The way Xu and colleagues implemented focus in English suggests a domain larger than the syllable for that function, given the behaviour of the $\mathrm{F_0}$ in the post-focal region in the case of contrastive focus [15]. Recently, [10] clearly stated that "the issue of syllable synchronization is certainly unresolved, and much further research is needed".

In the PENTA model, a pitch target refers to the underlying pitch trajectory typically associated with a syllable. Due to the model's assumptions, pitch targets have to be linked to some communicative functions, and can be either static (high, low or mid), or dynamic (rise or fall). At the end of the generation, the modelled $\mathrm{F_0}$ contour is implemented by the quantitative Target Approximation model (qTA) explained in [9].

The quantitative value of the melodic primitives are used together to specify an asymptotic contour given by equation 1.

$$F_0(t) = (c_1 + c_2.t + c_3.t^2).e^{-\lambda.t} + m.t + b \qquad (1)$$

The coefficients of the target line equation are the target slope (m) and the target height (b), whereas the λ parameter specifies the rate of approaching the target line [9]. The coefficients c_1, c_2 and c_3 are related to the initial conditions. Thus, given these initial conditions, three parameters completely specify the modelled $\mathrm{F_0}$ contour for a given linguistic segment. Thus, the qTA model generates a left-to-right local implementation of a melodic contour where the absolute $\mathrm{F_0}$ values, together with the first and second $\mathrm{F_0}$ derivatives in an utterance segment are defined as the last values of the immediately previous utterance segment [10]. The model parameters can be obtained from an analysis-by-synthesis learning process, which can operate in two modes: (1) identical search in the parameter space for all prespecified intonational functions, and (2) constrained search guided by the encoding scheme of the respective function realised in a particular utterance segment (e.g., the syllable or the word).

4. Modelling the melodic contour of reading and storytelling in German and BP

4.1. Corpora

The corpora consist of parallel productions in BP and German. Two native female and four native male speakers in both languages read a $1,500$-word text on the origin of the pastries *pastéis de Belém* (reading style, RE) in their own language. The BP text is an adaptation of a text in European Portuguese (EP), whereas the German text was translated from the BP text by the second author. The translation was made sentencewise to make comparisons easier. After the reading, all subjects told what the text was about (storytelling style, ST). All speakers were aged 30 to 45 years, and were students with a Linguistics or a Computer Science background. The analyses shown here are based on excerpts from 150 to 200 words in each language and style.

4.2. Analysis-by-synthesis modelling

All the utterances in the two speaking styles were modelled with the two production-based models. In the case of the Fujisaki model, the Praat [2] $\mathrm{F_0}$ tracker served as a front end for obtaining the $\mathrm{F_0}$ traces. The search was limited between 120 and 450 Hz for females, and between 70 and 350 Hz for males at a 10 ms pace in both cases. Modelled $\mathrm{F_0}$ contours were obtained using the *FujiParaEditor* [7, 6]. After the automatic extraction of the parameters, manual correction was done whenever necessary to assure that (1) phrase command positions precede a prosodic phrase, which is usually an intonational phrase, and that (2) accent commands correspond to accented linguistic units or allow the maintenance of a certain $\mathrm{F_0}$ level in upstepped chunks. These manual corrections were made by native speakers, the first author for BP, and the second for German.

In the case of the qTA model, the PENTA_trainer version 1.4 script [13], which runs onto Praat, was used for modelling the contours. The same genderwise $\mathrm{F_0}$ limits used for the Fujisaki model were used for the qTA model. Both duration and intensity of the respective linguistic unit were unchanged. For the sake of defining the encoding schemes of two particular functions, word prominence and terminal/non-terminal discourse function, all utterances were segmented into words, within which the search for the three qTA parameters, target height and slope, and target strength, took place. These words were labelled auditorily by the authors as prominent (p), or non-prominent (n) for prominence, and terminal (t) or non-terminal (c) for phrase boundary. Prominence is understood in the general sense, which is related to focus. There are no examples of the special case of contrastive focus in the data examined here.

One of the goals of this paper is to show that the word domain allows the implementation of the functions presented here, with an amendment with respect to secondary stress. An example of modelling result for the BP sentence "O que mais lhe custava, no entanto, era ter de se levantar no meio da noite para rezar as Matinas." (What cost more to him, however, was to have to get up in the middle of the night to pray the Mornings.) can be seen in Fig. 1. The German utterance corresponding to this sentence can be seen in Fig. 2.

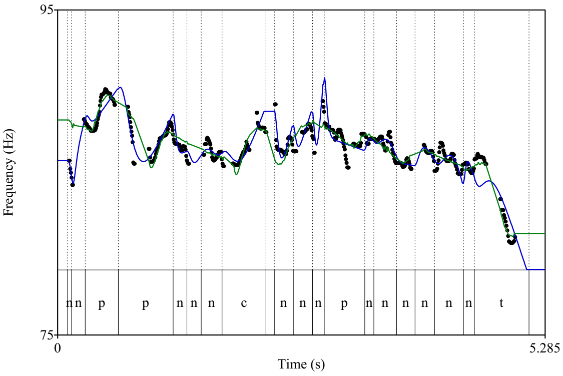

Figure 1: *F_0 traces in st re 1 Hz for the Fujisaki (green) and the qTA models (blue) for the read sentence (prominent words in bold): "O que **mais-lhe custava**, no ... no entanto (c), era ter de se **levantar** no meio da noite para rezar as Matinas (t)." by BP male speaker MT. Prominent words marked with label* p. *See text for further details. Original F_0 in black. Audio files: BMT-3Orig.wav, BMT-3Fuji.wav, BMT-3qTAGen.wav.*

Observe the sequence of six non-prominent BP words in Fig. 1 between the final *p* and the *t* labels, which presents a downdrifting $\mathrm{F_0}$ trace. Notice also that the non-terminal function is implemented by a rising-falling pattern in BP, whereas it is signalled by a rising pattern in German (Fig. 2).

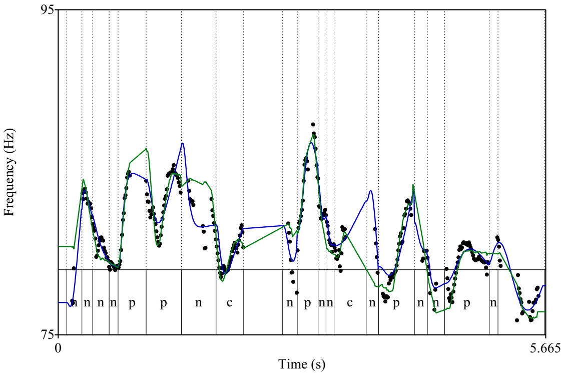

Figure 2: *F_0 traces in st re 1 Hz for the Fujisaki (green) and the qTA models (blue) for the read sentence: "Was ihn aber am* ***meisten Überwindung*** *kostete war (c), sich* ***mitten*** *in der Nacht(c) zu* ***erheben*** *, um das* ***Morgengebet*** *zu sprechen (t)." by German male speaker G1. See text and Fig. 1 for further details. Audio files: G01-3Orig.wav, G01-3Fuji.wav, G01-3qTA.wav*

For evaluating the qTA algorithm and examining the behaviour of the parameter values for the four labels, the two search modes presented in section 3 were used. First, a search unspecified for function was carried out with the default constraints of the qTA algorithm for the three parameters: target slope, height, and strength. This search is referred to as the general search. From the analysis of the parameters behaviour for the four labels, a second search mode was used, referred here as the constrained search.

5. Results

Table 1 shows two measures of performance, Root Mean Square Error (RMSE) and correlation medians and standard-deviations, for the first thirteen sentences of the read material (188 words), and ten utterances in the ST style for the two languages ($150 - 200$ words). The qTA algorithm is tested in the general search mode. Observe that both models perform very similarly. Correlations are slightly lower for the qTA model for the German data. This has to do with the realisation of secondary stress, because words bearing it also exhibit an additional F$_0$ peak in primary stress position, and the qTA can only generate one single peak per domain.

Table 1: Median (and STD) of RMSE values in semitones and correlation (R) for the first 13 utterances in RE style, and 10 utterances in ST style in BP and German (G). Results refer to the Fujisaki (F), and the qTA models in the general search mode and the word domain.

		RMSE		*R*	
L.	St.	F	qTA	F	qTA
BP	RE	1.2(0.6)	1.0(1.0)	0.99(0.02)	0.99(0.01)
G	RE	1.3(1.2)	1.4(0.5)	0.97(0.03)	0.90(0.13)
BP	ST	1.3(0.8)	1.2(0.6)	0.99(0.02)	0.99(0.01)
G	ST	1.2(0.6)	1.3(0.7)	0.98(0.01)	0.92(0.19)

A closer look at the qTA model parameters associated with the prosodic functions studied here is a first step towards the proposal of an encoding scheme for each function. In Fig. 1, the parameter triplet (slope, height, strength) for the three prominent words are $(8, 5, 60)$ for *mais*, $(12, 3, 19)$ for *custava*, and $(-2, 1, 80)$ for *levantar*. Observe that all prominent words have positive heights for the targets. The last one has a slope near zero because it is a weaker prominence. To the final word, *Matinas*, was attributed the terminal boundary function with parameters $(-17, -7, 20)$, with negative target height and slope. The non-terminal function is implemented by the triplet $(15, 3, 20)$ in that case. Observe, though, the slight fall after the rising. As for the German utterance in Fig, 2, the triplets are $(-3, 8, 47)$ for *meisten*, $(22, 10, 30)$ for *Überwindung*, $(-47, 9, 26)$ for *mitten* (which has a sharp fall after the peak), and $(-8, 2, 15)$ for *Morgengebet*, also with a fall after the peak. The non-terminals have positive slopes and heights: $(16, 5, 50)$ for *war*, and $(28, 7, 20)$ for *Nacht*. The last word has $(27, 1, 11)$, due to the short rising at the end.

After studying the distribution of target slope and height according to the four labels, the same trend was found for all subjects in the two speaking styles. Fig. 3 presents histograms for target slope for BP and German in the two speaking styles for the *n* label (non-prominent word). It is the only label for which the slope parameter has a bimodal distribution. All other labels have unimodal distributions for target slope and height. A left mode of negative slope values appears in both languages and styles, as well as a mode around zero with a considerable spread. This bimodality together with the variation is related to transitions between peaks and a baseline, as can be seen in Fig. 1 between the last p word and the last word associated with a terminal function. Negative slopes occur due to F$_0$ transitions from peaks towards the baseline, and positive slopes, from a lower, preceding position to F$_0$ peaks.

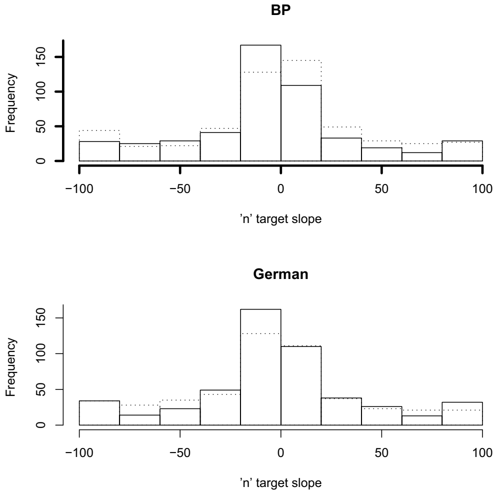

Figure 3: *Histograms for target slope (st/s) for BP (top) and German (bottom) for the non-prominent label. Plain line is ST style and dotted line, RE style.*

Tab. 2 shows the results for the target height parameter in both languages and styles. Observe how the target heights for the terminal function are mainly spread in the negative scale regardless of language and style. The non-terminal function target height is spread mainly in the positive scale for the RE style in German. It is practically symmetrical around zero in the ST style. In BP, it is perfectly symmetrical, regardless of style. This is related to the rising-falling pattern in BP for non-terminals, which are realised quite variably. Target heights for prominent words do not seem to concentrate in any region besides a mean close to zero. These figures motivated a restricted search for testing the encoding scheme for the terminal function in all cases, and a positive height for the *c* function in German RE style. A negative height with zero slope was then imposed to the former, whereas a positive height with zero slope for the latter. The effect of the restricted search can be seen in Fig. 4 for the final part of the BP utterance illustrated in Fig. 1. Observe the change in the qTA-modeled trace. The parameters triplet for the final word entailed a less closer contour in relation to the original one. The corresponding audio file, when compared with the one obtained with the general search mode, reveals a slight difference in degree for the same terminal function. A better result is obtained when a secondary stress is assigned to the syllable "Ma" in the last word.

Table 2: Target height mean (and 90 % CI) in st re 1 Hz for the four function labels (f) for both languages and styles.

	G		*BP*	
f	RE	ST	RE	ST
n	0(−9 8)	−2(−10 7)	1(−9 10)	0(−8 9)
p	3(−5 12)	1(−8 10)	2(−8 11)	2(−6 10)
t	−3(−15 9)	−3(−12 5)	−6(−18 5)	−3(−13 7)
c	5(−3 14)	3(−8 10)	0(−13 13)	0(−10 10)

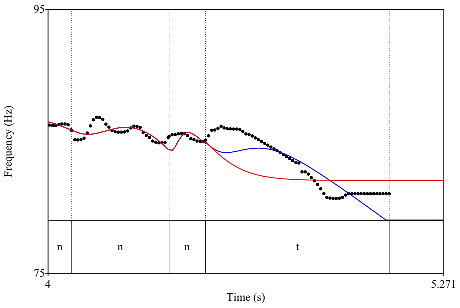

Figure 4: *Effect of restricted search in the final part of the Fig. 1 utterance. General search in blue, restricted search in red (slope for* t *is forced to zero). Audio file: BMT-3qTARst.wav*

6. Discussion and conclusion

The values of RMSE and correlation revealed that the qTA model has virtually the same performance as the Fujisaki model when working in the word domain. An exception is the implementation of F_0 peaks in secondarily stressed syllables. Since the qTA contours can only have one lobe within a prespecified domain, it's impossible to generate two F_0 peaks within the word limits. Being contour-guided, the analysis-by-synthesis for the Fujisaki model doesn't suffer from this restriction. To amend that, it's necessary to delimit the secondarily stressed syllable before using the qTA algorithm for this restricted domain. This in accordance with its principles. The target slope for the non-prominent words indicates that the absence of prominence is probably targetless: the F_0 trace drifts to a baseline after realising the peak. This could indicate that, if no function at the intonational level needs to be realised, no underlying qTA targets need to be specified. As it was shown here, the terminal function has some potential to be encoded as a communicative function, as well as the non-terminal function in German reading, aspects that need to be further explored.

7. Acknowledgements

The authors acknowledge the grant 490726/2008-9 from CNPq and 444 BRA 113 59 0-1 from DFG. The first author also acknowledges the CNPq grant 300371/2008-0. The BP corpus was recorded in the context of the FCT project PTDC/PLP/72404/2006, for which INESC-ID Lisboa had support from the "Quadro Comunitário de Apoio III". The authors thank J. Batella for adapting the original EP text to BP, all the speakers for their time, M. Tamashiro, M. A. Svicero and I. Santos for help with the BP data annotation, and E. Yahel and A. Hilbert for help with the German data. Y. Xu and S. Prom-on are thanked for support with the PENTA_trainer.

8. References

[1] Bailly, G., Holm, B., "Learning the hidden structure of speech: from communicative functions to prosody", Cadernos de Estudos Linguísticos, 43: 37-54, 2002.

[2] Boersma, P., Weenink, D., "Praat: doing phonetics by computer" (Version 5.1.37) [Computer program], Online: http://www.praat.org.

[3] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", Journal of the Acoustical Society of Japan (E), 5(4): 233–241, 1984.

[4] Hirst, D.J., Espesser, R., "Automatic modelling of fundamental frequency using a quadratic spline function", TIPA 15: 71-85, 1993.

[5] Leemann, A., Beat, S., "Intonational and temporal features of Swiss German", in Proc. ICPhS, Saarbrücken, 957–960, 2007.

[6] Mixdorff, H., "FujiParaEditor" Online: http://www.tfh-berlin.de/∼mixdorff/fujisaki_analysis.htm

[7] Mixdorff, H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", in Proc. of ICASSP, Istanbul, 3:1281–1284, 2000.

[8] Mixdorff, H. and Pfitzinger, H.R., "Analysing Fundamental Frequency Contours and Local Speech Rate in Map Task Dialogs", Speech Communication, 46:310-325, 2005.

[9] Prom-on, S., Xu, Y., "Articulatory-Functional Modeling of Speech Prosody: A Review" Proc. Interspeech 2010, Makuhari, 46–49, 2010.

[10] Prom-on, S., Xu, Y., Thipakorn, B., "Modeling tone and intonation in Mandarin and English as a process of target approximation" J. Acoust. Soc. Am., 125 (1): 405–424, 2010.

[11] Taylor, P. "Analysis and synthesis of intonation using the tilt model", J. Acoust. Soc. Am., 107: 1697-1714, 2000.

[12] Xu, Y., "Speech melody as articulatorily implemented communicative functions" Speech Communication, 46: 220-251, 2005.

[13] Xu, Y., Prom-on, S., "PENTAtrainer.praat", Online: http://www.phon.ucl.ac.uk/home/yi/PENTAtrainer/

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

[15] Xu, Y., Xu, C.H., "Phonetic realization of focus in English declarative intonation", J. Phon., 33: 159-197, 2005.