The qTA Toolkit for Prosody: Learning Underlying Parameters of Communicative Functions through Modeling

Santitham Prom-on[1], *Yi Xu*[2]

[1] Department of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand
[2] Department of Speech, Hearing, and Phonetic Sciences, University College London, UK

santitham@cpe.kmutt.ac.th, yi.xu@ucl.ac.uk

Abstract

This paper presents the qTA toolkit, a general-purpose research toolkit for studying speech prosody. The toolkit consists of analysis and visualization tools. The analysis tool processes F_0 and timing data together with annotation of communicative functions to estimate function-specific underlying pitch targets and their function-specific adjustments. The visualization tool generates illustrations of the synthesized F_0 contour and their pitch target input. As an initial test, the qTA toolkit is applied to a Mandarin corpus, and the results suggest that it can be effectively used for investigating prosody in terms of communicative functions.

Index Terms: quantitative target approximation, qTA model, pitch target, communicative function, research toolkit

1. Introduction

Prosody plays a crucial role in human speech communication. A widely used approach in speech prosody investigation involves forming a hypothesis about the prosodic events under scrutiny, observing the related speech features such as fundamental frequency (F_0), duration or intensity, and then postulating the mechanisms in the form of a computational model that describes the prosodic events. If the model has been thoroughly tested, it can then be used for either conducting further studies of other prosodic events or implementation in a text-to-speech (TTS) system. The continuation of this research cycle is important for the advancement of speech science and technology.

One of the major factors that limit the progress of speech prosody research is the lack of a general-purpose computational tool for studying prosodic events based on an articulatory-based phonetic framework. Such a tool will allow speech prosody researchers to investigate articulatory-based hypotheses without the need to write their own programs. The introduction of such a tool may facilitate further development in speech prosody research.

The quantitative target approximation (qTA) model [1] and its theoretical counterpart, the parallel encoding and target approximation (PENTA) model [2], offer a well-defined articulatory-based functional framework for modeling speech prosody. In this framework, directly observable prosodic events are regarded as the end results of implementing underlying targets associated with various communicative functions. For example, F_0 peaks are viewed not as directly linked to any communicative meanings, but as part of the F_0 responses that occur when functions such as tone and focus are implemented [2]. In the case of focus, the full response would include pitch range expansion at the on-focus position and pitch range compression after focus. Such functional specifications thus link the meanings not to the surface F_0 changes directly but to modifications of the parameters in the process of target approximation. Figure 1 shows an example F_0 contour of an utterance in Mandarin with the tone sequence HHFHH and narrow focus on the first two syllables. H and F denote the high and falling tones respectively. The encoding scheme of the focus function is to expand the pitch range of the focused word by raising the [high] pitch targets (or lowering the [low] targets when applicable) of the focused syllables while compressing the pitch range of the post-focus syllables by lowering their pitch targets [1].

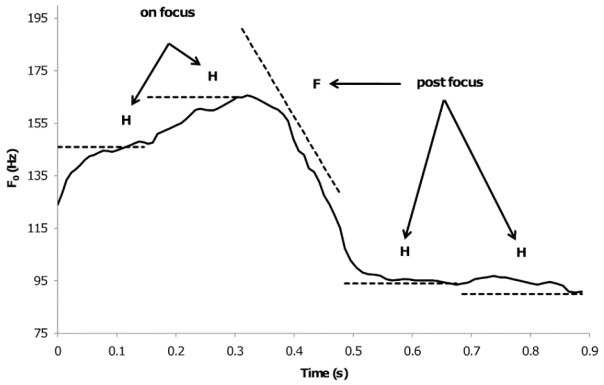

Figure 1: *An example of surface F_0 response and its corresponding pitch targets when interpreting focus as a communicative function. The solid line indicates the F_0 contour while the dashed line indicates the pitch target of each syllable.*

The qTA model thus allows researchers to look beyond the surface F_0 response to study the underlying prosodic representations of communicative functions. Communicative functions in qTA are associated not to F_0 values directly, but to an intermediate representation known as pitch targets and their adjustments. A pitch target can be thought of as a goal of laryngeal muscular control. A sequence of pitch targets are implemented consecutively in qTA to produce F_0 responses that correspond to specific communicative functions. Because critical articulatory mechanisms of pitch production are built into qTA as part of the intrinsic properties, researchers can more easily identify the underlying targets and their modifications as related to communicative functions.

In this paper we present the qTA toolkit, a general-purpose prosody research toolkit for studying communicative functions encoded by pitch variations. The toolkit processes the F_0 input data and their corresponding functional annotations to estimate the pitch target of each syllable. The communicative functions can then be computed from the estimated targets based on the known functional combinations. For visualization purposes, the shape of the F_0 contour synthesized from the given parameters can be inspected with the help of the qTA visualization tool.

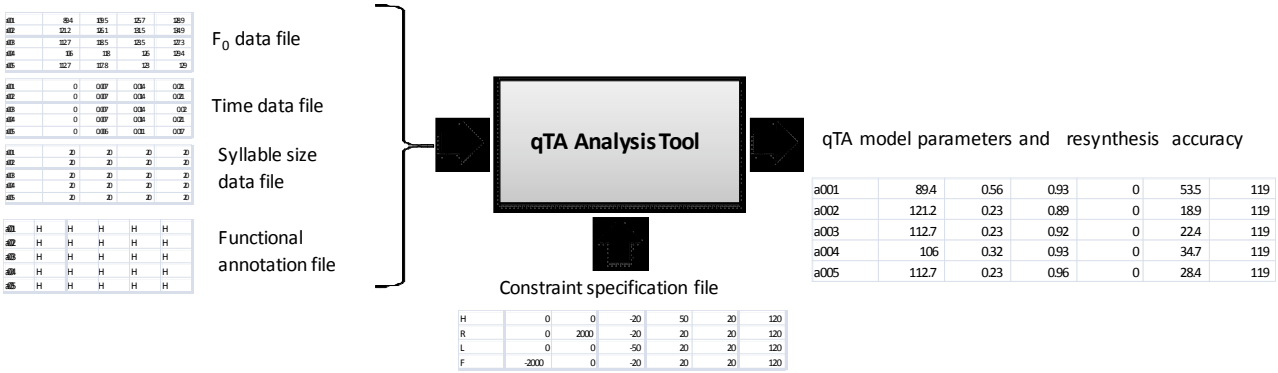

Figure 2: *Input and output specifications and requirements of the qTA Analysis Tool.*

Table 1. *The result file header format.*

Column	Result
1	Utterance ID
2	Utterance Initial F_0 Value
3	Resynthesis RMSE (semitone scale)
4	Resynthesis Correlation
5…	qTA Model Parameters (Group of 3 columns)

Table 2. *The qTA model parameter sequence.*

Column	Result
1	Pitch Target Slope, *m*
2	Pitch Target Height, *b* (relative to initial F_0)
3	Approximation Rate, λ

2. qTA Toolkit

The qTA toolkit consists of two main tools that perform two major functions: analysis and visualization. The qTA analysis tool allows researchers to estimate the underlying pitch targets of the primary communicative function from the F_0 contour. Researchers who want to inspect the synthesized F_0 contour along with the specified pitch targets can use the qTA visualization tool to interactively display and adjust the graph of both synthesized F_0 contour and pitch targets. Combining the power of both tools makes the qTA toolkit effective for investigating speech prosody in terms of communicative functions. The qTA toolkit has been implemented in JAVA, a powerful and versatile programming language, and can be downloaded from http://www.cpe.kmutt.ac.th/~promon/qta.

2.1. qTA analysis tool

To learn pitch targets of the given communicative function, the qTA toolkit requires the following input data:

- F_0 data file: This data file contains a tab-delimited sequence of F_0 values. The first column in this file contains a metadata of each utterance. This metadata should be unique since they are used as references to the specific utterances. Each line in this data file corresponds to an utterance in the speech corpus.
- Time data file: The time data file contains a tab-delimited sequence of sampling time. Each time value is paired with a corresponding F_0 value in the F_0 data file. Similar to the F_0 data file, each line in this file corresponds to an utterance in the speech corpus.
- Syllable size data file: This file contains a tab-delimited sequence of number of samples of each syllable. For each line, the summation of all number of samples must equal to the number F_0 and time samples in the F_0 and time data files.
- Functional annotation file: This annotation file contains a tab-delimited sequence of symbols representing different categorical input of communicative functions. These symbols are defined in the constraint specification file. In each line, the number of symbols has to be equal to the number of syllables in the syllable size data file.
- Constraint specification file: This file contains tab-delimited functional specifications. It specifies the categorical input of the communicative function as well as its corresponding qTA model parameter constraints. Each constraint defines a search space for a model parameter. The first column defines a unique symbol representing categorical outputs of a communicative function. The second and third columns define the minimum and maximum values of the pitch target slope. The third and forth columns define the minimum and maximum values of the pitch target height. The fifth and sixth columns define the minimum and maximum values of the target approximation rate. Each line in this file indicates distinct possible output of the function.

The qTA analysis tool reports the analysis results in terms of F_0 resynthesis accuracy and the sequence of pitch targets. The format of the result file is shown in Table 1 and Table 2. The result file consists of two main parts: the header and the parameter sequence. The header is located in the first four columns of each line as shown in Table 1. The utterance initial F_0 value is required as a result because the qTA F_0 synthesis depends partly on this value. The details of the qTA model will be discussed in the next subsection. The accuracy is measured in terms of root mean square error (RMSE) and Pearson's correlation coefficient. The RMSE indicates the average distance between the original and synthesized F_0 contour while the correlation indicates similarity of contour shapes between the two. These accuracy measures are located in the third and forth columns of each line in the result file. The remaining columns contain the qTA model parameters of each syllable grouped in a set of three columns. Table 2 shows the designated model parameters of each column in each set.

2.2. qTA model

The core of the toolkit is the qTA model [1]. The qTA model is an articulatory-oriented F_0 control model for simulating tone and intonation. The model represents F_0 as a response of a pitch target approximation process [1-3]. In the qTA model, a pitch target is defined as the underlying goal of the local prosodic event [1]. It can be represented by a simple linear equation

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

where $x(t)$ is a pitch target depicted as a dashed line in Figure 1. m and b are the slope and height of the pitch target, respectively. Because target approximation is always local to the host syllable [3], the time, t, is relative to the onset of the syllable.

In the qTA model, F_0 is represented as the response of the vocal fold tension control mechanism driven by the pitch target [1]. The core mechanism of the model is represented as a third-order critically damped linear system. Thus, F_0 can be expressed mathematically as

$$f_0(t) = x(t) + (c_1 + c_2 t + c_3 t^2)e^{-\lambda t} \tag{2}$$

where the first term, $x(t)$, is the forced response which is the pitch target and the second term, the polynomial and the exponential, is the natural response of the tension control system. The model parameter λ represents the rate of target approximation which controls the completeness of target approximation within a given amount of time. The transient coefficients c_1, c_2 and c_3 are determined by the initial conditions and other model parameters of the current syllable. The initial conditions of the articulatory process include initial F_0 level, $f_0(0)$, initial velocity, $f'_0(0)$, and initial acceleration, $f''_0(0)$. By solving the systems of linear equation determined from the initial conditions, the transient coefficients can be calculated from the following formulas

$$c_1 = f_0(0) - b \tag{2}$$

$$c_2 = f_0'(0) + c_1\lambda - m \tag{3}$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2 \tag{4}$$

2.3. Parameter extraction and constraints

For each syllable, the qTA model requires two pitch target parameters, m and b, and a rate of approximation, λ. These parameters are estimated using an automatic analysis-by-synthesis procedure. As shown in Figure 2, the qTA analysis tool reads the data files for F_0 and time, the annotation file for the functional annotation of each syllable, and the constraint specification file for the range of parameter search space of each form of the specified communicative function. The program automatically varies the parameters values in the specified search space, and returns the parameter set with the lowest sum square error between the synthesized and original F_0 contours. This procedure is executed one syllable at a time, from the first syllable to the last syllable of an utterance.

One of the important features of the qTA analysis tool is that it allows users to specify and adjust the parameter constraints for each of the communicative function. By adjusting these parameter constraints, researchers can test hypotheses about these functions. The hypotheses may be based on empirical research. For example, for Mandarin, there is empirical evidence that the H and L tones have static targets while the R and F tone have dynamic targets [3]. Such empirical knowledge can be further tested by setting the parameter constraints so that the slope of H and L tones is fixed to zero, the slope of F tone is always positive, and the slope of F is always negative, as done in [1]. The hypotheses can also be based on informal observations, and setting the parameter constraints derived from those observations would still be an effective way of testing the hypotheses. It is also possible to set a large constraint space so that target estimation is more dependent on the modeling process. This may be desirable in case where little is known about the prosody of the languages under scrutiny.

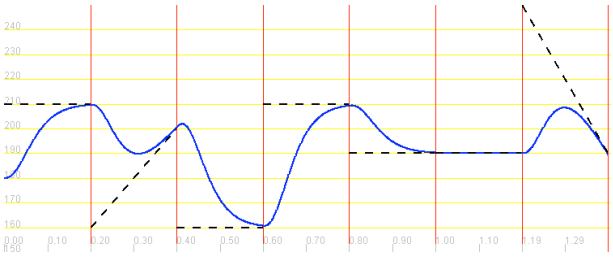

Figure 3: *The output example of the qTA visualization tool. The solid line indicates the synthesized F0 contour while the dashed line indicates the pitch target. The vertical line demarcates the syllable boundary.*

Table 3. *Mean and confidence interval of the qTA model parameter of the tone function in a Hertz scale.*

Tone	*m* (*Hz/s*)	*B* (Hz)	*λ*
H	0	-14.4±10.86	74.13±2.28
R	150.63±52.56	-37.21±14.44	56.85±5.41
L	0	-91.40±36.80	35.96±5.44
F	-295.70±137.21	-21.29±18.87	58.84±5.87

2.4. qTA visualization tool

As part of the toolkit, the qTA visualization tool provides a means to inspect the variation of the synthesized F_0 along with the pitch target input. The tool relies on the core qTA model to synthesize the F_0 contour according to the specified sequence of qTA model parameters. Figure 3 shows an example of F_0 contours and the underlying pitch targets plotted by the visualization tool. Users can adjust and visualize the synthesized F_0 contour, the pitch target, and the target approximation rate interactively. With this tool one can therefore perform informal testing of various hypotheses before conducting any systematic experiments.

3. Results

For demonstration purposes, a Mandarin corpus that was collected in an acoustic study of tone and focus [4] and quantitative modeling [1] was used. The corpus consists of extracted F_0 and time values of 3840 five-syllable utterances recorded by 4 male and 4 female native Mandarin speakers.

Using the qTA analysis tool, we extracted the model parameters and resynthesis accuracy and saved them in the output files. Summarizing the 95% confidence interval of RMSE and correlation of all speakers together, we obtained an average RMSE of 0.78±0.20 semitone and an average correlation of 0.96±0.01. These low RMSE and high correlation values indicate that the pitch targets obtained through the automatic parameter extraction procedure can accurately represent the underlying goals in F_0 control.

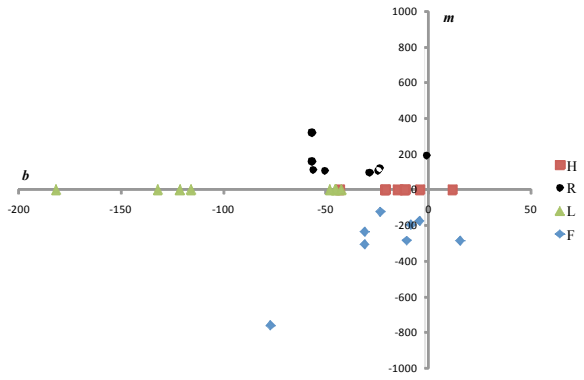

Figure 4: *Pitch target distribution of the tone function. The square, round, triangle, and diamond symbols denote the H, R, L, and F tones, respectively.*

To summarize the general shape of the pitch target representing each output of the tone function, including the H, L, F and R tones, we calculated the averaged values and the 95% confidence intervals of the qTA model parameters, as shown in Table 3. The values of the model parameters seem to be consistent with the conventional observation of the form of the pitch target of Mandarin tone [3], but it should be noted that the zero slope of H and L tones is the result of the slope constraints. Interestingly, the higher λ of the H tone compared to the L tone suggests that target approximation of the H tone is more complete than that of the L tone, which is something that can be more closely examined in future research.

Figure 4 shows the distribution of the pitch target slope and height of different tones estimated from different speakers using the qTA tools. Despite speaker variability, distinct clusters of pitch target can be clearly observed from the figure. These clusters indicate the general trend of the pitch targets corresponding to the tone function.

4. Discussion

Quantitative modeling of prosodic events provides a means to investigate hypotheses and assumptions of prosody generation and manifestation. Tools for quantitative modeling are thus very important to the development of prosody research. There are a number of existing tools that can be used for studying the prosody-related speech feature such as F_0 and duration, or modeling the surface F_0 variation [5-9], but only one of them allows researchers to investigate the underlying articulatory representations of prosodic events [9].

As the core of the new toolkit, the qTA model can generate F_0 contours quite accurately as shown earlier. This demonstrates that the model can capture the essential components of F_0 control. While it is able to achieve comparable synthesis accuracy to previous modeling studies [1, 8, 10-12], the qTA model requires fewer parameters than do other systems in F_0 synthesis. This indicates the effectiveness of the qTA approach.

Beside synthesis accuracy, the form and constraints of the communicative functions can also be customized in the qTA analysis tool. This flexibly allows researchers to apply the toolkit for studying different communicative functions. This customizability can be achieved because the model allows a clear separation of articulatory-based prosody generation mechanism from the communicative meanings encoded by prosody. From the pitch target parameters estimated by the qTA toolkit, researchers can then identify the encoding characteristics of specific communicative functions by performing condition-specific averaging and then computing the differences from the condition that contains the smallest number of functions [1].

The current version of the qTA toolkit still has certain limitations. The first is that the qTA analysis tool is still semi-automatic, so that the investigation of the communicative functions requires much manual control. The feature for fully automatic analysis will be added to the qTA analysis tool in subsequent development. The second limitation is that, in order to properly set optimal values of the constraints, researchers need to know, prior to the study, the general forms of the communicative function. Thus, the current version of the toolkit still lacks discovery capability. Again, such capability will be added in further development of the toolkit.

5. Conclusion

This paper presents a qTA toolkit for modeling communicative functions in terms of underlying pitch targets. The toolkit relies on the core mechanism of the qTA and PENTA models. It allows researchers to investigate underlying pitch targets of the prosodic events under scrutiny. The qTA toolkit includes analysis and visualization tools which can be used to extract pitch targets, estimate communicative functions, measure synthesis performance, and visualize the synthesized F_0 along with the pitch targets. In a test run, the toolkit was applied to a Mandarin speech corpus. The results show that it can accurately simulate the F_0 contours and estimate the patterns of pitch targets underlying the tone function in the language.

6. References

[1] Prom-on, S., Xu, Y. and Thipakorn, B., "Modeling tone and intonation in Mandarin and English as a process of target approximation", J. Acoust. Soc. Am., 125(1):405-424, 2009.

[2] Xu, Y., "Speech melody as articulatorily implemented communicative functions", Speech Commun., 46(3-4):220-251, 2005.

[3] Xu, Y. and Wang, Q. E., "Pitch targets and their realization: Evidence from Mandarin Chinese", Speech Commun., 33(4):319-337, 2001.

[4] Xu, Y., "Effects of tone and focus on the formation and alignment of F_0 contour", J. Phonetics, 27:55-105, 1999.

[5] Boersma, P., "Praat, a system for doing phonetics by computer", Glot. International, 5(9-10):341-345, 2001.

[6] Huang, Z., Chen, L. and Harper, M. P., "An open source prosodic feature extraction", in the Proceedings of Language Resource and Evaluation, May 2006.

[7] Cutugno, F., D'Anna, L., Petrillo, M. and Zovato, E, "APA: towards an automatic tool for prosodic analysis", in the Proceedings of Speech Prosody 2002, 231-234, April 2002.

[8] Hirst, D. and Espesser, R. "Automatic modelling of fundamental frequency using a quadratic spline function," Travaux de l'Institut de Phonétique d'Aix, 15:75–85, 1993 [http://www.icp.inpg.fr/~loeven/Praat/momel_english.html]

[9] Mixdorff, H., Fujisaki, H., Chen, G. P. and Hu, Y., "Towards the automatic extraction of Fujisaki model parameters for Mandarin", in Proceedings of EUROSPEECH 2003, 873-876, September 2003. [http://public.tfh-berlin.de/~mixdorff/thesis/fujisaki.html]

[10] Fujisaki, H., Wang, C., Ohno, S. and Gu, W. "Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model", Speech Commun., 47:59-70, 2005.

[11] Kochanski, G. and Shih, C. "Prosody modeling with soft templates", Speech Commun., 39:311–352, 2003.

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