Modeling tone and intonation in Mandarin and English as a process of target approximation

Santitham Prom-on, Yi Xu, and Bundit Thipakorn (2009)

Presented by: Barbara Helene Schmehl

Table of contents

I.

Introduction

Brief Introduction to Tonal Languages: Tone and Intonation in English vs Mandarin

- Mandarin is a **tonal language** considered to have five distinct tones:
  - **High** (H) - constant high pitch
  - **Rising** (R) - starts mid, then rises
  - **Low** (L) - dips low before rising slightly
  - **Falling** (F) - starts high, then drops
  - **Neutral** (N) - short and unstressed with minimal pitch variation
- English uses **intonation** to convey meaning
  - Highlight a particular word or information structure
  - Emphasis

Goal

The paper outlines an attempt to simulate tone, stress, and focus in Mandarin and English with a quantitative model that generates surface **F0 contours** through the process of target approximation in order to

- create a robust model for use in speech technology
- test current understandings of tone and intonation

Vocabulary:

F0 - pitch or the fundamental frequency

Motivation & previous models

Direct F0 models

- Mainly based on **pitch contours** with minimal consideration for **articulatory process**
- **Examples**: quadratic spline (Hirst and Espesser, 1993), Pierrehumbert (Pierrehumbert, 1981), tilt (Taylor, 2000), SFC (Bailly and Holm, 2005)
- **Limitations**: No differentiation between information-carrying patterns and articulatory effects, “superficial”

Underlying mechanism models

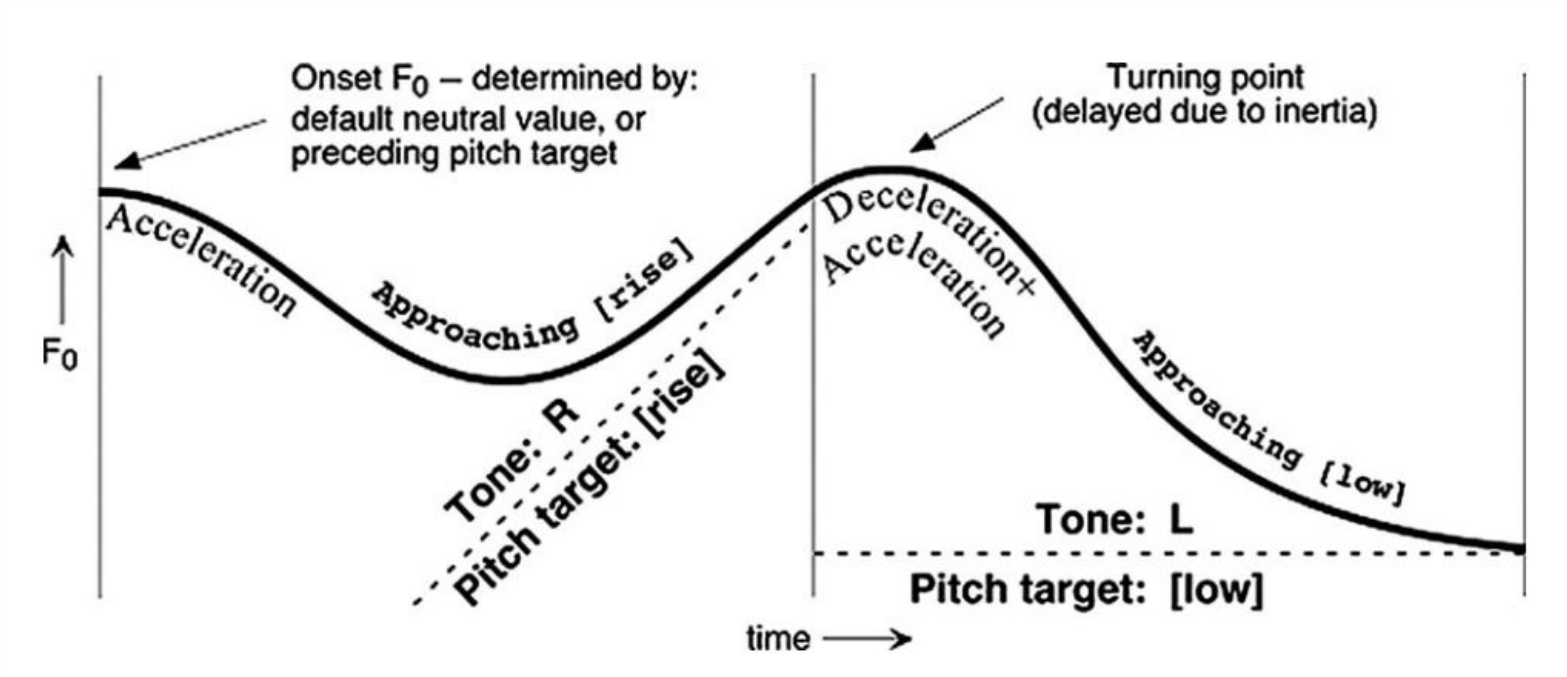
- Based on assumptions about the **process of F0 production**
- **Examples**: soft-template (Kochanski et al., 2003), command response (CR) (Fujisaki et al., 2005)
- **Limitations**: Assumes individual muscle control inconsistent with physiology, requires too many parameters per syllable

TA model (i)

The shortcomings of the SoA models were the motivation of the TA (target approximation) model proposed in 2005 (Xu and Wang, 2005):

- Assumes that observable F0 contours are the result of the **implementation of pitch targets**, which are linear functions that can be either
  - **static** - neither rising nor falling; **slope of zero** (**Ex**: low pitch)
  - or **dynamic** - rising or falling; **positive** or **negative slope** (**Ex**: rising pitch)
- The implementation of the pitch targets is synchronized with the syllable, beginning at syllable onset and ending at syllable offset
- Pitch contour is influenced by the preceding syllable's articulatory state (F0, velocity, and acceleration), creating a carryover effect that explains phenomena like F0 peak delay

TA model (ii) (Fig 1)

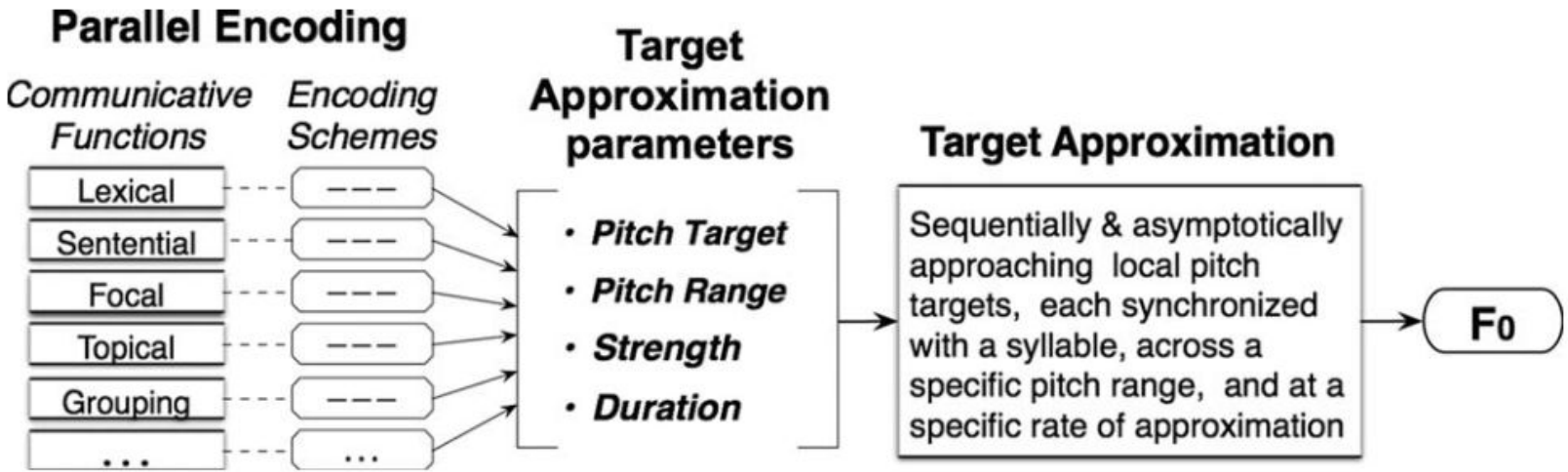


PENTA (i)

Parallel Encoding and Target Approximation model

- **Parallel Encoding** - Multiple communicative functions are **encoded simultaneously** in speech prosody (e.g. tone, intonation, focus, stress) without hierarchy
- **Encoding schemes** - Each communicative function is mapped to **distinct encoding schemes** which specify **language-specific or universal** rules for determining pitch target, range, strength, and duration
- **Target Approximation (TA)** - The model uses TA to execute these encoded functions, shaping F0 contours

PENTA (ii) (Fig 2)



qTA model

TA and PENTA are theoretical approaches that need to be **tested quantitatively** - enter the **quantitative target approximation model**, the outcome of quantifying both TA and PENTA!

II.

Modeling Biophysical Mechanisms of F0 Production

A: Background Assumptions (i)

1. Vocal cord control as a third-order critically damped linear system

- **Third order** - The order of a mathematical model refers to the number of independent variables; here the variables are initial pitch, velocity of pitch change, and acceleration of pitch change
- **Critically damped** - Damping describes how a system handles movement towards a target
    - overdamped - overshoots
    - **critically damped** - perfect aim
    - underdamped - undershoots
- **Linear system -** F0 production can be modelled as a linear process where the input is vocal fold tension and the output is the corresponding pitch (F0)
- **Test simulations** of F0 contours revealed that a third-order critically damped system effectively balanced sufficient complexity and accuracy with computational efficiency

A: Background Assumptions (ii) - Fig 3

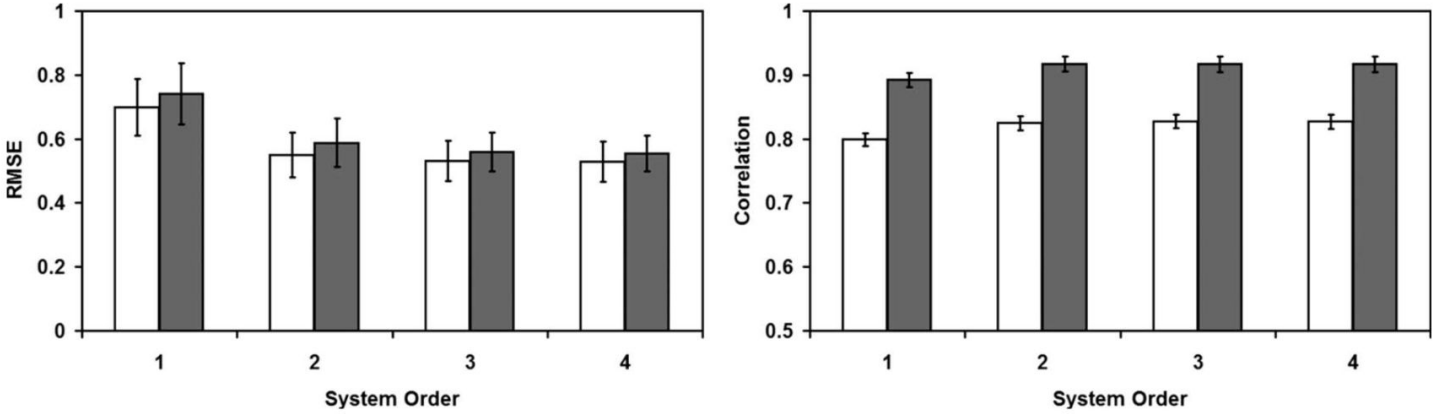


FIG. 3. Average RMSE (left) and correlation (right) of resynthesis results comparing between damping conditions and model order. White bars indicate results from the overdamped system and dark bars indicate results from the critically damped system. Vertical lines show standard errors of the mean.

A: Background Assumptions (iii)

2. Sequentiality and syllable synchronization

- **Syllable synchronization** - F0 movement towards an underlying pitch target is synchronized with the syllable, beginning at syllable onset and ending at syllable offset
- **Sequentiality** - Tone and intonation are sequential in articulation; all movements unidirectionally approach one target without any returns
- **Syllable-based modeling** has been shown to be **more accurate than accent-based models**

B: Model (i)

1. Pitch target

- Underlying goal of tone or intonation
- Modeled with the equation **x(t) = m(t) + b**, where ***m*** defines whether the target is static or dynamic, ***t*** represents time, and ***b*** represents the height of the pitch target.

2. F0 realization

- A third-order critically damped linear system models F0 transitions, where the **forced response** represents the pitch target itself, and the **natural response** accounts for transient effects during the transition to the target. The total F0 is a combination of these Components, influenced by initial F0 level, velocity, and acceleration.
- The model's parameters include ***m*** (slope) and ***b*** (intercept) for the pitch target, and ***λ*** (approximation rate).

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

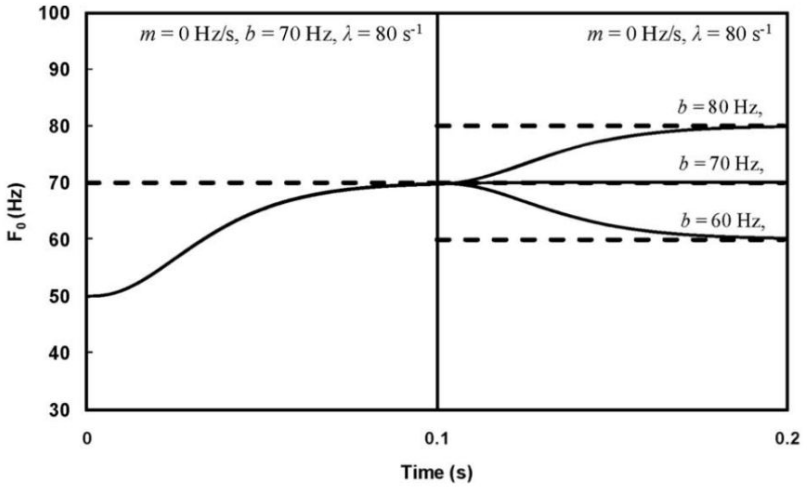
$$c_1 = f_0(0) - b,$$

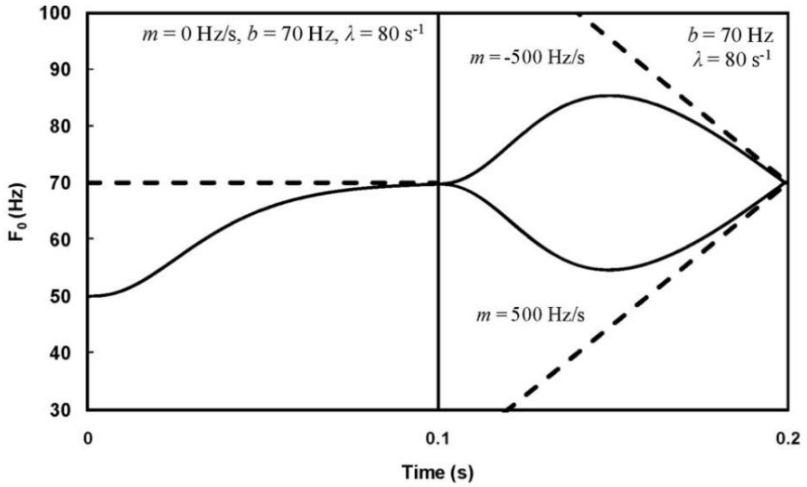
$$c_2 = f_0'(0) + c_1\lambda - m,$$

$$c_3 = (f_0''(0) + 2c_2\lambda - c_1\lambda^2)/2.$$

B: Model (ii): Fig 4 (example)

A





Automatic parameter extraction

- Parameter extraction was done with an automatic optimization-by-synthesis optimization algorithm (see Fig 5 and Table II below)

TABLE II. Constraint violation rates and relative changes in standard deviation when removing each parameter constraint in the automatic parameter extraction.

Constraint on	Constraint violation rate (%)	Relative change in standard deviation (%)		
		$\Delta\sigma_m$	$\Delta\sigma_b$	$\Delta\sigma_\lambda$
m	52.35	200.82	−0.67	81.22
b	17.15	−4.43	13.75	10.84
λ	7.21	1.74	−0.71	20.22

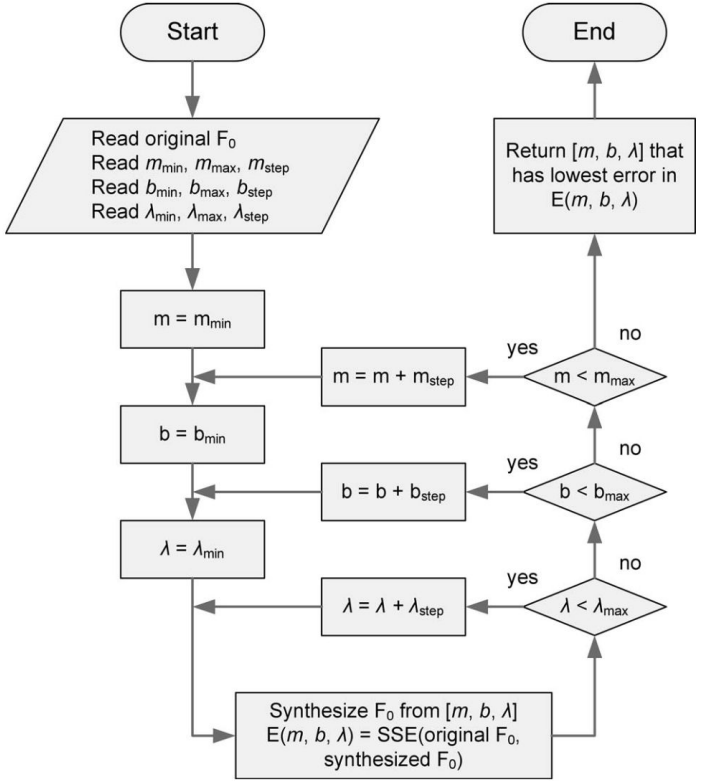


FIG. 5. A flowchart of automatic parameter extraction of qTA model. The algorithm optimizes for the suitable model parameters that, when implemented by the qTA model, generate F_0 contours that closely approximate the original ones.

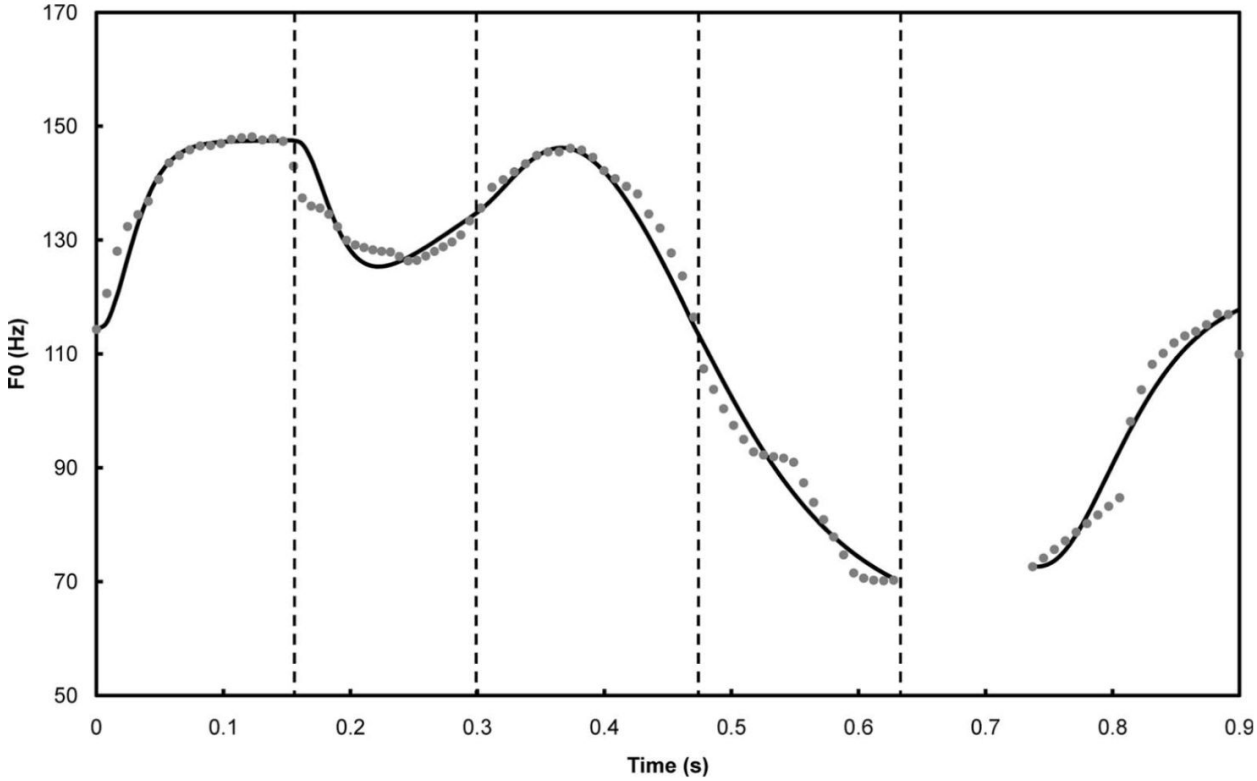


FIG. 7. An example of resynthesized F_0 for the tone sequence HRFLH. The solid line represents the synthesized F_0 contour while the gray dotted line indicates the original F_0 contour. The dashed vertical lines show the syllable boundaries. The discontinuity of F_0 at the beginning of the last syllable is due to the voiceless initial stop consonant [t] in the last syllable.

III.

Modeling Tone and Focus as Communicative Functions

Modeling communicative functions

How are communicative functions modeled in qTA?

- **Encoding in qTA** - Communicative functions are translated into parametric vectors for each syllable. These vectors define pitch targets, rates of approximation, and adjustments for stress and focus.
- **Additive combination** - Multiple functions are encoded in parallel. $p_j = \begin{bmatrix} m_j \\ b_j \\ \lambda_j \end{bmatrix}$
- The end result is a prosodic vector of the sentence that is a combination of the parametric vectors for each syllable.

$$s = \{p_1, p_2, \ldots, p_N\}$$

A. Lexical tone and lexical stress

Lexical tone in Mandarin

- Each tone is generated by a parametric vector generated by a tone function
  - (x: neutral (N), high (H), rising (R), low (L), falling (F))

$$p = \text{tone}(x)$$

- The parameters for a tone are derived by averaging vectors extracted from all occurrences of that tone in training data

Lexical stress in English

- Pitch values related to stress are assigned postlexically in English
- All syllables are assigned specific pitch targets, whether stressed or unstressed
- Stress is represented by the following function where x is either stressed or unstressed

$$p = \text{stress}(x)$$

B. Focus

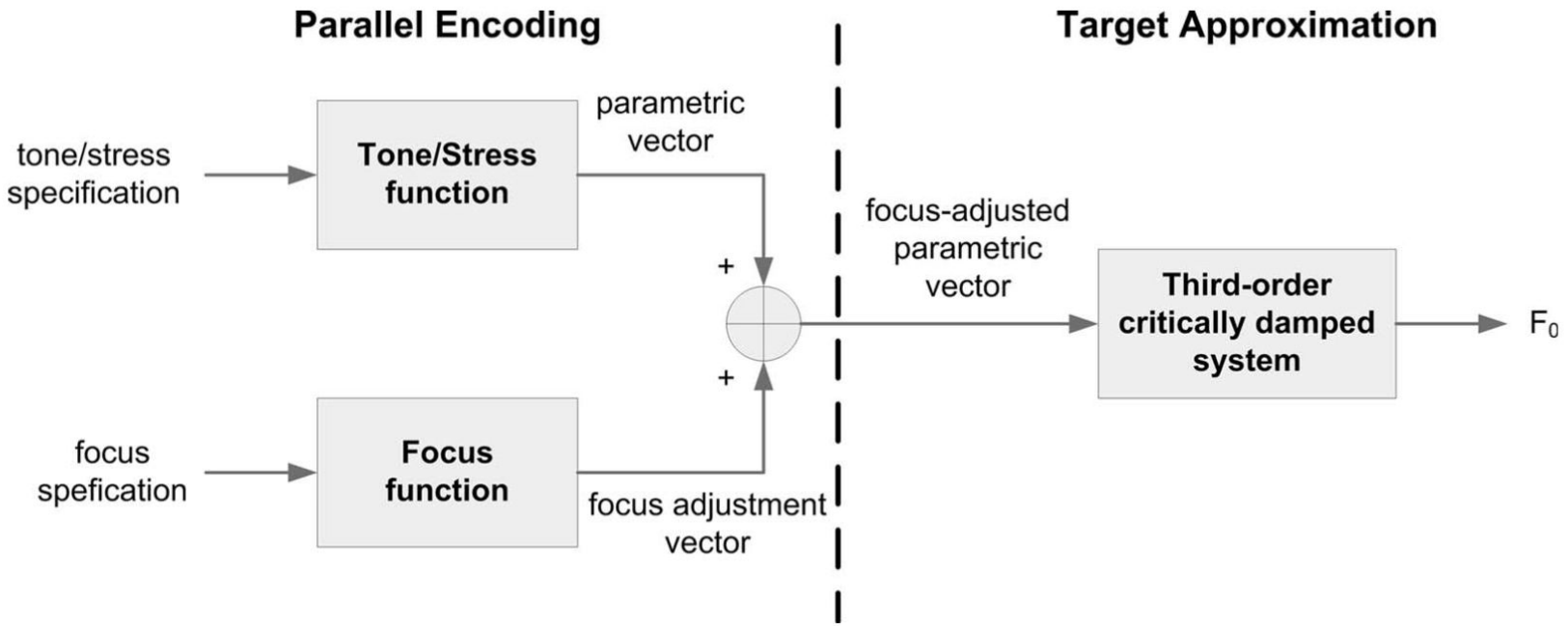
- Focus is a discourse function to highlight a particular piece of information; in Mandarin and English, focus expands the pitch range of the focused syllables and compresses the pitch range of the post focused syllables.
- Syllables in each sentence can be mapped into four regions: prefocus, on focus, postfocus, and final focus
- Computationally, it's an adjustment function which maps the given prosodic vector s with length N and focus point K to the output prosodic vector

$$\hat{\mathbf{s}} = \text{focus}(\mathbf{s}, \mathbf{K})$$

$$\hat{\mathbf{s}} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M + \Delta\mathbf{p}_{\text{on}}, \mathbf{p}_{M+1} + \Delta\mathbf{p}_{\text{post}}, \ldots, \mathbf{p}_N + \Delta\mathbf{p}_{\text{post}}\},$$

- The focus-adjusted parametric vector is the output of the encoding process and serves as the input to the TA model

Fig 8: qTA as the quantitative counterpart of PENTA



IV.

Experimental Evaluation

A. Corpora

Mandarin dataset

- 3840 five-syllable utterances
- 4 female speakers, 4 male speakers
- First two and last two syllables are disyllabic; third syllable monosyllabic

Word 1			Word 2			Word 3		
HH	mao1 mi1	"kitty"	H	mo1	"touches"	HH	mao1 mi1	"kitty"
HR	mao1 mi2	"cat-fan"	R	na2	"takes"	LH	ma3 dao1	"sabre"
HL	mao1 mi3	"cat-rice"	F	mai4	"sells"			
HF	mao1 mi4	"cat-honey"						

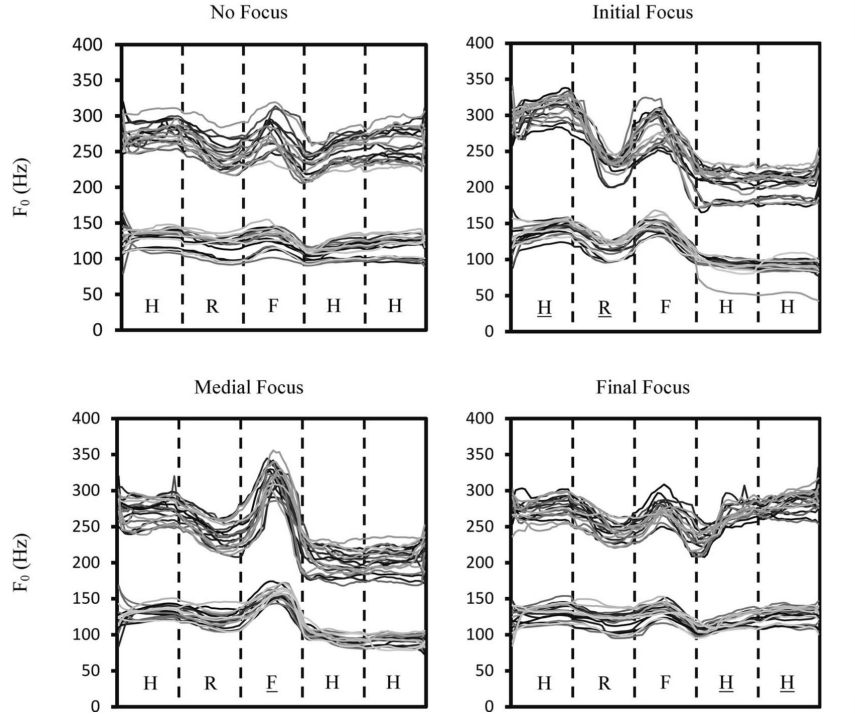
English dataset

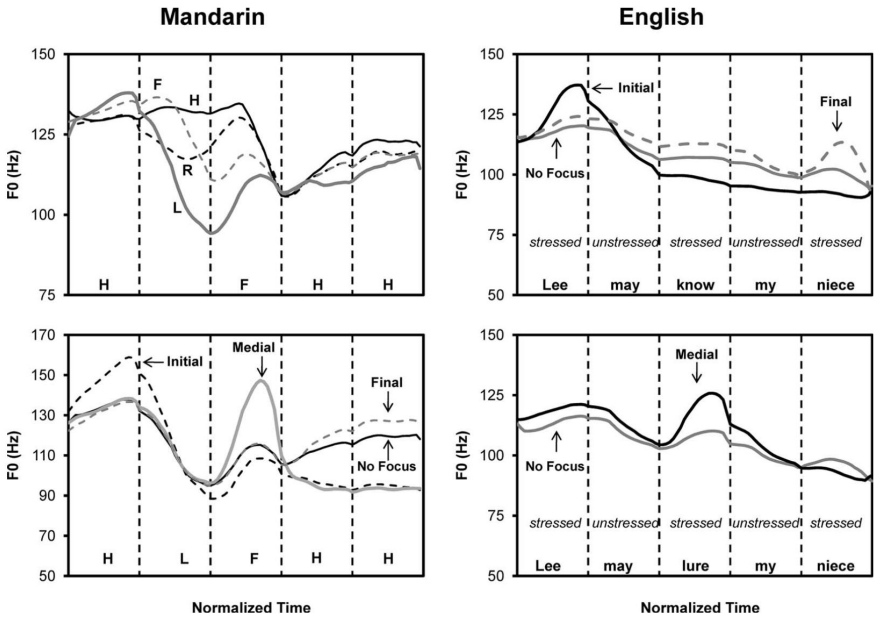
- 1176 short declarative utterances
- 4 female speakers, 4 male speakers
- Each sentence said in: no focus, sentence-initial focus, sentence-medial focus, and sentence-final focus, as well as with both speeds: normal and fast

TABLE IV. A list of sentences in the English dataset (Xu and Xu, 2005).

Word 1	Word 2	Word 3	Word 4	Word 5
Lee/Nina/Lamar/ Emily/Ramona	May	Know	My	Niece
Lee		Lure/mimic/ minimize		Niece
Lee		Know		Niece/nanny/ mummy

A. Corpora Figs 9 and 10





B. Parameter analysis (i)

- **Prosodic vectors** were extracted from each sentence in both datasets using analysis by synthetic optimization algorithm

TABLE V. Means and confidence intervals of extracted parametric vectors of the tone function obtained from the Mandarin dataset. Because different speakers have different average F_0, the utterance-onset F_0 is subtracted from b, the pitch target height, so that the values of b in the table are relative to the utterance-onset F_0.

Tone	m (st/s)	b (st)	λ (s^{-1})
H	0	0.0 ± 1.0	54.5 ± 5.4
R	93.4 ± 3.4	-2.2 ± 1.1	40.7 ± 3.8
L	0	-8.9 ± 0.6	34.1 ± 5.0
F	-106.4 ± 3.0	-2.5 ± 1.3	39.3 ± 3.2

TABLE VII. Means and confidence intervals of parametric vectors of the stress function obtained from the English dataset.

Syllable position	Stress	m (st/s)	b (st)	λ (s^{-1})
1	Unstressed	0	0.1 ± 0.8	72.8 ± 6.7
	Stressed	0	1.3 ± 0.9	48.1 ± 9.1
2	Unstressed	0	-1.2 ± 0.9	41.0 ± 5.9
3	Unstressed	0	-2.0 ± 0.7	51.4 ± 7.2
	Stressed	0	-1.2 ± 0.6	62.6 ± 11.8
4	Unstressed	0	-2.8 ± 0.6	43.3 ± 6.7
5	Unstressed	0	-5.3 ± 1.2	58.3 ± 9.9
	Stressed	0	-1.3 ± 1.2	49.8 ± 12.0

B. Parameter analysis (ii)

TABLE VI. Means and confidence intervals of focus adjustment vectors of on-focus, postfocus, and final-focus regions obtained from the Mandarin dataset. These focus adjustment vectors are relative to the parametric vector of the tone function in Table V.

Focus location	Tone	Δm (st/s)	Δb (st)	$\Delta\lambda$ (s^{-1})
On focus	H	0	2.3 ± 1.1	-1.6 ± 3.5
	R	11.8 ± 3.6	0.6 ± 1.0	-3.9 ± 3.7
	L	0	-2.4 ± 1.8	1.0 ± 6.7
	F	-6.7 ± 2.5	1.2 ± 1.8	-1.2 ± 3.2
Postfocus	H	0	-5.6 ± 1.0	-11.2 ± 4.2
	R	-7.3 ± 2.9	-4.1 ± 0.7	7.8 ± 7.0
	L	0	-4.1 ± 1.4	-3.2 ± 5.4
	F	4.5 ± 3.3	-2.8 ± 1.5	2.5 ± 5.3
Final focus	H	0	-0.2 ± 0.8	-16.0 ± 2.5
	L	0	-2.1 ± 1.5	-4.6 ± 5.4

TABLE VIII. Means and confidence intervals of adjustment vectors of the focus function obtained from the English dataset. They are derived relative to the parametric vectors in Table VII.

Focus location	Stress	Δm (st/s)	Δb (st)	$\Delta\lambda$ (s^{-1})
On focus	Unstressed	0	-1.1 ± 0.7	-5.2 ± 3.7
	Stressed	0	2.9 ± 1.1	-10.6 ± 2.4
Postfocus	Unstressed	0	-1.9 ± 0.9	14.3 ± 5.0
	Stressed	0	-2.9 ± 1.0	1.4 ± 11.1
Final focus	Unstressed	0	-1.7 ± 2.7	0.9 ± 13.5
	Stressed	0	2.3 ± 1.6	-24.0 ± 8.7

B. Parameter analysis (iii)

TABLE IX. Means and confidence intervals of parametric vectors of the stress function and the adjustment vectors of the focus function for the exceptional cases where the pitch target is dynamic. These parametric vectors are also derived from the English corpus. The symbol Δ indicates that the parameters in that row are relative to the prefocus parameters.

Stress	Focus	Δm (st/s)	Δb (st)	$\Delta\lambda$ (s^{-1})
Focused word-final stressed syllable	On focus (Δ)	-81.1 ± 14.8	3.5 ± 1.4	-20.5 ± 1.3
Sentence-final monosyllabic word	Prefocus	-90.2 ± 11.8	-4.3 ± 1.5	38.4 ± 7.9
	Postfocus (Δ)	27.1 ± 14.6	-1.0 ± 1.9	-4.6 ± 5.6
	Final focus (Δ)	-22.8 ± 9.7	0.3 ± 1.6	-11.6 ± 2.7

C. Model evaluation by assessing synthesis quality (i)

- **Effectiveness of the qTA model was evaluated in two ways:**
  - (a) **numerical assessment** of closeness of fit between synth and natural F0 and
  - (b) **perceptual identification** of tone and focus as well as judgement of naturalness by native speakers

1. Numerical assessment

- Model tested using leave-one-out cross validation, repeated 8 times to assess inter-speaker variability and detect the worst-case errors
- Two metrics were used to compare synthesized and natural F0 contours
  - Root Mean Square Error
  - Pearson’s correlation coefficient

C. Model evaluation by assessing synthesis quality (ii)

2. Perceptual identification

Mandarin

- 12 utterances were synthesized with varying tones for the second and third syllables, as well as with varying focus conditions
- Listeners identified tones of the second and third syllables and the focused word
- Listeners were asked to judge whether the sentences were natural or synthesized

English

- 40 test utterances, 8 natural and 32 synthesized across duration and focus conditions
- 4 female speakers, 4 male speakers
- Each sentence said in: no focus, sentence-initial focus, sentence-medial focus, and sentence-final focus, as well as with both speeds: normal and fast

C. Model evaluation by assessing synthesis quality (iii)

3. Evaluation results: Numerical (Mandarin)

Focus is key!

TABLE X. Average RMSE and correlation coefficients in the Mandarin simulations. Synthesized F_0 used in these comparisons are those generated by resynthesis, tone function, focus function, and positional effect.

Imposed function	No. of parametric vector	RMSE (st)	Correlation
Resynthesis	19 200	0.56	0.92
Tone	4	2.84	0.74
Tone+position	20	2.46	0.75
Tone+focus	16	2.24	0.77
Tone+focus+position	80	2.16	0.78

TABLE XI. Average RMSE and correlation coefficients for adding context dependency in the simulation of the Mandarin dataset.

Imposed function	No. of parametric vector	RMSE (st)	Correlation
Tone	4	2.84	0.74
Tone+preceding context	16	2.57	0.76
Tone+following context	16	2.50	0.77
Tone+focus	16	2.24	0.77
Tone+focus+preceding context	68	2.17	0.78
Tone+focus+following context	68	2.21	0.77

C. Model evaluation by assessing synthesis quality (Fig 11)

3. Evaluation results: Perceptual (Mandarin)

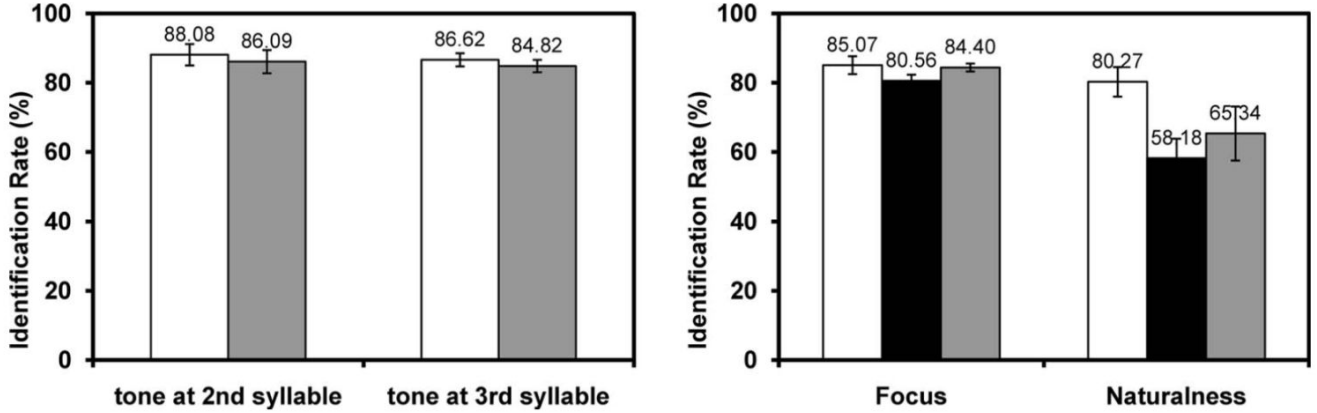


FIG. 11. Means (bars and the numbers above them) and standard errors (vertical lines) of identification rates in the Mandarin perceptual evaluation. The left graph shows averaged tone identification results while the right graph shows results of focus identification and naturalness evaluation. In the left graph, the white and gray bars indicate rate of tone identification for natural and synthetic F_0. In the right graph, the white, black, and gray bars indicate the results of focus identification and naturalness rating for natural F_0, synthetic F_0 without focus duration, and synthetic F_0 with focus duration, respectively.

C. Model evaluation by assessing synthesis qualit

3. Evaluation results

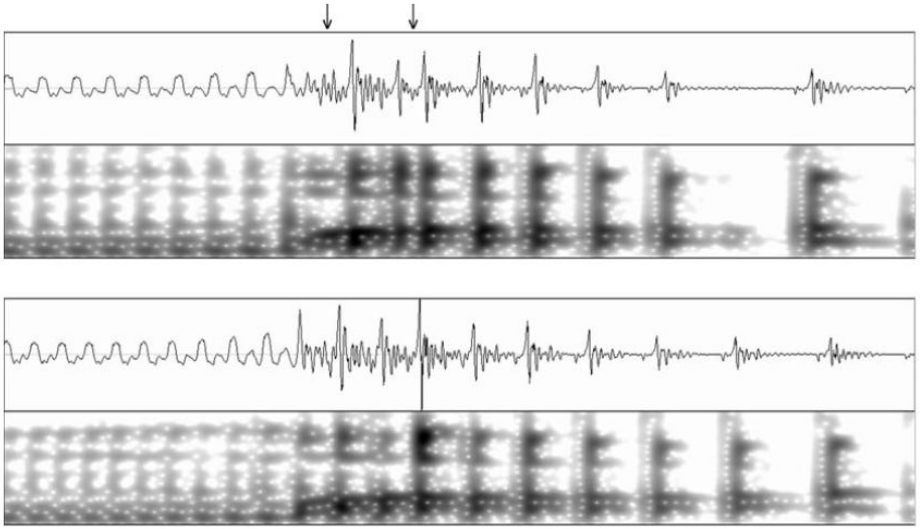


FIG. 12. Upper panel: Aperiodicity in a synthetic sentence. The arrows point to the two locations where the periods are either exceptionally long or exceptionally short. Lower panel: The resynthesized original sentence, where no strong aperiodicity is seen in the same locations.

C. Model evaluation by assessing synthesis quality (Fig 13)

3. Evaluation results: English

TABLE XII. Averaged RMSE and correlation coefficients in the simulation of English dataset.

Imposed function	No. of parametric vector	RMSE (st)	Correlation
Resynthesis	14 224	0.32	0.83
Stress	4	1.93	0.75
Stress+position	11	1.71	0.78
Stress+focus	12	1.68	0.77
Stress+focus+position	18	1.57	0.78

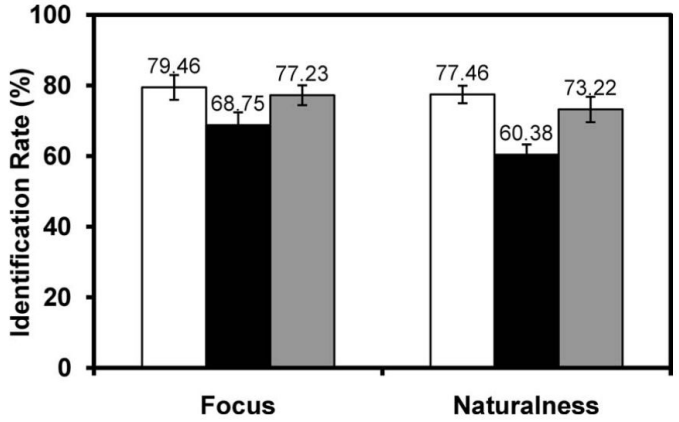


FIG. 13. Means (bars and the numbers above them) and standard errors (vertical lines) of focus identification rate and naturalness evaluation in the English perception tests. White, black, and gray bars correspond to sentences that are natural, synthetic without focus duration, and synthetic with focus duration, respectively. The vertical line in each bar indicates standard error.

C. Model evaluation by assessing synthesis quality (

3. Evaluation results

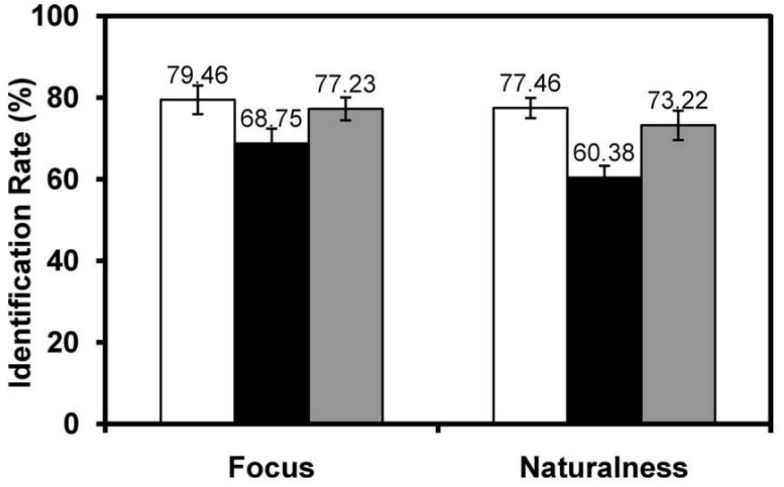


FIG. 13. Means (bars and the numbers above them) and standard errors (vertical lines) of focus identification rate and naturalness evaluation in the English perception tests. White, black, and gray bars correspond to sentences that are natural, synthetic without focus duration, and synthetic with focus duration, respectively. The vertical line in each bar indicates standard error.

V. Discussion

qTA as a quantification of PENTA

- The PENTA model represents tone and intonation as communicative functions encoded via a syllable-synchronized target approximation mechanism, using a third-order critically damped system with three key parameters to generate F0 contours.
- It ensures a biophysically plausible link between articulatory control and communicative functions.

Validation of the qTA model

- qTA can synthesize natural-sounding contours, supported by both numerical and perceptual evaluations.

V. Discussion

Applications to speech technology

- qTA's relative simplicity, (using just three parameters: slope, height, and rate of approximation) makes it easier to implement than more complex models.
- Articulatory constraints during parameter extraction reduce search space, aiding practical applications like intonation synthesis for speech processing systems.

Limitations and future directions

- qTA models only the communicative functions of tone, intonation, stress, and focus, but cannot synthesize emotion (for example)
- Not fully automated: It still relies on predefined parameter constraints.

VI. Conclusion

qTA is an effective tool for the theoretical study of tone and intonation and for generating F0 contours in automatic speech systems

Language	RMSE	Correlation coefficient
Mandarin resynthesis	0.56	0.92
Mandarin synthesis	0.32	0.83
English resynthesis	2.24	0.77
English synthesis	1.68	0.77

Thanks!
Let’s discuss!