Toward Effectively Simulating Post-low Bouncing in Mandarin Chinese by PENTAtrainer2: Theory-testing and Model Fine-tuning[i]

Abstract

Post-low bouncing is often treated as a tonal phenomenon in Mandarin Chinese. It broadly refers to the abrupt pitch rise in the neural tone after the low tone. Recent studies found it might be caused by an independent, possibly universal, articulatory mechanism due to perturbation of antagonistic forces maintained in the laryngeal muscles (balance-perturbation hypothesis). Based on such theory, Prom-on et al (2012) proposed a simulation mechanism by adjusting the acceleration at the initial state of the first neutral tone after the low tone. This paper aims to verify the efficiency of such bouncing mechanism and fine-tune the model for an optimal parameter set to synthesize post-low bouncing in PENTAtrainer2.

Key words: prosodic theories; post-low bouncing; speech synthesis; model calibration

Under the Supervision of Professor Yi Xu

Candidate Number: MZSZ9

Word Count: 9910

1. Introduction

Post-low bouncing generally refers to the phenomenon in Mandarin Chinese that after producing a low lexical tone (L), the F_0 of the next neutral tone (N) bounces up (Chao, 1968; Lin and Yan, 1980; Shen 1992; Shih, 1988). Subsequent studies found there might be an articulatory mechanism causing F_0 to rise after a low pitch, and it is more readily observed in the N tones immediately follow the L tone in Mandarin and Cantonese (Chen and Xu, 2006; Gu and Lee, 2009, Shen, 1994). The Target Approximation (TA) model proposed by Xu and Wang (2001), which addresses the issue of contextual prosodic variability, does not account for this phenomenon. TA model is a representation of successive tone production as asymptotically approaching the tonal targets. In such process, each syllable needs to start from the F_0 of the previous syllable offset, which well accounts for the carryover effect reported in Thai and Mandarin (Gandour et al.,1994; Xu, 1997, 1999). The computational implementation of the TA model, quantitative Target Approximation model (qTA model, Prom-on et al., 2009) simulates the assimilation by transferring the F_0 dynamic states of the proceeding tone to the next tone as its initial dynamic states. Hence the qTA model would always simulate a carryover lowering at where post-low bouncing is supposed to occur. Post-low bouncing as a possible independent articulatory mechanism from the TA mechanism, has been examined by acoustic analysis as being a temporal loss of balance in the antagonistic laryngeal control after producing a low pitch (Chen and Xu, 2006; Prom-on et al., 2012). Prom-on et al. (2012) simulated this bouncing effect by iteratively searching for the optimal acceleration at the initial state of the N tone after L in a modified qTAtrainer[1] (Xu and Prom-on, 2010-2021) and a general parameter set

[1] Hereafter used interchangeably with PENTAtrainer1. Note PENTAtrainer refers broadly to the computational model and PENTAtrainer2 refers specifically to the current version PENTAtrainer.

was developed for the boost. For the current study, the goal is to determine if it is better to pre-learn the optimal post-low acceleration boost parameter and keep it fixed in PENTAtrainer2. The study contains 6 parts. Section 2 is a general overview of existing prosodic theories, PENTA model, and its computational model PENTAtrainer2. It discusses how PENTA model and PENTAtrainer could account for prosodic variability and synthesize speech by learning from natural speech data. Section 3 reviews the possible underlying mechanism of post-low bouncing and the original study of Prom-on et al. (2012), which verifies the balance-perturbation theory and developed a general equation for post-low bouncing. Section 4 introduces the methodology of the experiment including the training corpus used and the evaluation methods concerned. Numeric reports are analyzed in detail in Section 5. Correlation analysis and visual inspection are reported and discussed in section 6. The final section is a conclusion of findings in this study.

2. Review of Prosodic Theories

2.1 Three-way Distinction of Current Prosodic Theories

The study of speech, specifically the study of prosody, meets obstacles when it comes to the issue of reference. As Xu (2011) points out, the lack of orthographic representation of the prosodic units causes disputes in terms of whether or not, or in what degree linguists should rely on prosodic units such as F_0 peaks/valleys, F_0 turning points or the size of the F_0 movements. Depending on the degree of awareness of such difficulty, Xu (2011) made a three-way distinction among the existing prosodic theories: linear versus superpositional, formal versus functional, and acoustic versus articulatory.

2.1.1 Linear versus Superpositional

Linear prosodic models assume the surface forms of F_0 events correspond to actual prosodic units. Examples of linear models are the British nuclear tone tradition (Crystal, 1969; O'Connor and Arnold, 1961; Palmer, 1922), the Autosegmental Metrical theory (AM theory, Ladd, 2008) or in the same vein, the Pierrehumbert Model (Pierrehumbert, 1980), and the IPO model ('t Hart el al., 1990). For the linear models, labels are assigned only to the local F_0 events with one prosodic unit at one temporal location. The linear events are considered as phonologically contrastive, be it nucleus (and their affinities) or pitch accents, while non-linear effects such as pitch range are not considered as phonologically meaningful. The superpositional models would assume that directly observed prosodic forms should be decomposed into layers of categories. Examples of superpositional models are the SFC model (Bailly and Holm, 2005) and Fujisaki model (Fujisaki, 1983). In these models they assume the surface prosodic contours are added together with layers of F_0 contours. For example, in the Fujisaki model, the global surface contours are added by the Accent command and the Phrase command.

2.1.2 Acoustic versus Articulatory

The distinction between acoustic and articulatory model is based on the source of the surface prosody. Acoustic models assume the source of surface form is direct acoustic manipulation while articulatory models assume the source as direct articulatory manipulation. In short, it is a division of whether articulatory mechanisms are incorporated in the model. For the acoustic models, they either use polynomial curve-fitting (Andruski, 2004; Liu et al., 2006) or non-polynomial curve-fitting (Pierrehumbert[2], 1981), which can generate reasonably good fitting for individual

[2] Pierrehumbert (1981) implements the AM model quantitatively by fitting between adjacent F_0 turning points, or assumed pitch accents, with linear and parabolic interpolations.

utterances in a case-by-case manner. While they need different parameterizations for fitting different utterances, henceforth *ad hoc curve fitting*. Generally, acoustic models are not tested for its predictive capability, with the exception of SFC model[3]. On the other hand, articulatory models take into account the articulatory mechanism. Models which are articulatory are the Fujisaki model (Fujisaki, 1983), the Stem-ML model (Kochanski and Shih, 2003) and the PENTA model (Prom-on et al., 2009; Xu, 2005; Xu and Prom-on, 2014). Take the Fujisaki model for example, the surface prosodic forms are generated by the linguistically meaningful commands; Accent command and Phrase command, which cause responses such as on/off ramps of step and pulses responses of a second-order linear system. The results from the two commands are then passed onto the articulatory mechanism: glottal oscillation mechanism to generate the surface F_0 contour. The PENTA model, which will be reviewed in section 2.2, assumes the F_0 contour as the realization of the laryngeal movements to either static or dynamic pitch targets.

2.1.3 Formal versus Functional

The division of formal and functional refers to if the model uses the prosodic forms or the communicative function to define the prosodic components. AM theory (Ladd, 2008), IPO model (‘t Hart el al., 1990) and the Pierrehumbert model (Pierrehumbert, 1980) are formal as they rely on system of phonological representation for the observed F_0 contours and utilized the system to compare phonologically distinctive F_0 patterns. According to Pierrehumbert (1980:59), there is no intention of offering a theory of such a system, but they merely proposed it for convenience of investigating prosodic

[3] Bailly and Holms (2005) uses a neural-network controlled contour generator in order to choose contours that fit the superpositional functional profile from a set of learned contours.

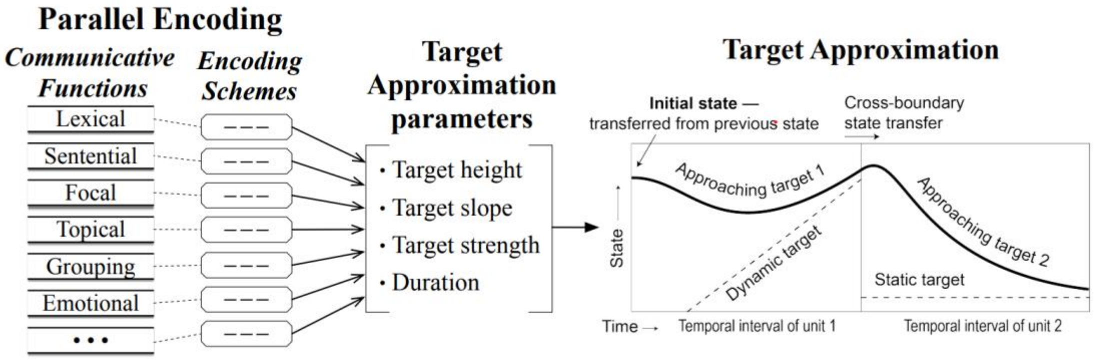
meaning. This is exactly the divergence from functional models, as formal models still define the phonological units by F_0 event, though there is awareness of functional meanings behind the appearance of F_0. Functional models such as SFC model (Bailly & Holm, 2005) takes metalinguistic functions into consideration. More specifically, it assumes the surface forms are composed by multiple underlying contours each carrying a metalinguistic function such as segmentation, hierarchization, emphasis and attitude. The PENTA model reviewed in the next section is also a functional model.

2.2 PENTA (Parallel Encoding and Target Approximation) Model

2.2.1 PENTA Model

PENTA model is, according to the three-way distinction of Xu (2011), a quasi-superpositional (quasi-linear), articulatory, functional model. The quasi-superpositional nature is because, for one thing, it assumes the observed F_0 is resulted from a set of parallel communicative functions and for another, PENTA assumes communicative functions are encoded and they modify the pitch target parameters. So, the F_0 contour is not controlled by simple linear addition of functions. Henceforth quasi-linear/quasi-superpositional. PENTA is also functional based because it uses communicative functions to define the prosodic units. Though all theories acknowledge the importance of communicative functions in speech prosody, the main departure of PENTA from form-based theories is that it assumes the prosodic units being defined in terms of communicative functions in parallel. Being also an articulatory model, at the core of PENTA's articulatory mechanism is *syllable synchronized sequential target approximation* (Xu, 2005). The mechanism treats speech as a string of syllables in linear sequences, in which each syllable has its vocalic, consonantal, and laryngeal targets to be synchronically approached within the domain of the syllable[4]. Target Approximation process (briefly mentioned in section 1) is shown in the rightmost box of Figure 1. Each syllable is assigned either a dynamic or a static target (Xu and Wang, 2001). The initial state of one syllable is transferred from the final state in the previous syllable. In Figure 1, the leftmost block contains the parallel communicative functions in speech that are encoded by the Encoding Schemes (the second left block). Target Approximation parameters in the third block from the left is a set of target parameters determined by the encoding system. The parameters then control the Target Approximation process in which each syllable will be assigned a target to generate the surface F_0 contour.

Figure 1: PENTA model (in Xu, 2005)



2.2.2 Basic Biophysical Mechanisms Behind PENTA

The articulatory mechanism of PENTA is based on the empirical evidence that the transitional movements between laryngeal states takes time and should be considered in any prosodic theory. Firstly, the time for F_0 movements is not negligible like previous theories assume, for time is needed in the transition of the laryngeal state (Sundberg,

[4] Except in the case of CV co-onset where the initial consonant approaches its target earlier than other target approximation (see the time structure model of the syllable in Xu & Liu, 2006).

1979; Xu & Sun, 2002). According to the experiment conducted by Xu and Sun (2002), the relation between the F_0 excursion and the minimum time needed is as follows:

$$\text{Rise: } t_r = 89.6 + 8.7d$$

$$\text{Fall: } t_f = 100.04 + 5.8d$$[5]

The finding shows that for any noticeable pitch movement to happen, it would take a minimum of 100ms. In this sense, a larger half of each syllable would be used to do the pitch transition, or in other words, to approach its assigned target. Hence, theories of prosody should take the transitional movement and the significant time it consumes into consideration.

Regarding the obligatory synchronization of tone and syllable, several pieces of evidence have exemplified that the range between the onset and offset of a syllable is where the implementation of pitch targets resides in (Xu, 1998[6]; Ohala & Roengpitya, 2002: p.2285; Xu et al., 2003[7]; Xu, 2004[8]; Xu & Wallace, 2004[9]). Yet there is another piece of biophysical evidence which is possibly the mechanism behind PENTA model in terms of the assumption about the synchronization of tone and syllable. Kelso (1984) and Schmidt et al. (1990) discover a deep-rooted tendency of human to coordinate motor movements. Further, in high-speed motor movement Kelso (1984) found the

[5] The duration *t* is measured in milliseconds, the size of F_0 excursion is measured by semitone.

[6] It was found in this study that the alignment of F_0 does not change on the condition if there is a nasal coda or not.

[7] In Xu et al., (2003) it has been reconfirmed that the local perturbation caused by voiceless consonants does not influence the approximation of tonal targets, with R tone following H being already low and F following H being quite high after the local perturbation. This also exemplifies the syllable is the host of the implementation of tonal targets.

[8] This study found the undershoot of pitch target is not ideal in conflicting tonal contexts, nevertheless it does happen because of the obligatoriness of synchrony of tone and syllable.

[9] It was found in Xu & Wallace (2004) that the F_0 alignments of Mandarin dissyllabic words with no nasal coda, short nasal coda and long nasal coda show virtually no difference, so the tone is synchronized with the entire syllable.

only option for the subjects was to keep the coordination at the most stable state, as it is unrealistic to maintain other phase relations between two movements other than keep them synchronized[10]. Analogously, in human speech, which often reaches maximum speed (Xu & Sun, 2002; Xu & Prom-on, 2019), it is largely possible to be driven by the same biophysical mechanism. The movement of laryngeal and supralaryngeal movements are controlled separately to compose speech as the melodic[11] and the segmental components[12]. The movements, following the biophysical tendency, should then be coordinated at the most stable state, as it is difficult for human due to the speed of natural speech, to maintain complex phase relation between the pitch realization and the syllable.

2.3 Two Types of Prosodic Variability and the Mandarin N Tone

There are two types of prosodic variability in natural speech. Contextual variations refer to the variation of a tone category under the influence of its adjacent tones. For example, each tone category of Mandarin Chinese, when proceeded by four tones, generates F_0 patterns with extensive variability (Liu & Xu, 2005; Xu 1997, see Figure 2), especially at the early portion of the syllable, but the F_0 trajectory would approach its tonal target gradually.

The second type of variability is non-contextual which often is found at a large scale of temporal domains. It is caused by many-factored modifications of speech functions including intonational, emotional and attitudinal ones. They differ from the contextual ones, which is made mandatory in the speech mechanism of target approximation (see

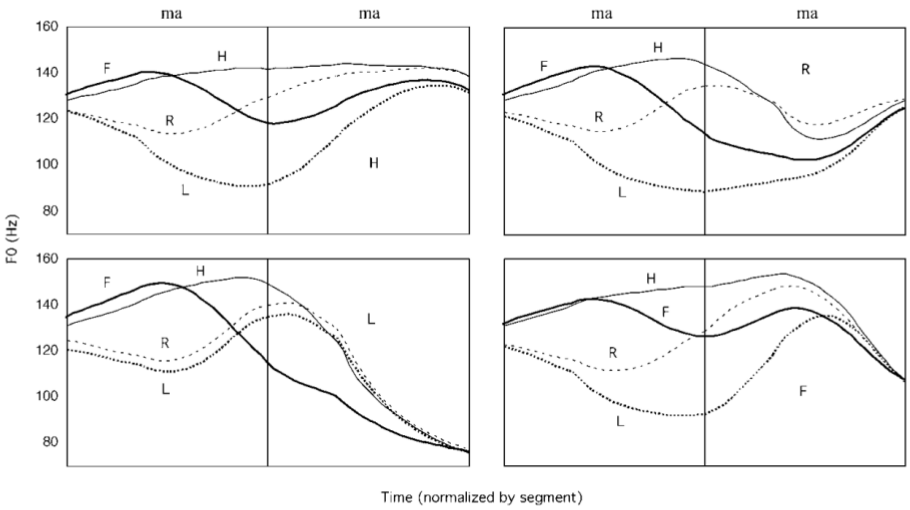
[10] In Kelso (1984) subjects were asked to perform a task of waging fingers with both hands in two manners, either simultaneously or one finger starts to perform earlier than the other by half cycle. They were repeated under slow and high speed. Results show under high-speed condition, subjects are only capable of maintaining the motion at a fully synchronized pattern.

[11] F_0 pattern.

[12] spectral pattern.

section 2.4) and varies under the influence of adjacent tones.

Figure 2: the effect of the proceeding tones on the following tones (in Xu, 1997)



Post-low bouncing is not non-contextual as there does not seem to be communicative functions driving it. However, it could be said as contextual as the phenomenon surfaces with an adjacent L tone, or more accurately, a low pitch, and the rising excursion of F_0 contour seems to vary with degrees of F_0 lowering (pursued in section 3). For the two kinds of prosodic variability, many of the existing models of speech prosody either do not separate the two (linear and acoustic models[13]) or cannot handle complex combinations of prosodic functions (Fujisaki model[14], Fujisaki et al., 1990). The PENTA framework on the other hand, assumes the necessity of identifying these two variabilities in its separate nature as it assumes one is made mandatory in speech and

[13] The IPO model (‘t Hart et al., 1990) and the AM, along the same line, Pierrehumbert model (Beckman & Pierrehumbert, 1986; Pierrehumbert, 1980; Ladd, 2008) are incapable of separating two types of variability. The IPO model assumes intonation as concatenated linear sections and ignore details of F_0 pattern. AM model treats intonation as a combination of pitch accents, phrase accents and boundary tones and between these events are filled with linear or curve interpolations. This way the separation of two types of variability is largely ignored.

[14] The Fujisaki model (Fujisaki et al., 1990) as described in section 2, models intonation as phrase and accent commands, is capable of synthesizing F_0 pattern of tonal variations and sentence type but is not tested for more complex situations.

the other should be functionally encoded.

There is also an issue of greater contextual variability shown in N tone in Mandarin Chinese. In phonological terms, N is assumed as the weak element that is unspecified or toneless, just like the English schwa. N tone shows greater surface variability in speech, especially dependent on the syllable proceeds it. Previous studies usually treat the variability as a piece of evidence for the lack of phonological specification, and N would vary under tone spreading (Yip, 1980) or tone interpolation (Shih, 1987). While under the framework of the TA model (Xu and Wang, 2001), the N tone also has its target to asymptotically approach like full tones in Mandarin. The difference is that the N tone is rather short and often half as long as the full tones (Lin and Yan, 1980). Recall in Xu and Sun (2002) that the maximum speed of pitch change articulatorily constricts the TA process, hence the target of N tone does not fully surface with limited time, showing more contextual variations. Detailed acoustic analysis (Chen and Xu, 2006) identified that the variability of N tone is largely due to the tone proceeds it. Further they have found firstly, the variability of N reduces over time and the F_0 direction and magnitude of the N tone is largely predictable according to the final state of the proceeding syllable. It was also found the F_0 direction of three consecutive N tones[15] would turn away from the influence of the proceeding full tone at the first N and show a gradual convergence to a mid-target. This supports the TA process and suggests the N tone in Mandarin Chinese might have a mid-target implemented with weak force. The weak implementation also leads to the key issue of the current study, post-low bouncing. Chen and Xu (2006) show F_0 of the N tone often is raised after a L. Considering the pitch gradually drops during the implementation of the previous L, the

[15] In Chen and Xu (2006), utterances with three N tones after four full tones are tested. In this way, more time is allowed for N tone to reach its target.

phenomenon is unpredicted by the TA model.

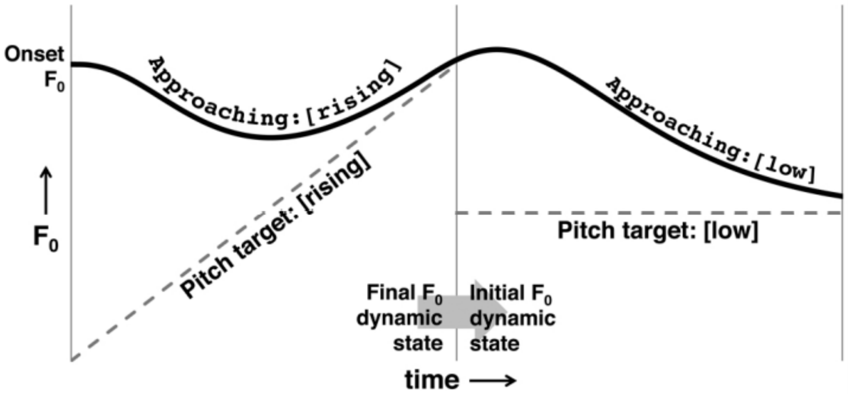
2.4 PENTA Model's Response in Addressing the Issue of Prosodic Variability

Section 2.2 reviews the core concept of PENTA model and the articulatory mechanisms behind it. In answering the issue of prosodic variability raised in 2.3, the mathematical model of the TA process and the Parallel Encoding scheme are reviewed in the next two subsections.

2.4.1 qTA Model

The qTA model (Prom-on et al., 2009) is a quantitative model based on the TA (Target Approximation) model (Xu and Wang, 2001). The illustration of the TA process is presented in the zoom-in of the rightmost column of Figure 1 (see Figure 3), with the dashed lines representing the pitch targets driving force to generate the surface F_0 contour. The final F_0 dynamic state is transferred to the next syllable as its initial F_0 dynamic state. The TA process is synchronized within each host syllable and is implemented sequentially syllable by syllable. Pitch targets in qTA is thus defined as a force driving F_0 movement in the form of a linear equation;

Figure 3: Target Approximation process (in Xu, 2005)



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

in which *m* and *b* represent the slope and height of the pitch target and *λ* stands for the strength of the F_0 movement forced by TA. The total F_0 response $f_0(t)$ is the combination of the pitch target *x(t)* and the natural response of the system in the form of,

$$\boldsymbol{f_0(t) = x(t) + (c_1 + c_2t + c_3t^2)e^{-\lambda t}} \qquad \textbf{(2)}$$

where c_1, c_2 and c_3 are the transient coefficients calculated from the final F_0 state of the previous syllable,

$$\boldsymbol{c_1 = f_0(0) – b} \qquad \textbf{(3)}$$

$$\boldsymbol{c_2 = f_0'(0) + c_1 \lambda – m} \qquad \textbf{(4)}$$

$$\boldsymbol{c_3 = (f_0''(0) + 2c_2 \lambda – c_1 \lambda^2)/2} \qquad \textbf{(5)}$$

The dynamic states transferred include the final F_0 position, the initial velocity of F_0 movement the initial acceleration of F_0 movement by taking $f_0(0)$ and its second and third derivatives. In this way, qTA could capture the F_0 movement by controlling the slope ***m***, the height ***b*** and the strength *λ* of target approximation**.** The algorithm then readily resolves the issue of the contextual variability as it captures the initial state and target of each syllable without needing to know the state of the proceeding syllable. In Prom-on et al. (2011), the variability issue of N tone has also been largely accounted for under this model except post-low bouncing. They successfully simulated N tone in Mandarin Chinese and exemplifies it is possibly a mid-level target (***b***) implemented with weak force (*λ*).

2.4.2 Parallel Encoding

Other than the TA model, another component of PENTA framework is Parallel Encoding. In the illustration of the PENTA framework shown in Figure 1, functions are encoded by the encoding schemes and then used to determine parameters for TA process, which renders the prosody of speech as a process of encoding communicative functions in the form of target approximation. Since PENTA model assumes pitch targets are modified by various communicative functions, the non-contextual variations can be automatically generated by realization of *syllable-synchronized target approximation* along with the contextual ones.

2.5 Computation realization of PENTA Model: PENTAtrainer2

Computational modeling of speech prosody can bring advance in prosody research because of its rigorous power in theory testing, and the rigor lies in the fact that computational models can generate continuous prosodic units containing fine details that are comparable to real speech. At the core of each computational model, quantitative algorithms need to be developed based on different assumptions of the underlying mechanism behind speech prosody. PENTAtrainer2 (Prom-on & Xu, 2013), is such a computational model based on the PENTA framework with qTA as the quantitative algorithm. In spirit of PENTA, PENTAtrainer2 is a data-driven system aiming to extract as few representations (functional combinations) as possible to generate many pitch variants by learning from real speech and achieve predictive synthesis. It is thus critically different from computational models that are based on acoustic prosodic theories, as those models pair many representations of surface F_0 contour to their phonetic representations which hardly could achieve predictive synthesis. PENTAtrainer2 on the other hand requires investigators to annotate the underlying functions of real speech data and it can then extract optimal parameters for each underlying functional combinations by stochastic learning. The parameters then control the TA process to synthesize speech.

2.5.1 Functional Annotation

The functional annotation of PENTAtrainer2 follows three strategies: (1) *layered functional annotation*, (2) *pseudo-hierarchical combination* and (3) *edge synchronization*. Strategy (1) specifies that each functional layer should be annotated independently, and the functional categories are named by the investigator. Strategy (2) refers to the boundaries of the first layer are projected to the layers below and they together form functional combinations. Functional combinations that are the same are treated as one functional representation to avoid redundancy. *Edge synchronization* then specifies each layer is synchronized with the smallest temporal domain. A detailed annotation is shown in section 4.2.

2.5.2 Analysis-by-synthesis

In the qTA algorithm developed by Prom-on et al. (2009) and PENTAtrainer1[16] (Xu & Prom-on, 2010-2021), the target parameters are learned through a local exhaustive search, syllable by syllable, for the optimal parameter sets that result in the lowest sum of square errors (SSE) between the input F_0 and the synthesized F_0. The learned parameters are then averaged across utterances of each functional category. By doing so, optimal parameter sets are extracted for all functional combinations that has been annotated. Due to the fact that the estimated parameter sets are not necessarily optimal for the functional categories and the λ (target strength) estimation is often not satisfactory (see a detailed discussion in Xu & Prom-on, 2014), PENTAtrainer2 is critically different from PENTAtrainer1 by replacing the local optimization by a stochastic global optimization (simulated annealing, Kirkpatrick et al., 1983) to directly estimate the optimal parameters of each functional combination from the entire corpus.

This way PENTAtrainer2 can also resolve the issue for weak prosodic functions such as English unstressed syllables and N tone in Mandarin. Recall that the issue of prosodic variability can be easily addressed by two schemes of PENTA, namely TA and parallel encoding. Thanks to the computational implementation of PENTA, the variability can then be synthesized by learning from real speech. However, the prosodic variations caused by post-low bouncing cannot be addressed by current version PENTAtrainer2.

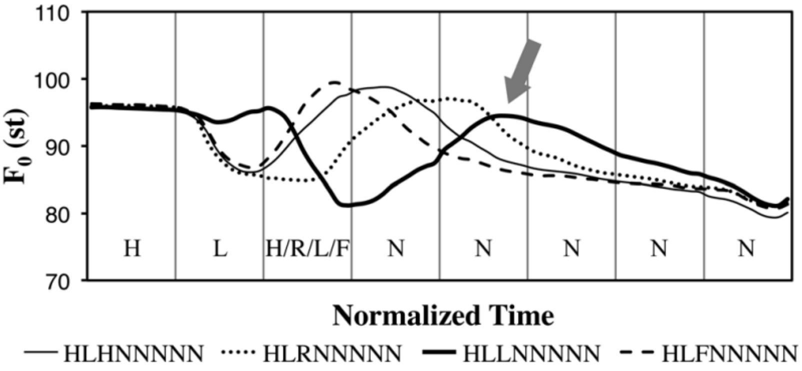
3. Post-low Bouncing: Phenomenon, Hypothesis, Theory-testing and Modeling

3.1 Phenomenon and Hypothesis

Post-low bouncing initially was described as an abrupt rising of F_0 in the N tone immediately after a L tone (Chao, 1968; Lin and Yan, 1980; Shen, 1992; Shih, 1988). Subsequent studies (Chen and Xu, 2006; Gu and Lee, 2009; Shen, 1994) found that the raising effect is not limited to N tone, only that it is more readily observed in N tone than full tones. This might be due to the strength of target approximation of N tones is not as strong as full tones, as it was found to take several consecutive N tones for it to reach its mid target (Chen and Xu, 2006; Liu and Xu, 2007). It was further discovered in those studies that the raising effect carries into the N-tone sequence after the L tone and the effect becomes stronger when the L is under focus. As shown in Figure 4, the F_0 contour starts to rise from a low point in the first N tone after the L and carries into the next several syllables where at the time of the third N the F_0 is higher than the

Figure 4: Time-normalized F_0 contour of utterances averaged across eight speakers with four full tones on the third syllable (in Prom-on et al., 2012, adapted from Liu

and Xu, 2007)



other three conditions. This phenomenon is exceptional as it does not fall under the scope of TA model (Xu and Wang, 2001). In the effort to explain such phenomenon, Yip (1980) treats it as one of the conditions where the L shows its full dipping contour (the other condition being when it is pronounced in isolation). However, it was reported that the effect is evident not only in L-N sequences but also in L-H sequences when L is under prosodic focus (Chen and Xu, 2006). Similar effect is also reported in English that after the L* accent, there is a constant time interval between the F_0 valley and the following peak (Pierrehumbert, 1980). This could lead one to question if the bouncing effect is a phenomenon only exists Mandarin Chinese and Cantonese, as chances are it might be caused by a universal mechanism when low pitch is involved. Along this line, Chen and Xu (2006) hypothesized that the bouncing effect might be due to a special articulatory mechanism, which is activated when F_0 reaches a certain threshold and is independent of the target approximation articulatory mechanism. The hypothesis is based on several pieces of evidence. Firstly, the extrinsic laryngeal muscles, particularly the infrahyoid muscles including the sternohyoid (SH), sternothyroid (ST) and thyrohyoid (TH) muscles, are contracted in producing F_0 below mid-level (speaker's mean F_0) (Atkinson, 1978; Erickson, 1993, 2011; Halle, 1994). The reason is, as reported in Honda el al. (1999), that the lowering of the larynx due to the contraction of the infrahyoid muscles[17] would lead to the shortening and relaxation[18] of the vocal folds for a low F_0 production. When the extra laryngeal muscles suddenly stop contracting after producing a low pitch, it might then cause a perturbation of balance and increases the tension of the vocal folds as the cricothyroids that lengthen the vocal folds are in contraction (Zemlin, 1988).

3.2 Theory-testing and Modeling

As reviewed in section 2.2, one of the principles of PENTA model is that it is based on the recognition of articulatory mechanism behind surface F_0 contour. In the spirit of such principle, efforts were made to identify the independent mechanism behind post-low bouncing and simulate the effect (Prom-on et al., 2012). Prom-on et al. have found acoustic evidence of the perturbation hypothesis and were able to simulate post-low bouncing by adding an acceleration on the initial state of the bouncing syllable.

Firstly, Prom-on et al. (2012) run a multi-factor correlation analysis to identify the most related F_0 dynamic features within the F_0 trigger group (prior to the bouncing event), F_0 event group (during the bouncing event) and between the two groups to seek out for possible causal relations. A set of evidence in support of the perturbation hypothesis were found. Within the F_0 event group, strong positive correlation was found between initial velocity (V_I) and the initial acceleration (A_I)[19], and, between A_I and peak rising acceleration (A_P)[20]. This indicates that there is no deceleration at the syllable onset and the driving force of bouncing starts at the post-low syllable onset. Most importantly,

[17] As reported in Atkinson (1978) and Shipp (1975), the infrahyoid muscles control the vertical position of the larynx.
[18] It was reported in Erickson (2011) that the activation of TH and SH muscle when producing a low F_0 rotates the cricothyroid joint in a way to release tension of the vocal fold.
[19] V_I and A_I refer to the initial rate of F0 change or F0 velocity change at the start of the N tone, respectively.
[20] A_p refers to maximum F0 rising acceleration during the bouncing event.

the peak rising acceleration time (T_{AP})[21], which shows an exponential distribution, achieves to its peak quickly and is smaller than the peak time (T_P)[22]. This further hints on the driving force triggered by the perturbation are transient and it is highly probable to be the only force involved in causing the effect. Between group test also shows strong correlation between F_0 lowering in L and the initial dynamic states of the post-low syllable. This further exemplifies the possible causal relation between F_0 lowering and the bouncing effect.

Having established possible causal relation between F_0 lowering and the bouncing effect, Prom-on et al. (2012) modeled the F_0 raising force when F_0 lowers to a certain threshold. Specifically, the driving force was simulated as a modification of velocity $\boldsymbol{v_s}$, acceleration $\boldsymbol{a_s}$, and both, on the initial dynamic states. The optimal value of extra velocity or acceleration were found by PENTAtrainer1 by iteratively search for values which results in the lowest SSE (Error Sum of Squares) during parameter estimation. The modification with both $\boldsymbol{a_s}$ and $\boldsymbol{v_s}$ shows the best result but only slightly outperforms the one with only $\boldsymbol{a_s}$. Further analysis shows during the implementation of parameter estimation only a small portion (3%) involves only $\boldsymbol{v_s}$ and overwhelmingly, a significantly large portion involves only $\boldsymbol{a_s}$ (55%). This leads to the final decision of using the acceleration only as the modification (see Table 1), which is reasonable as the extra force is only proportional to acceleration and theoretically should be represented

Table 1: Modification on F_0 dynamic states on the first post-low syllable in qTAtrainer

F_0 dynamic states	Initial qTA	qTA with bouncing effect
Height	$\mathbf{f_0(0)}$	$\mathbf{f_0(0)}$
Velocity	$\mathbf{f_0'(0)}$	$\mathbf{f_0'(0)}$

[21] Time calculated from syllable onset to the point F_0 reaches the maximum rising acceleration.

[22] Time calculated from syllable onset to the point F_0 reaches the maximum bouncing height.

Acceleration	**$f_0''(0)$**	**$f_0''(0) + a_s$**

that way in modeling. A general rule of the bouncing effect was then calculated by linear regression of the best-fit acceleration ***a_s***[23] as a function of the F$_0$ lowering (f$_L$),

$$a_s = \begin{cases} -412.82fl + 191.17, & fl \leq 0.46 \\ 0, & fl > 0.46 \end{cases} \qquad \textbf{(6)}^{24}$$

As can be seen from the equation, the threshold for bouncing is 0.46st, which is in line with the balance-perturbation hypothesis. In this study, PENTAtrainer2 will be used as the modeling tool. The general purposes, as illustrated in the introduction, are to test if post-low bouncing can be effectively simulated in PENTAtrainer2, and if the answer is yes, what are the optimal values for the bouncing slope, intercept, and threshold as in Equation (6). Further investigation is also made to examine if there is appropriate bouncing with the most optimal parameter set by correlation analysis and visual inspection.

4. Methodology

4.1 Corpus and Parameter Setting

The corpus used in this study was collected for studying N tones in question intonation in Liu and Xu (2007). It recorded 32 eight-syllable sentence structures (see Table 2) with at least three consecutive neutral tones at the 4th, 5th and 6th syllable from eight native Mandarin Chinese speakers (4 males and 4 females). Each sentence structure is repeated by each speaker five times. The total corpus contains 1280 utterances. Syllable 1 and 2 mean *he bought* and carries high and low tone respectively. Syllable 3-4 stand

Table 2：Sentence structure of the corpus, numbers represent five tones in Mandarin,

[23] They were obtained by iteratively search for the extra acceleration by qTAtrainer.

[24] Equation (5) in Prom-on et al. (2012).

0 stands for the neutral tone.

Focus	**Family**	**Boundary**	**Sentence Type**
1: on 2nd syllable	1: ma1 ma0	0: ends with le0 ma0	1: statement
	2: ye1 ye0		
2: on 3rd syllable	3: nai3 nai0	1: ends with mao1 mi1	2: question
	4: mei4 mei0		

Naming scheme indicated at the left of the colons

for *mother*, *grandfather*, *grandmother* and *sister* with the first syllable carrying four full tones of Mandarin and the second always carrying N. Syllable 5-6 also carries N tone, which literally is the equivalent of the English possessive s'. The last two syllables are either *goody* (le0 ma0) or *kitten* (mao1 mi1). The prosodic focus of the utterance falls either on syllable 2 or syllable 3. The naming scheme of the corpus is also shown in Table 2 with numbers marked on the left of the colons. For example, 11011 is the first repetition of the utterance "ta1 mai3 ma1 ma0 men0 de0 le0 ma0.", which is the statement *he bought mom's stuff* with focus on "mai3".

The fine-tuning of bouncing parameters follows the grid-search procedure where a set of candidate parameters are selected and evaluated. There are three candidate tuning parameters for the bouncing mechanism in PENTAtrainer2, namely, Threshold, Intercept and Slope. Parameter set that renders the best performance is to be fixed in PENTAtrainer2. The tuning parameters thus have 28 sets of conditions. The baseline condition is the general equation defined in Prom-on et al. (2012), that is to configure the parameters for bouncing slope, intercept, and threshold to -412.82, 191,19 and

0.46. For each parameter three candidates around the baseline condition are chosen (see Table 3) and all possible combinations formed by the three candidate tunning parameters are evaluated (3*3=27 conditions). They are named as C1-C27 in Table 4. The original condition is the one with bouncing mechanism is added is used for comparison. The evaluation of model performance follows the method of objective evaluation by measuring Root Mean Squared Error (RMSE) in semitones[25] (Prom-on et al., 2009; Raidt et al.,2004) and Pearson's correlation coefficient between the original utterance and synthesized utterance.

Table 3: Candidate parameters for Equation (6)

Threshold	**Intercept**	**Slope**
-2.5	100	-350
0	200	-450
2.0	300	-550

Table 4: Bouncing parameters for all conditions

Condition	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Threshold	-2.5	-2.5	-2.5	-2.5	-2.5	-2.5	-2.5	-2.5	-2.5	0
Intercept	100	100	100	200	200	200	300	300	300	100
Slope	-350	-450	-550	-350	-450	-550	-350	-450	-550	-350
Condition	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20
Threshold	0	0	0	0	0	0	0	0	2	2
Intercept	100	100	200	200	200	300	300	300	100	100
Slope	-450	-550	-350	-450	-550	-350	-450	-550	-350	-450
Condition	C21	C22	C23	C24	C25	C26	C27	Baseline	original	_
Threshold	2	2	2	2	2	2	2	0.46	1000	_
Intercept	100	200	200	200	300	300	300	191.17	NA	_
Slope	-550	-350	-450	-550	-350	-450	-550	-412.82	NA	_

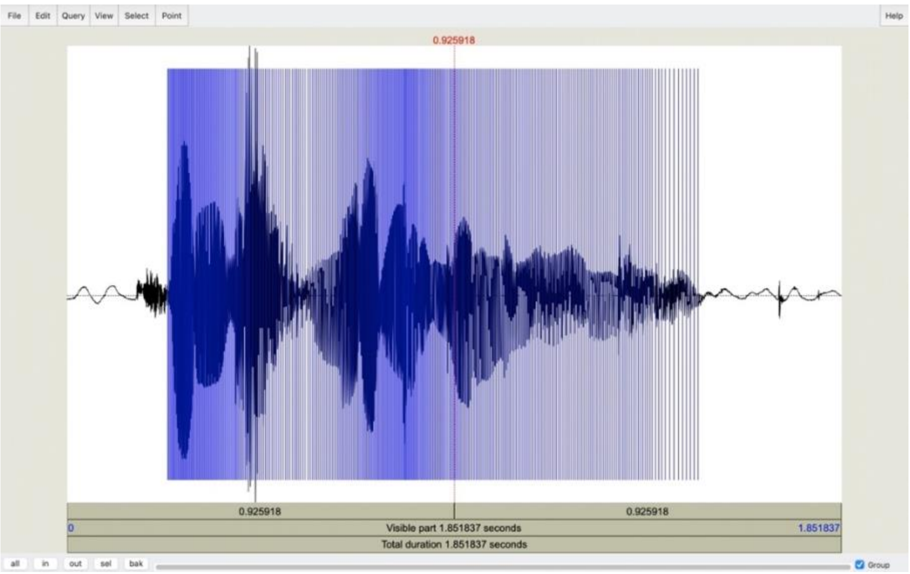
4.2 Workflow of PENTAtrainer2

The workflow of PENTAtrainer2, follows such steps: pulse marking, annotation, and

[25] See Xu (2011) for a detailed discussion explaining RMSE in Hz needs to be interpreted on the average mean F_0 of each speaker.

training. Figure 5 shows the spectral pattern in the form of waveform. On top of that, the blue vertical lines are pulses indicating where the speech signals repeat. The gap between each pulse marking refers to the pitch period that is inversely proportional to F_0. Pulse marking is used in PENTAtrainer2 for pitch estimation, instead of relying on the pitch contour in Praat, which often leads to error. In this study, the pulse has already been marked and rectified by ProsodyPro (Xu, 2013).

Figure 5: Pulse marking window in ProsodyPro (Xu, 2013)

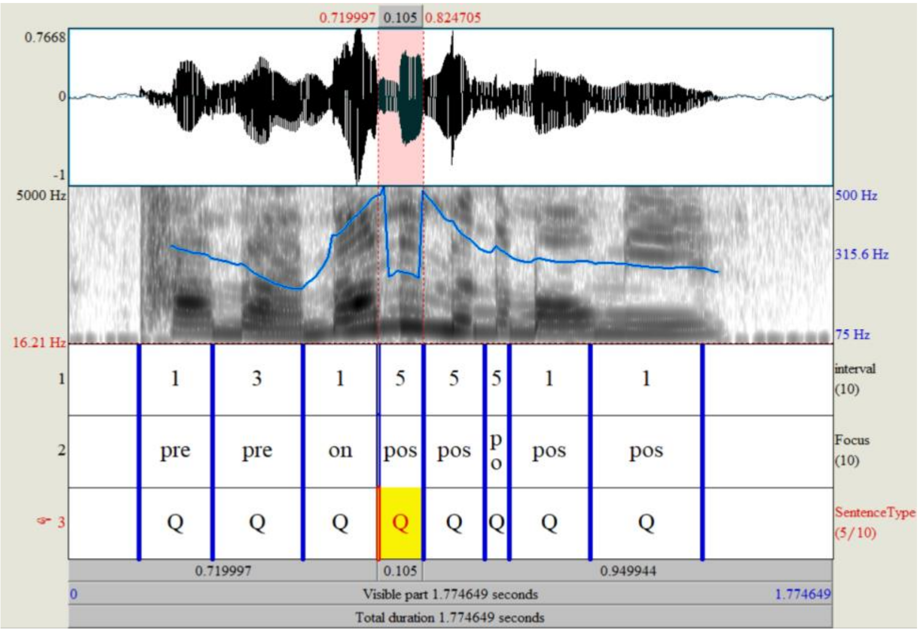


Annotation is the process consists of adding boundaries, functional layers, and labeling. The boundaries in the first tier were already added. The only adjustment was to move the original marked left boundary of the first syllable, indicating the onset of voicing of the vowel, to the initial perturbation of F_0, which is the syllable co-onset of consonant and vowel[26]. Then the functional layers: Focus and Sentence Type were added into PENTAtrainer2 aligning with the boundaries of the first tier. A short Praat script then could automatically fill the intervals with corresponding functions according to their names. In the annotation, four full tones of Mandarin are represented by numbers 1-4.

[26] See Xu and Liu (2006) for a full discussion of the time structure model.

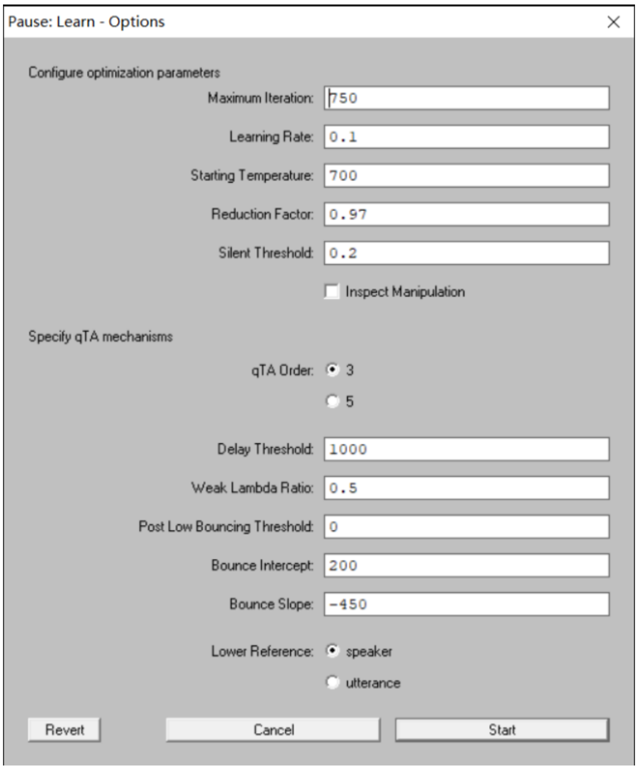
0 stands for the neutral tone. One thing to note is that in all utterances which involve L on the 3rd syllable, the L tone forms L-L tone sandhi with the 2nd syllable *mai3*. Therefore the 2nd syllable needs to be annotated as LS (low sandhi) rather than 3. Else it will result in low synthetic accuracy in these two syllables, as PENTAtrainer2 would not differentiate the two combinations unless explicitly annotated. Pre, on, post, indicate if the syllable is pre-focus, on-focus, or post-focus. Q or S is short for Question or Statement as in sentence modality. The names of the function carry no meaning other than differentiating categories in PENTAtrainer2 for parameter extraction. A full annotation of an utterance is shown in Figure 6 where there are five unique functional combinations.

Figure 6: Full annotation



After the annotation is complete, parameters for each condition were configured in the operation window of task-learn in PENTAtrainer2 and then trained. Figure 7 shows the configuration of one condition. For the original condition, the bouncing threshold was set to 1000. Each condition was trained once for each speaker[27]. The results are shown in the next section.

Figure 7: Parameter configuration in PENTAtranier2-learn-options



Pause: Learn - Options

Configure optimization parameters

Maximum Iteration: 750

Learning Rate: 0.1

Starting Temperature: 700

Reduction Factor: 0.97

Silent Threshold: 0.2

□ Inspect Manipulation

Specify qTA mechanisms

qTA Order: ● 3 ○ 5

Delay Threshold: 1000

Weak Lambda Ratio: 0.5

Post Low Bouncing Threshold: 0

Bounce Intercept: 200

Bounce Slope: -450

Lower Reference: ● speaker ○ utterance

5. Accuracy of Synthesis: Numeric Results

The RMSE and Correlation between the synthesized utterances and the original utterances are calculated and generated by PENTAtrainer2 automatically. For each condition, accuracy terms are averaged across the entire corpus (1280 utterances) and the utterances with L on the third syllable (320 utterances, hereafter L set). The task is two-fold, first to show if Equation (6) is adequate to simulate the bouncing effect. This can be verified either by directly measuring the performance of the baseline condition or measuring the average performance of all parameter sets excluding the original condition (hereafter collectively referred to as “bouncing conditions”[28]), as

[27] This is due to post-low bouncing is related to speaker’s mean F_0. Running the entire corpus altogether for each condition would not be meaningful.

[28] Including the baseline condition

the candidates are chosen based on Equation (6). The second task is trying to find the optimal parameter set and to determine if it can then be fixed in PENTAtrainer2. This can be done by looking directly for the condition that result in the lowest RMSE and highest correlation for (1) the whole utterance and (2) the three post-low N tones. The accuracy of synthetic N tones is evaluated by RMSE and correlation averaged across three N-tone syllables. Once the best parameter set is located, simulated F_0 can be analyzed by either checking the time-normalized F_0 or inspecting the synthesized contour, which is shown in section 6.

5.1 Accuracy Based on Full Corpus[29]

The accuracy of the whole utterance measured by RMSE and correlation is visualized in the line plots (Figure 8 and 9). Immediately can be spotted from the plots is firstly, C1 is an outlier with higher RMSE and lower correlation (RMSE=3.993, Corr=0.823). Hence the average performance of the bouncing conditions (RMSE=2,277, Corr=0.888) is calculated excluding C1. Results of the bouncing conditions do not vary much around the mean accuracy (SD_{RMSE}=0.119, SD_{Corr}=0.018). Secondly, C26 (Threshold=2.0, Intercept=300, Slope=-450) seems to realize the most optimal synthesis accuracy in terms of utterance-specific RMSE and correlation (RMSE=2.075, Corr=0.914). The baseline condition achieves above-average results (RMSE=2.212, Corr=0.896). The original condition without bouncing slightly underperforms the average accuracy (RMSE=2.305, Corr=0.885).

[29] For comparison of performance of all conditions measured on the full corpus, refer to the line plots. For comparison of accuracy measured on the full corpus between the bouncing conditions, the baseline condition, the original condition, and condition with the best performance, see Table 5 for quick reference.

Figure 8: Utterance-specific RMSE (averaged across the entire corpus)

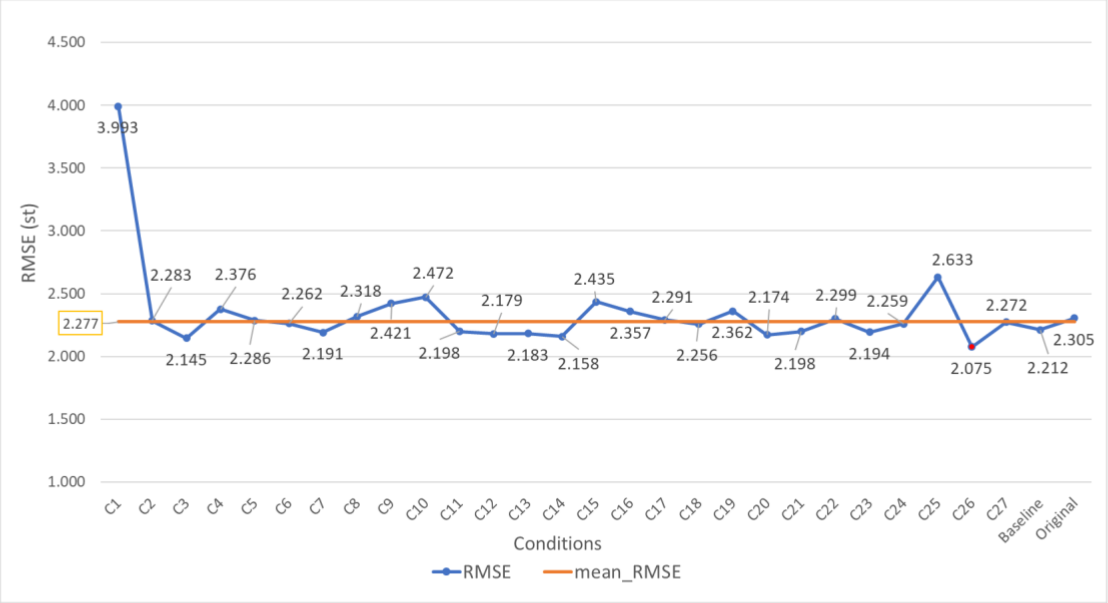
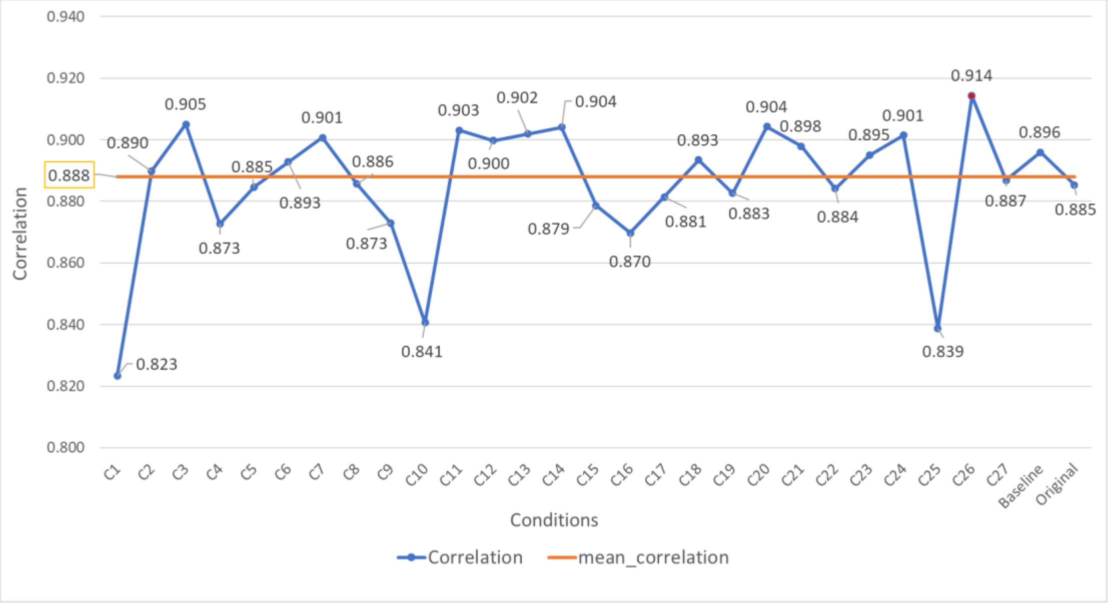


Figure 9: Utterance-specific Correlation (averaged across the entire corpus)



Similar pattern could be found in the accuracy reports of the three N tones (see Figure 10 and 11). Again, C1 is an outlier of all bouncing conditions (RMSE=3.537, Corr=0.670). The bouncing conditions excluding C1 achieve an average RMSE of 1.982 (SD=0.134) and correlation of 0.721 (SD=0.043) with small dispersion. The baseline condition performs above-average (RMSE=1.911, Corr= 0.742), and the original condition underperforms the average only in terms of RMSE (RMSE=2.036, Corr=0.725). C26 still yields the highest synthesis accuracy (RMSE=1.738, Corr=0.786).

Figure 10: Syllable-specific RMSE for N tones (averaged across the entire corpus)

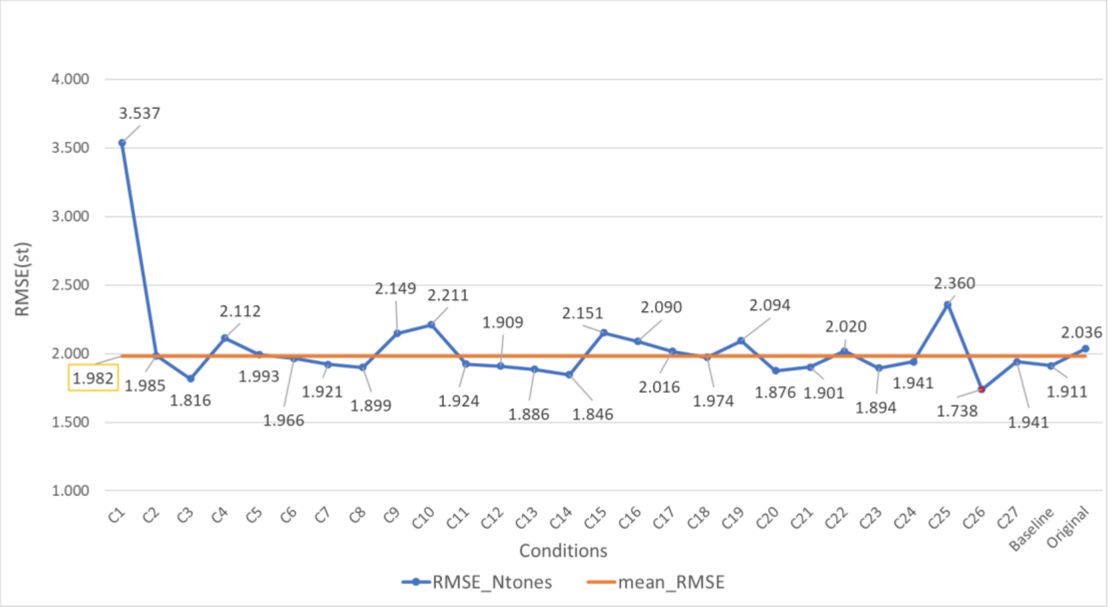
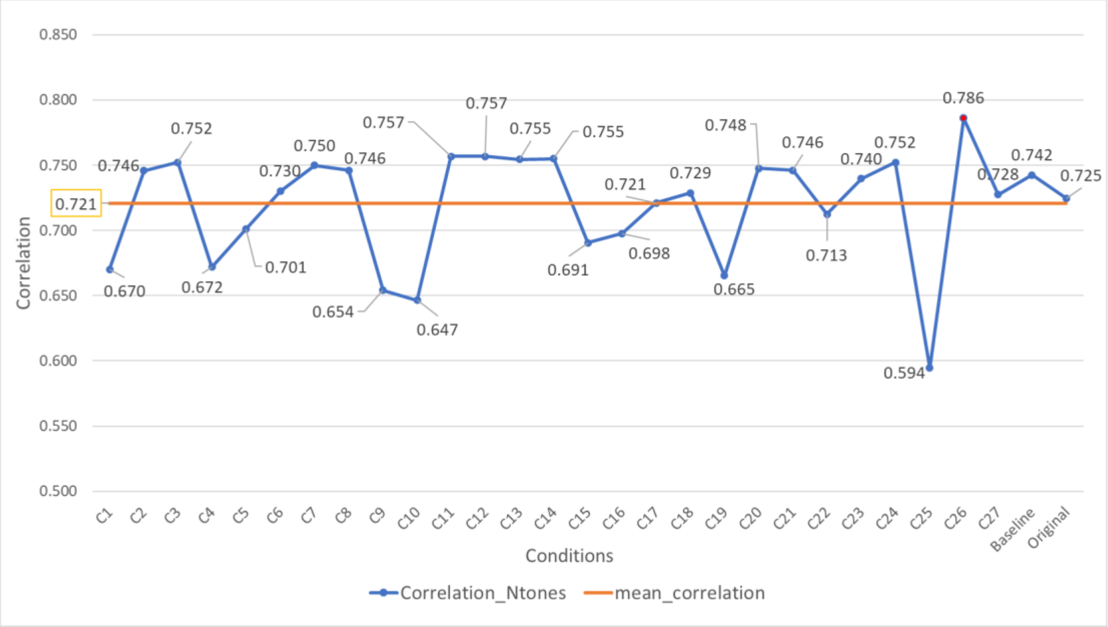


Figure 11: Syllable-specific Correlation for N tones (averaged acoross the entire corpus)



According to the utterance and syllable specific accuracy based on the entire corpus, it seems reasonable to conclude that firstly, the efficiency of the baseline condition is largely verified and secondly, original condition seems to underperform the average accuracy by a marginal difference. Also, C26 gets the overall best performance in synthesizing. It should be noted that though accuracy measured on the entire corpus is key in deciding the optimal condition, it is not yet deterministic. Due to simulated annealing, the synthesis accuracy could vary per training cycle, which introduces randomness. It is still likely that C26 has better performances in synthesizing utterances containing the other three full tones on the third syllable, which leads to an overall higher accuracy, especially when the accuracy of all bouncing conditions[30] does not show large dispersion. Evaluation based on L set could further check the performance where post-low bouncing is concerned.

Table 5: Accuracy terms averaged across the entire corpus by conditions

Conditions/Accuracy	Utterance-specific Accuracy		Ntones Accuracy	
-	RMSE(st)	Correlation	RMSE(st)	Correlation
Bouncing Conditions	2.277$_{(SD=0.119)}$	0.888$_{(SD=0.018)}$	1.982$_{(SD=0.134)}$	0.721*$_{(SD=0.043)}$
Baseline Condition	2.212	0.896	1.911	0.742
Original Condition	2.305*	0.885*	2.036*	0.725
C26	**2.075**	**0.914**	**1.738**	**0.786**

*Bold for best performances. * for the least ideal performances*

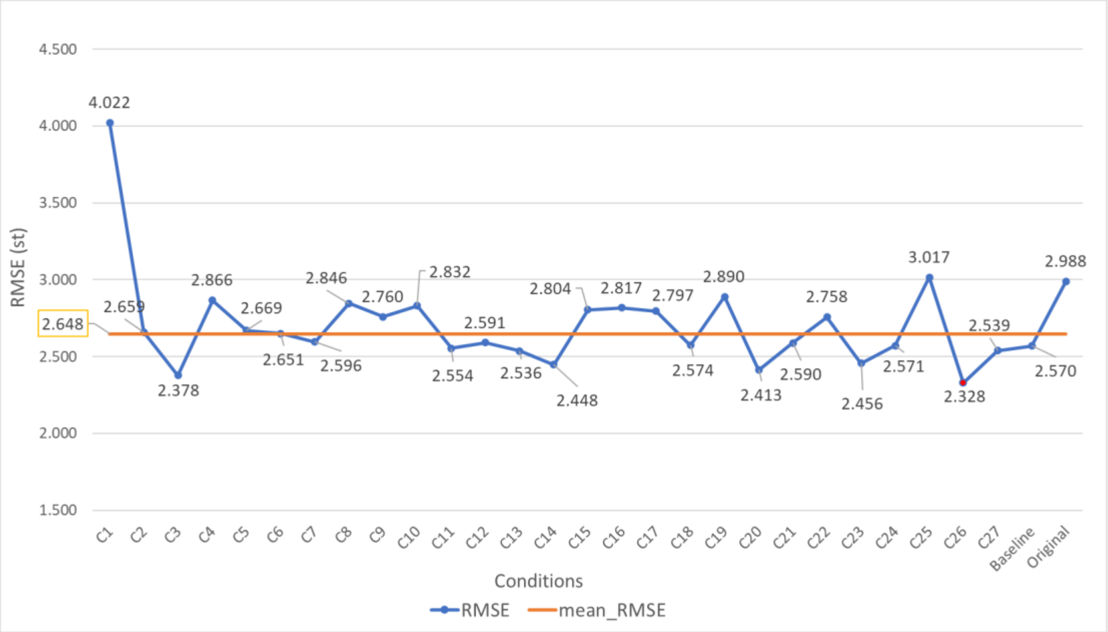
5.2 Accuracy Based on L set[31]

[30] Excluding C1

[31] For comparison of performances of all conditions measured on L set, refer to the line plots. For comparison of accuracy measured on the L set between the bouncing conditions, the baseline

The utterance-specific results of L set in term of RMSE and correlation are shown in Figure 12 and 13. C1 is again excluded for its low accuracy (RMSE=4.022, Corr=0.764) and C26 (RMSE=2.328, Corr=0.882) achieves the highest accuracy. The mean accuracy across the bouncing conditions excluding C1 is RMSE=2.648 (SD=0.174), Corr=0.830 (SD=0.025).

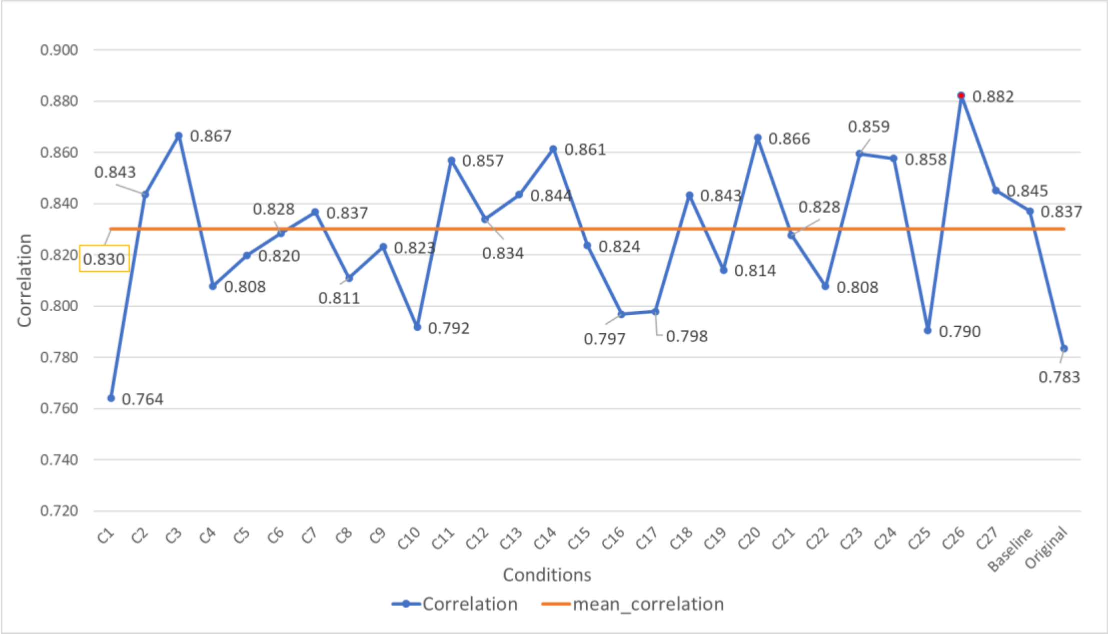
Figure 12: Utterance-specific RMSE (averaged across L set)



The original condition performs much less ideal (RMSE=2.988, Corr=0.783) in synthesizing the L set, which is largely expected as the 320 utterances directly concern post-low bouncing. The accuracy of the baseline condition (RMSE=2.570, Corr=0.837) still shows a marginal advantage over the average of all bouncing conditions. Accuracy in synthesizing the three post-low N tones is shown in Figure 14 and 15. C1 still has the lowest RMSE as an outlier hence it is excluded from the bouncing conditions for calculating the average performance (RMSE=2.569, SD=0.243, Corr=0.447,

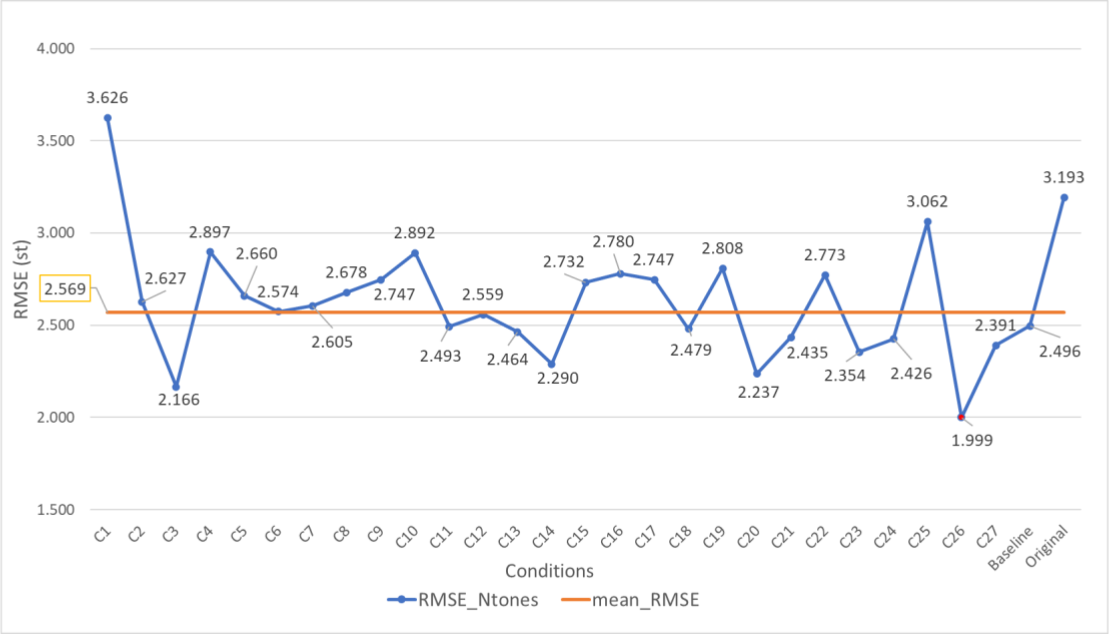
condition, the original condition, and condition with the best performance, see Table 6 for quick reference.

Figure 13: Utterance-specific Correlation (averaged acorss L set)



SD=0.058). C26 again outranks with RMSE of 1.999 and correlation of 0.566. The outcome of the baseline condition outperforms (RMSE=2.496, Corr=0.457) the average score. The incapability of the original condition to simulate post-low bouncing is more apparent with larger RMSE (3.193) and much lower correlation (0.336).

Figure 14: Syllable-specific RMSE for N tones (averaged across L set)



Accuracy terms averaged across the full corpus and the L set so far could lead to three tentative conclusions; Firstly, the original condition achieves the lowest accuracy rate comparing with the bouncing conditions, and it could be more readily observed based on the L set, where post-low bouncing is concerned. On the other hand, the baseline condition could achieve higher accuracy rate than the average performance of the bouncing conditions, which shows the efficiency of Equation (6). Also, the data invariantly show C26 has the highest synthesis accuracy among all conditions.

Figure 15: Syllable-specific Correlation for N tones (averaged across L set)

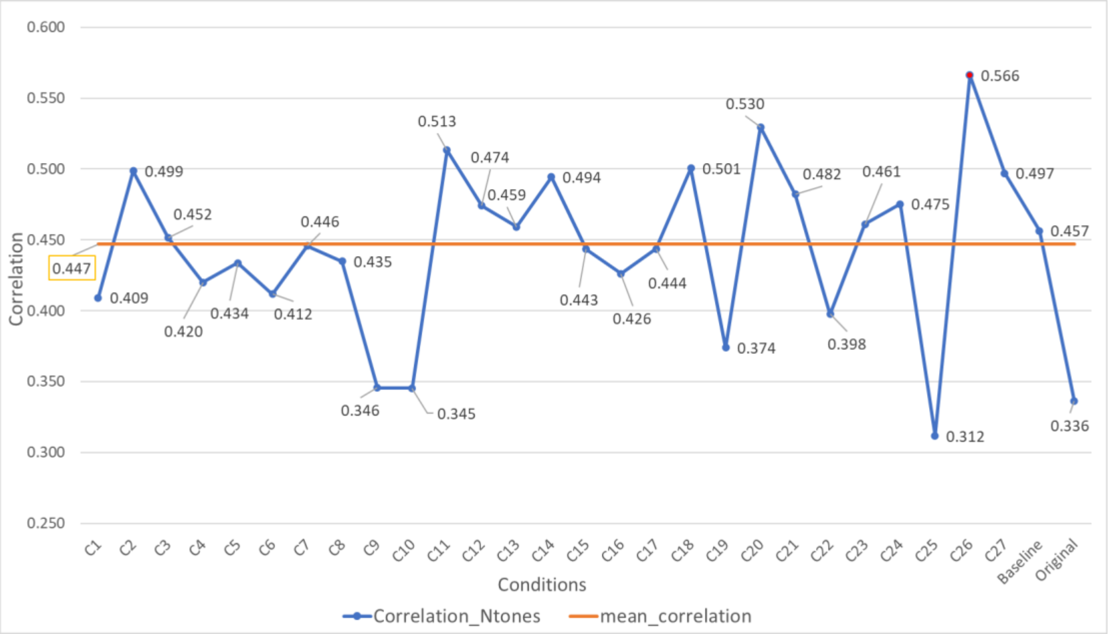


Table 6: Accuracy terms averaged across L Set by conditions

Conditions/Accuracy	Utterance-specific Accuracy		Ntones Accuracy	
-	RMSE(st)	Correlation	RMSE(st)	Correlation
Bouncing Conditions	2.648$_{(SD=0.174)}$	0.830$_{(SD=0.025)}$	2.569$_{(SD=0.243)}$	0.447$_{(SD=0.058)}$
Baseline Condition	2.570	0.837	2.496	0.457

Original Condition	2.988*	0.783*	3.193*	0.336*
C26	**2.328**	**0.882**	**1.999**	**0.566**

*Bold for best performances. * for the least ideal performances*

However, before concluding that C26 is the optimal condition, there are still two issues to address. Firstly, the standard deviation of accuracy among the bouncing conditions only shows little dispersion (excluding C1). One needs to then confirm if indeed the difference of RMSE and correlation of the bouncing conditions measured on the entire corpus is largely caused by different simulation accuracy of post-low bouncing, or could it be caused by the accuracy of utterances containing the other three full tones, as the randomness of results brought by simulated annealing is uncertain. The fact that the largest dispersion is witnessed in the accuracy rates of post-low N tones based on L set (SD_rmse=0.243, SD_corr=0.058) is one piece of evident supporting the former hypothesis. A simple calculation of coefficient of variation (CV) for results measured on the entire corpus and the L set is shown in Table 7 for direct comparison.

Table 7: Coefficient of variation calculated for RMSE and Correlation based on the full Corpus and L Set

Accuracy terms	CV (RMSE)		CV(Correlation)	
	Utterance-specific	N tones	Utterance-specific	N tones
Full Corpus	5.23%*	6.76%*	2.03%*	5.96%*
L set	**6.57%**	**9.46%**	**3.01%**	**12.98%**

*Bold for larger CV, * for smaller CV.*

It then becomes apparent that despite all bouncing conditions (excluding C1) do not show large dispersion, accuracy measured on the L set show larger fluctuation than that of the full corpus. Most importantly, accuracy of N tones based on the full corpus and L set show significant difference with much larger dispersion in the post-low N tones in the L set, especially in terms of correlation (5.96% against 12.98%). This is strong evidence that the fluctuation of accuracy rates in simulating post-low bouncing (post-low N sequences in L set) is a highly possible cause for different results based on the entire corpus. The second issue is there are many competing bouncing conditions that perform close to optimal (C26) and it is not yet known if C26 achieves the highest only by chance. The strategy then is to compare 27 bouncing conditions that vary with three parameters, namely Threshold, Intercept and Slope based on the accuracy of simulating L set where post-low bouncing is concerned, as it has been shown L set is largely the source of variance. From Table 8 it seems the three groups which set Threshold=0, Intercept=200 and Slope=-550 could achieve better performance, while again the difference is marginal. Three-way ANOVA tests[32] shows that only Slope has a

Table 8: Accuracy terms grouped by the variation of Threshold, Intercept and Slope

Parameters	Threshold			Intercept			Slope		
Candidates	-2.5	0	2.0	100	200	300	-350	-450	-550
RMSE (utterance)	2.926*	**2.575**	2.606	2.770*	**2.640**	2.697	2.828*	2.661	**2.618**
RMSE (N tones)	2.878*	**2.454**	2.501	2.649*	**2.574**	2.610	2.731*	2.604	**2.498**
Correlation (utterance)	0.806*	**0.844**	0.839	0.829	**0.834**	0.825*	0.822*	0.828	**0.839**
Correlation (N tones)	0.399*	**0.486**	0.454	**0.453**	0.444	0.441*	0.428*	**0.456**	0.455

*Best performances marked Bold. Least ideal performance marked by **

[32] Threshold, Intercept and Slope as factors. No interaction is tested as they are all independent from

significant influence on utterance-specific RMSE (F(2,20)=4.228, p=0.029) and correlation (F(2,20)=6.718, p=0.006). Post-hoc tests using Tukey's HSD for multiple comparisons found there is significant difference between the -350 and -450 group in terms of RMSE (P=0.041, 95%, C.I = [0.013, 0.690]) while in terms of correlation there are significant differences between the -350 and -550 groups, -350 and -450 groups (p=0.022, 95%, C.I = [-0.061, -0.004]; P=0.008, 95%, C.I = [-0.067, -0.010])). Regarding N tones, Slope still shows a true impact on both RMSE (F(2,20)=3.943, p=0.036) and correlation (F(2,20)=4.062, p=0.033) while other parameters do not. Similarly, Tukey HSD test shows only the -350 and -450 group differ significantly in RMSE (p=0.038, 95%, C.I = [0.017, 0.647]) and correlation (p=0.037, 95%, C.I = [-0.081, -0.002]. Overall, it seems only Slope is a significant factor determining the varying accuracy rates, and the difference between conditions with different Slope settings mainly lies in the significant difference between the -350 and -450 groups. On the other hand, -450 and -550 group do not differ significantly. Hence, according to the post-hoc tests and Table 7, one can see setting Slope to -450 and -550 would render equally good results with advantage over -350. Also, switching Threshold between -2.5, 0 and 2.0, or Intercept between 100, 200 and 300 does not yield significantly different accuracy rates. It is now safe to conclude that, despite the fact that there are candidates could potentially reach similar or even higher accuracy, C26 (2.0, 300, -450) could indeed be one of the optimal conditions, and it is largely right that it did not reach such accuracy rate by chance. The statistical analysis also supports the balance-perturbation theory as training the corpus by parameters around Equation (6) largely yields equally good results. In the next section, results of correlation analysis based on the time normalized synthetic F_0 and visual inspection of the synthetic contours under

each other.

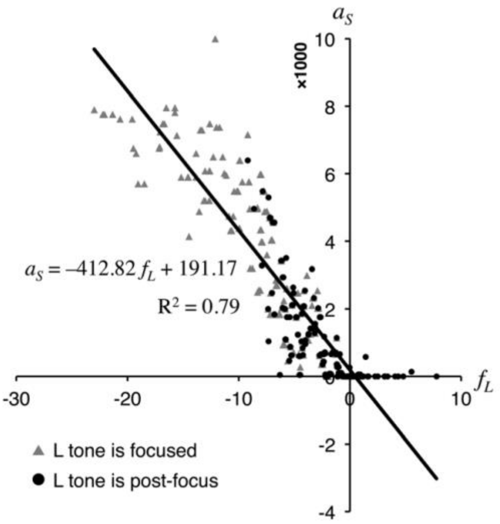
C26 are reported to further check the accuracy of simulating post-low bouncing.

6 Correlation Analysis and Visual Inspection

6.1 Bouncing Effect Visualization and Correlation Analysis

As shown in the linear regression performed in Prom-on et al. (2012), there are many points siting around the x-axis showing little to no F_0 lowering and initial acceleration (see Figure 16). On the other hand, utterances with focused L tone shows significant amount of bouncing. This makes sense according to the balance-perturbation hypothesis as the duration of the L tone lengthens when under prosodic focus. It allows more time for L to approach its low target and trigger the bouncing mechanism. To show the synthetic performance of post-low bouncing in a full scale, the strategy is to visualize the bouncing effect by scatter-ploting F_0 rising excursion (F_E) as a function of F_0 lowering (f_L) and calculate Pearson's correlation coefficient (r) based on the full L set (320 utterances) and the utterances with on-focus L tones (160 utterances, hereafter focused L set or on-focus L set). F_E refers to the magnitude of peak F_0 of three post-low

Figure 16: Linear regression of the simulated acceleration modified on the initial states of the first post-low N tone as a function of F_0 lowering (f_L) (in Prom-on et al. 2012)



syllables relative to the valley F_0 in the first post-low N tone[33]. In the synf0 file generated by PENTAtrainer2, the time-normalized F_0 can be directly extracted without extra measurement. Similarly, correlation is calculated on the original F_0 for comparison. f_L is simply the subtraction of the minF_0 in the L tone by each speaker's mean F_0. Each speaker's synthesized mean F_0 is calculated by taking the average of the time-normalized F_0 and the original mean F_0 is taken from the qTA file PENTAtrainer2 generates. They are very close to each other as shown in Table 9.

Table 9: Synthesized mean F_0 and original mean F_0 for eight speakers

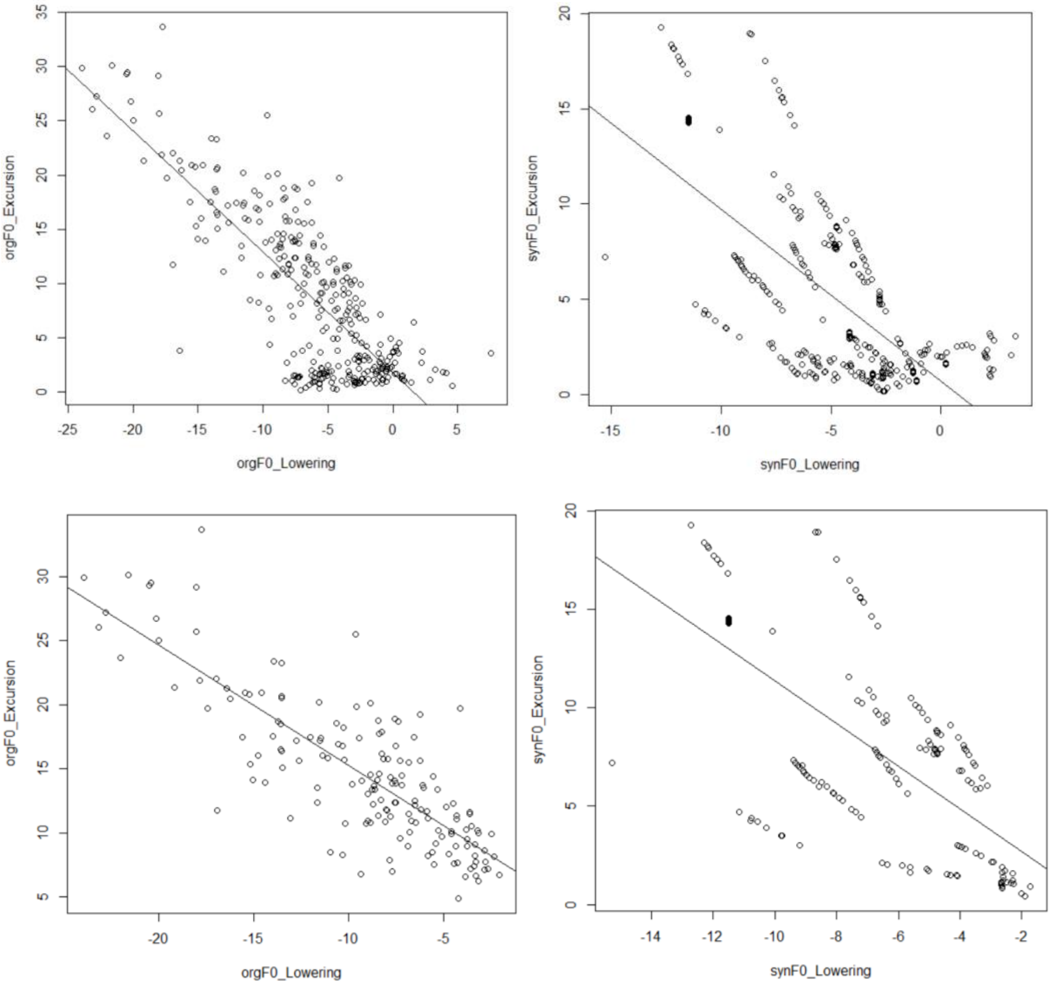
meanF_0	S1	S2	S3	S4	S5	S6	S7	S8
Syn	98.41	90.34	93.16	95.23	88.40	84.38	83.59	96.04
Org	98.53	90.41	93.23	95.24	88.29	84.38	83.63	96.10

The scatter plots and the best fit linear lines are shown in Figure 17. The correlation between F_L and F_E in the original utterances is $|r| = 0.80$ and for the synthesized utterances it is $|r| = 0.67$, which is close to strong and indicates good synthesis of the bouncing effect. Correlation analysis based on the focused L set also show similar results. The correlation between f_L and F_E for the original utterances is $|r| = 0.82$ and for the synthesis is $|r| = 0.65$, which is close to ideal as well. One thing immediately can be observed from Figure 17 is that it seems there is not enough amount of bouncing in the synthesized utterances as F_E would not go above 20. In fact, the synthetic F_0 does not go as low as the original utterances in the L tone, specifically, f_L seldom goes below the mean F_0 by 10st in the synthesized utterances, which leads to insufficient amount

[33] This is difference from Prom-on et al. (2012) as they calculated F_E by the subtraction of the initial F_0 of the first post-low syllable from the peak F_0 in the post-low N sequences.

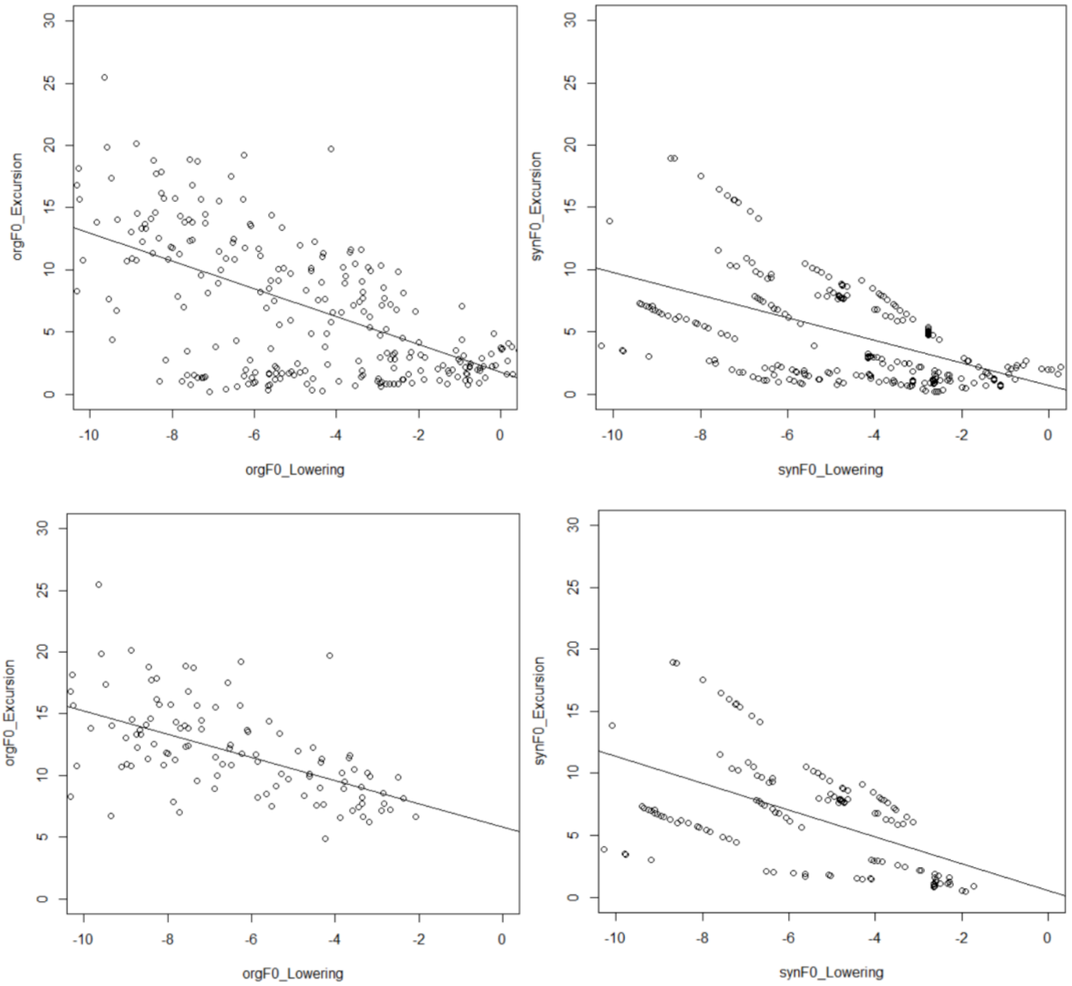
of bouncing. As shown in the zoom-in of the scatter plots (see Figure 18), range [-10, 0] of f_L seems to be where most of the synthesized utterances are densely distributed. Both the original and the synthesized utterances show similar amount of bouncing within the f_L range of [-8, 0]. For f_L range of [-10, -8], the simulated bouncing would still be slightly weaker but could still be ideal. According to the scatter plots of Figure 17 and Figure 18, it could be concluded that PENTAtrainer2 under C26 can produce largely sufficient bouncing, especially when the F_0 lowering in the L tone is within the range of [-8, 0]. While it seems to be slightly less bouncing when f_L goes below the

Figure 17: Scatter plots of f_L and F_E from the original and the synthesized utterances (left column for original utterances, right column for synthesized utterances; upper row for L set, lower row for focused L set)



mean F_0 by 8st, and even worse performance of the synthetic F_0 are observed in Figure 17 when f_L goes below mean F_0 by 10st[34]. This then needs to be carefully examined by visual inspection of the synthetic contours to see how it differs from the original.

Figure 18: Zoom-in of Figure 18 (f_L range = [-10, 0])(left column for original utterances, right column for synthesized utterances; upper row for L set, lower row for focused L set)



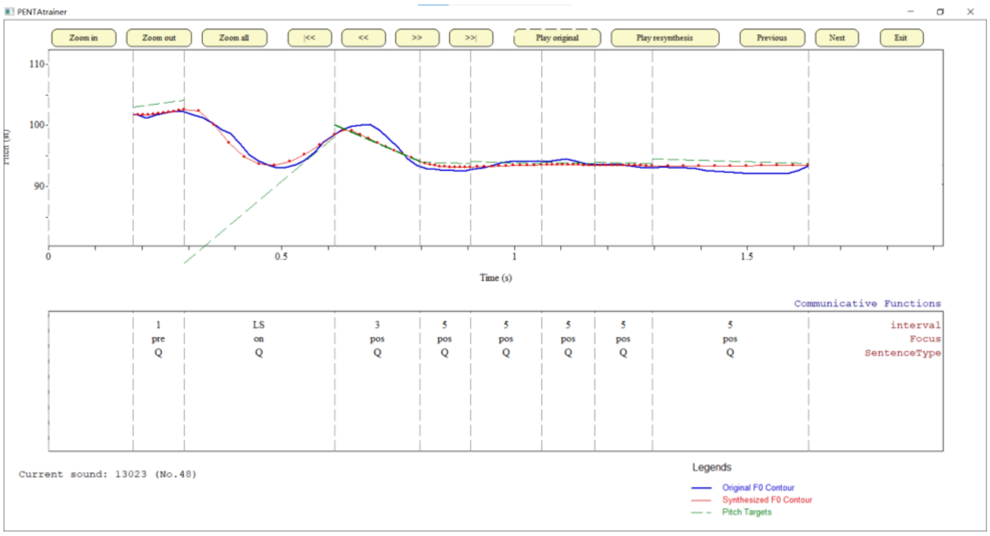
6.2 Visual Inspection

[34] The bouncing does not extinct when f_L reaches the threshold of -10. As can be seen from Figure 17, there are cases that the synthetic L surpasses the -10 threshold and achieves decent amount of bouncing. But again, F_E never exceeds 20st beyond.

After checking all synthetic contours in the L set, the results are then presented as follows; first this section shows a representative good fit of a post-focus L utterance and an on-focus L utterance. Then it shows a representative inaccurate fit and tries to locate the source of such inaccuracy.

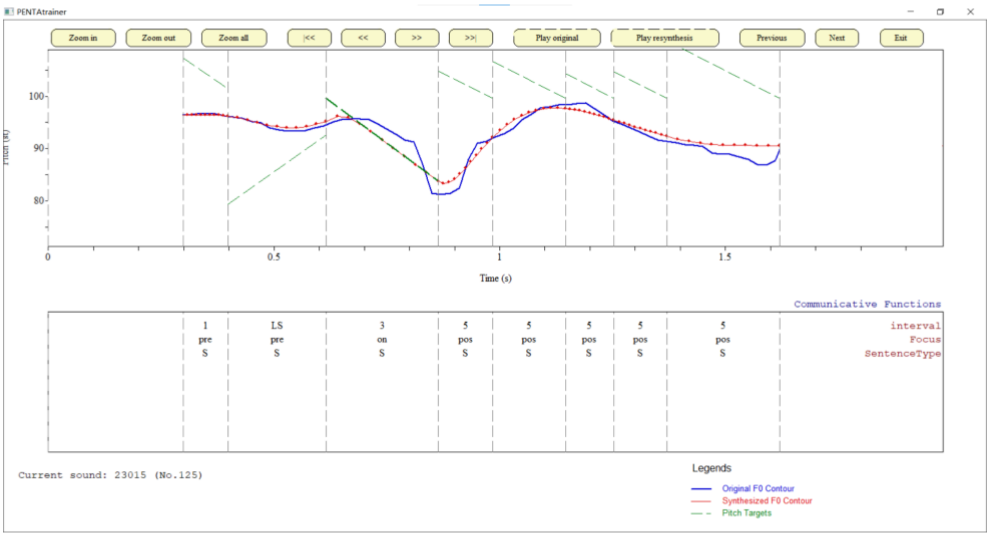
Figure 19 shows utterance 13023 with a post-focused L tone from speaker 4 (mean F_0 around 95st). Both the original and the synthetic contour show only a slight bouncing effect and the synthesis seems to work well. In Figure 20 where the L tone is under prosodic focus (utterance 23015), the original F_0 goes around 13st below the mean F_0 of speaker 4 and one can see the iconic bouncing contour starting from the onset of the first N tone after L, and the raising effect seems to carry into the consecutive three N tones. Around the time of the third N tone, the F_0 reaches the highest among the whole utterance. The synthetic contour could still fit well with the original by going around 10st below mean F_0, though it triggers the bouncing effect a bit earlier than the original

Figure 19: Synthesized contour compared with the original contour (utterance 13023 from speaker 4)



Current sound: 13023 (No.48)

Figure 20: Synthesized contour compared with the original contour (utterance 23015 from speaker 4)



Current sound: 23015 (No.125)

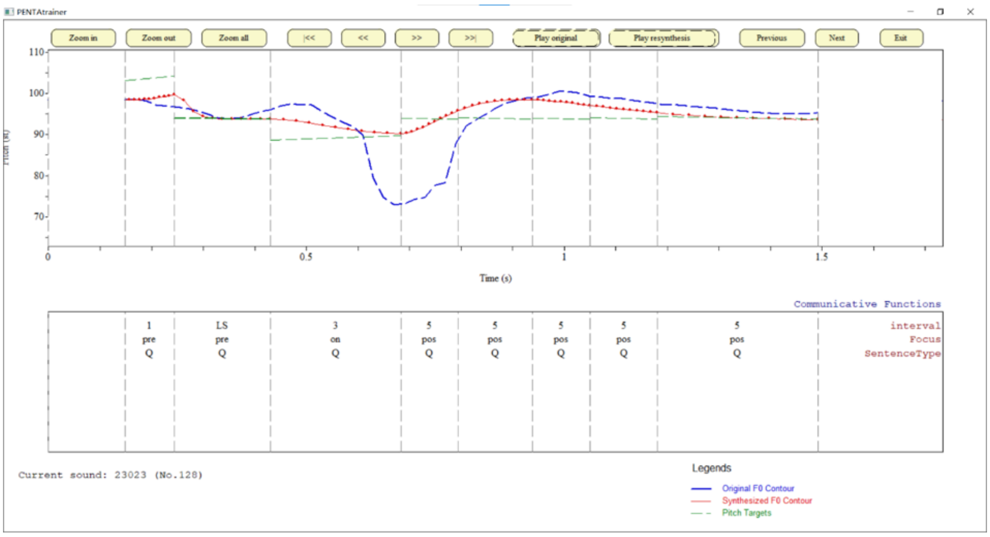
utterance. Since the L tone does not go as low as the original contour, it triggers a slightly less raising force, which is also visible in the first N tone. Overall, it seems the synthetic contour can sufficiently capture the bouncing effect, especially looking at the second and third post-low N tones. In the on-focus L set of speaker 4, there are some utterances that obtain exceptionally high RMSE and low correlation in the N tones. Upon checking the accuracy reports and the synthetic contours, same happens in the focused L set in a few other speakers (speaker 1, 2, 3). For example, utterance 23023 from speaker 4 is one example of an inaccurate fit, as shown in Figure 21. The original L goes below mean F_0 by 23st but the synthetic L only lowers 5st relative to the mean, this is the immediate trigger of less bouncing in the next post-low N tone. While utterance 23123 (L also under focus) from the same speaker shows a much more ideal fit as shown in figure 22. Both the original and the synthetic L goes below mean F_0 by around 5st which triggers similar raising force, as evident in the next three post-low N tones. Notice both utterances contain the same functional combination for the L tone (3, on, Q), representing an on-focus third tone realized in a question. Thus, they are assigned the same target by PENTAtrainer2, which is (**m**=4.38, **b**=-5.54)[35]. This kind of comparison could also be found in the other three speakers (speaker 1, 2 and 3) in the (3, on, Q) combination. A speculation then would be, due to such inconsistency of on-focus L production of some speakers, the system finally decides the target for the functional combination (3, on, Q) should be a slightly higher static target, for example, around 6st below the mean F_0 of speaker 4 as shown in Figure 21 and 22, than it is supposed to be for some of original extreme-lowering cases (Figure 21), and likewise for other three speakers. In comparison, speakers with more consistent production of focused-L tones does not show such occasional high RMSE and low correlation, nor can the investigator see such extremely contrasting patterns like Figure 21 and 22 in those speakers. For example, in speaker 6 (mean F_0 around 84), the synthetic contours more or less look like Figure 23 where the target for the third syllable is around 6st lower relative to the speaker's mean F_0 and the simulated contour seems to fit well with the original. One evidence in support of such speculation is there are only 16 out of 80 L tones encoded as (3, on, Q) (20% of the entire (3, on, Q) combination set) that goes lower than -10st relative to speaker's mean F_0. Since PENTAtrainer2 learns in a data-driven manner, it is then largely expected that it does not assign a lower target for (3, on, Q) to approach. This would also partially explain the finding in section 6.1 as to why the synthesized L largely does not go more than 10st below the mean F_0[36].

The visual inspection and the correlation analysis could lead to the conclusion that PENTAtrainer2 is largely capable of faithfully simulating post-low bouncing under

[35] Obtained from parameter reports in PENTAtrainer2

[36] Those do go below -10st in Figure 17 seem to be the L tones that are labelled as (3, on, S), for example the L in Figure 20. There are a total of 56 utterances in the focused L set that go more than 10st below speaker's mean F_0, 16 of them are labeled as (3, on, Q) and the rest 40 are labeled as (3, on, S), hence PENTAtrainer2 would naturally assign lower target for (3, on, S). For example (3, on, S) is assigned a height of -11.5 in Speaker 4.

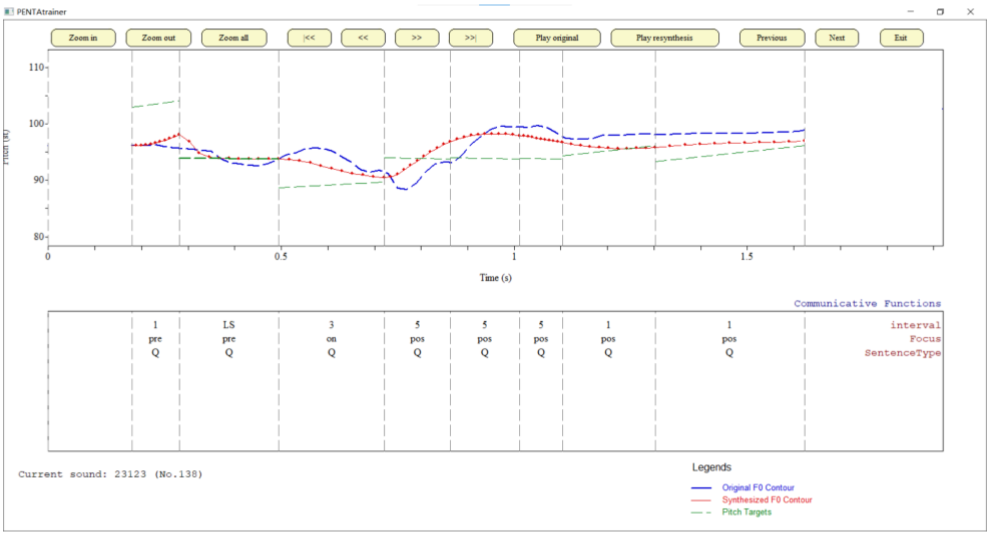
Figure 21: Synthesized contour compared with the original contour (utterance 23023 from speaker 4)



1	LS	3	5	5	5	5	5	interval
pre	pre	on	pos	pos	pos	pos	pos	Focus
Q	Q	Q	Q	Q	Q	Q	Q	SentenceType

Current sound: 23023 (No.128)

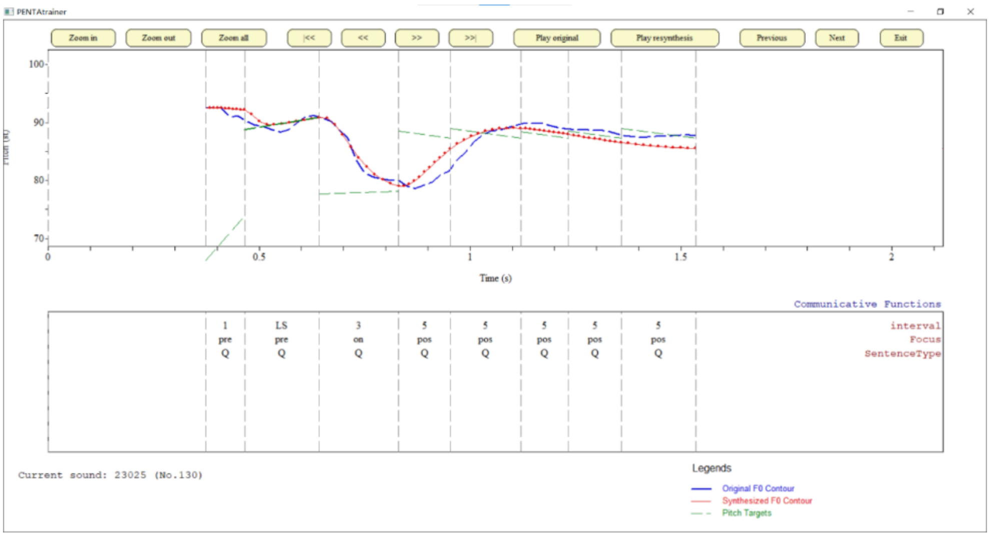
Figure 22: Synthesized contour compared with the original contour (utterance 23123 from speaker 4)



1	LS	3	5	5	5	1	1	interval
pre	pre	on	pos	pos	pos	pos	pos	Focus
Q	Q	Q	Q	Q	Q	Q	Q	SentenceType

Current sound: 23123 (No.138)

Figure 23: Synthesized contour compared with the original contour (utterance 23025 from speaker 6)



```
Current sound: 23025 (No.130)
```

C26. The only problem the investigator can notice so far is there is insufficient amount of bouncing in simulating some extreme cases of F_0 lowering (below -10st relative to mean) and the subsequent bouncing (above 20 st from the valley F_0 in the initial N), possibly due to a higher target assigned to the L labeled as (3, on, Q), which in turn triggers less bouncing.

7 Discussion and Conclusion

It now can be concluded that setting the Threshold, Intercept and Slope as 2, 300 and -450 for the general bouncing equation defined in Prom-on et al. (2012) can yield largely fine synthetic contours in PENTAtrainer2, based on objective evaluation of RMSE and Correlation, correlation analysis between F_0 lowering and F_0 excursion, and visual inspection of synthetic contours. It is also largely safe to fix the parameters in PENTAtrainer2 for future use. For future study, what immediately could help further check the reliability of synthesis accuracy is a perception test, especially in f_L range of [-10, -8] where the investigator noticed slightly weaker bouncing in simulation, whether it is ideal cannot be subjected determined by the investigator but requires further perception test. Secondly, as PENTAtrainer2 possibly encounters issue to capture the inconsistency of the production of on-focus L tones (3, on, Q), what could help to eliminate such error is for the investigators to add another functional layer. For example, one might want to categorize the extreme F_0 lowering and the milder F_0 lowering for the L tone, since even they are both under prosodic focus, some speakers would still vary significantly in their style of realizing such focus[37]. According to the correlation analysis and the visual inspection, it seems -8st or -10st could be a reasonable reference for splitting extreme lowering and mild lowering for L, as the synthesized utterances does not perform as well when f_L goes below the mean F_0 by 8st, and they largely do not go below the mean by 10st, which directly causes the problem of insufficient bouncing relative to the original utterances.

This study exemplifies two things, firstly, the success of simulating post-low bouncing using parameter sets based on Equation (6) verifies the general mechanism developed by Prom-on et al. (2012) is efficient and it further supports the balance-perturbation hypothesis as highly reliable. With the TA model and the bouncing mechanism, PENTA model now seems to be able to account for the prosodic variability of N tones. Secondly, the study shows the rigorous power of PENTAtrainer2, as a computational implementation of PENTA model, in theory testing, as it checks the synthesized sounds syllable-by-syllable against original utterances with all fine details generated to directly compare with natural speech, which can then be visually checked by the investigators or perceptually tested by human participants.

[37] Whether it is related to sentence modality is unknown and will not be pursued in this study.

References

Andruski, J., & Costello, J. (2004). Using polynomial equations to model pitch contour shape in lexical tones: an example from Green Mong. *Journal of the international Phonetic Association*, *34*(2), 125-140. doi: 10.1017/s0025100304001690

Atkinson, J. (1978). Correlation analysis of the physiological factors controlling fundamental voice frequency. *The Journal of The Acoustical Society of America*, *63*(1), 211. doi: 10.1121/1.381716

Bailly, G., & Holm, B. (2005). SFC: A trainable prosodic model. *Speech Communication*, *46*(3-4), 348-364. doi: 10.1016/j.specom.2005.04.008

Chen, Y., & Xu, Y. (2006). Production of Weak Elements in Speech – Evidence from F_0 Patterns of Neutral Tone in Standard Chinese. *Phonetica*, *63*(1), 47-75. doi: 10.1159/000091406

Crystal, D. (1969). Prosodic System and intonation in English. Cambridge: Cambridge University Press.

Erickson, D. (1993). "Laryngeal muscle activity in connection with Thai tones," Ann. Bull. Res. inst. Logoped. Phoniatr. Univ. Tokyo 27, 135–149.

Erickson, D. (2011). "Thai tones revisited," J. Phon. Soc. Jpn. 15, 74–82.

Fujisaki, H. (1983). Dynamic Characteristics of Voice Fundamental Frequency in Speech and Singing. *The Production of Speech*, 39-55. doi: 10.1007/978-1-4613-8202-7_3

Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics*, *22*(4), 477-492. doi: 10.1016/s0095-4470(19)30296-7

Gu, W., & Lee, T. (2009). "Effects of tone and emphatic focus on F0 contours of Cantonese speech—A comparison with standard Chinese," Chin. J. Phonetics 2, 133–147.

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of Vertical Larynx Movement and Cervical Lordosis in F0 Control. *Language and Speech*, *42*(4), 401-411. doi: 10.1177/00238309990420040301

Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, *39*(3-4), 311-352. doi: 10.1016/s0167-6393(02)00047-x

Ladd, D. R. (2008). intonational Phonology, Cambridge University Press.

Lin, M.; Yan, J. (1980). Beijinghua qingsheng de shengxue xingzhi. Dialect 3: 166–178

Liu, F., Surendran, D., & Xu, Y. (2006). Classification of statement and question intonations in Mandarin In: Proceedings of Speech Prosody 2006, Dresden, Germany. PS5-25_0232.

Liu, F., & Xu, Y. (2007). "The neutral tone in question intonation in Mandarin," in Proceedings of INTERSPEECH2007, Antwerp, pp. 630–633.

O'Connor J. D., & Arnold G. F. (1961). Intonation of Colloquial English, London: Longmans.

Palmer, H. E (1922). English Intonation, with Systematic Exercises, Cambridge: Heffer;

Pierrehumbert J. (1980) The Phonology and Phonetics of English Intonation [Ph.D. dissertation]. MIT, Cambridge, MA. [Published in 1987 by Indiana University Linguistics Club, Bloomington].

Pierrehumbert, J. (1981). Synthesizing intonation. *The Journal of The Acoustical Society of America*, *70*(4), 985-995. doi: 10.1121/1.387033

Prom-on, S., Liu, F., & Xu, Y. (2012). Post-low bouncing in Mandarin Chinese: Acoustic analysis and computational modeling. *The Journal of The Acoustical Society of America*, *132*(1), 421-432. doi: 10.1121/1.4725762

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of The Acoustical Society of America*, *125*(1), 405-424. doi: 10.1121/1.3037222

Prom-on, S., & Xu, Y. (2013). Modeling speech melody as communicative functions with PENTAtrainer2, Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France, 2013. 82-85.

Shen, J. (1994). “Hanyu yudiao gouzao he yudiao leixing (intonation structures and patterns in Mandarin),” Fangyan (Dialect) 3, 221–228.

Shih, C. (1987) The phonetics of the Chinese tonal system (AT&T Bell Labs, Murray Hill).

Shih, C. (1988). “Tone and intonation in Mandarin” in Working Papers of the Cornell Phonetics Laboratory, edited by N. Clements, No. 3, pp. 83–109.

Shipp, T. (1975). Vertical Laryngeal Position During Continuous and Discrete Vocal Frequency Change. *Journal of Speech and Hearing Research*, *18*(4), 707-718. doi: 10.1044/jshr.1804.707

't Hart J., Collier, R., & Cohen A. (1990). A perceptual Study of intonation — An experimental-phonetic approach to speech melody, Cambridge: Cambridge University Press.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*(1), 61-83. doi: 10.1006/jpho.1996.0034

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0contours. *Journal of Phonetics*, *27*(1), 55-105. doi: 10.1006/jpho.1999.0086

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*(3-4), 220-251. doi: 10.1016/j.specom.2005.02.014

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication*, *46*(3-4), 220-251. doi: 10.1016/j.specom.2005.02.014

Xu, Y. (2013). ProsodyPro — A tool for large-scale systematic prosody analysis. in Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France: 7-10.
Xu, Y., & Emily Wang. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Communication*, 33(4), pp.319-337.

Xu, Y., & Liu, F. (2006). Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics 18*: 125-159.

Xu, Y. & Prom-on, S. (2010-2021). qTAtrainer. praat.
Available from: http://www.homepages.ucl.ac.uk/~uclyyix/qTAtrainer/.

Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *The Journal of The Acoustical Society of America*, *111(3)*, 1399-1413. doi: 10.1121/1.1445789

Yip, M. (2008). The tonal phonology of Chinese; PhD diss. MIT, Cambridge.

Zemlin, W. R. (1988). Speech and Hearing Science — Anatomy and Physiology. Englewood Cliffs, New Jersey: Prentice Hall

[i] Acknowledgement

Upon completion of this dissertation, I would like to thank the people who helped me along the way. First, I would like to thank my supervisor, Professor Yi Xu. Having never met him in person this year due to the pandemic, he nevertheless gave me his warmest encouragement and best support. Being a top phonetician and a good programmer, he offered me insights both theoretically and technically, and mentored me how to do research independently. Without his patience and knowledge, the study wouldn’t have gone smoothly, and I wouldn’t be so fulfilled. I’d also like to thank my parents and my cousin Jie, who is studying linguistics at Reed College. They supported me wholeheartedly for pursuing my interest. Finally, I’d like to thank my friends, Boyan, Wenxiang, Yi’an and Yichuan. They kept me company when I was ill and made London feel like home for me. Playing Mahjong with them was the most enjoyable thing during the worst time.