Journal of Information Hiding and Multimedia Signal Processing ©2020 ISSN 2073-4212
Ubiquitous International Volume **11**, Number **1**, March **2020**

A New Computational Method for Determining Parameters Representing Fundamental Frequency Contours of Speech Words

Yen Thai Ta

Hanoi University of Business And Technology, Hanoi, VietNam.
thaity@hubt.edu.vn

Hoang Huy Ngo

Electric Power University of the Vietnam Ministry of Industry and Trade
235 Hoang Quoc Viet str., Co Nhue, Tu Liem, 129823, Hanoi, VIETNAM
huynh@epu.edu.vn

Van Hung Nguyen

Military Institute of Science And Technology, Hanoi, Vietnam.
nvhnt73@gmail.com

Received October 2019; revised December 2019

Abstract. *In tonal languages, one of the basic parameters of speech is the quantitative target approximation (qTA) for generating fundamental frequency contours (F_0 contours) of words. It is not easy to automatically estimate the representational parameters or model the variations in these parameters because of the tone sandhi in the context of sentences, where F_0 contours have a complex shape. Based on continuous function approximation by polynomials, in this paper, we propose a computational method to calculate automatically the quantitative target approximation vector (qTAV) parameters in an extended model of the Xu framework to model the variations of the representational parameters for two-syllable tones, which appear with the highest frequency in tonal languages. The experimental results show that the proposed methods are able to present qTAV parameters to generate F_0 contours of two-syllable tones with complex shapes.*
Keywords: F_0 contours, Xu model, qTA, Polynomial approximation.

1. **Introduction.** Nowadays, Text to speech (TTS) systems are increasingly used by the radiologist to create radiology study reports. With the integration of the Laboratory Information System (LIS) and Radiological Information Systems (RIS), patient identifiers and examination can be automatically mapped into examination reports. There are many potential benefits of report automation to radiologists including improvements in efficiency, accuracy, and avoiding fatigue [1].

Besides, TTS systems can help people with disabilities integrate into the community by being able to use computers more easily. For example, the JAWS software (Job Access With Speech), is the world's most popular screen reader, developed for computer users whose vision loss prevents them from seeing screen content or navigating with a mouse. JAWS provides speech as an output for the most popular computer applications on your PC such as Microsoft office, Web browsers etc [2].

Due to its application demand, more and more researches have been going into speech representation. Despite that, the estimation and modeling of F_0 contours are still open

issues at this point. A speech sound contains an important type of frequency, namely fundamental frequency (F_0), which relates to vocal cords function and reflects the rate of vocal cords' vibration during pronunciation process.

Prosody is employed to express attitude, assumptions and attention in daily speech communication and has been studied by linguists, phoneticians, speech therapists. In recent artificial intelligence developments, people seek to communicate effectively with intelligent machines on a more personal and human level. To synthesize natural and human-sounding speech by computers, prosody plays an important role, which is related to pauses, pitch, speech rate and loudness. Among the factors which weave the prosody, pitch or fundamental frequency (in this paper, we consider pitch and fundamental frequency (F_0) the same) is the most characteristic.

From a modeling perspective, a model that is rarely used if it is not predictive. To make a model predictive, however, it is critical to first determine what the predictors should be. If, as suggested above, communicative functions like accent, focus and sentence type and their interactions are directly behind the complex surface F_0 contours, these communicative functions should then be the predictors. An alternative to such functional modeling is to simulate F_0 with predictors whose functional status is ambiguous, or whose definition includes characteristics of observed F_0 patterns, e.g., pitch accents, F_0 turning points, etc. From a theoretical perspective, functional modeling provides a powerful tool for hypothesis testing. That is, by assessing how well the surface F_0 contours are generated based on a set of hypothesized predictors, investigators can validate or falsify both general and specific theoretical assumptions about tone and intonation. Such a process is known as analysis-by-synthesis [3].

Parametric representation of speech often implies F_0 contours as a part of the model. There have been many attempts over the past decades to build a robust model capable of simulating various prosodic phenomena through F_0 modeling [4, 5, 6, 7, 8]. These approaches can be divided into two general categories, namely, those that model F_0 contours directly and those that attempt to simulate the underlying mechanisms of F_0 production. Models belonging to the first category are derived mainly based on the shape of the F_0 contours, with minimal consideration about the articulatory process of F_0 production.

The Fujisaki model is an effective model for approximating the contour of the fundamental frequency precisely for the source model of speech synthesis, The model represents the coarticulation of spectral frequencies by making an equation for a target model of speech perception and so on [5, 6, 7].

Quantitative modeling is one of the most rigorous means of testing our understanding of a natural phenomenon. This is particularly true if the model is built directly on assumptions that closely reflect the contested view about the mechanisms underlying the phenomenon. Modeling can also help to improve our knowledge by forcing us to make our theoretical postulations as explicit as possible. Therefore, to improve our understanding of human words, quantitative modeling is also indispensable. In [8], the author simulated tone, stress, and focus in Mandarin and English with a Quantitative Target Approximation(qTA) model that generates surface F_0 contours through the process of target approximation.

The qTA model simulates the production of tone and intonation as a process of syllable-synchronized sequential target approximation. It adopts a set of biomechanical and linguistic assumptions about the mechanisms of speech production. The directly modeled communicative functions directly modeled are lexical tones in tonal languages and lexical stress in non tonal ones and focus in both type languages. The qTA model is evaluated by extracting function-specific model parameters from natural speech via supervised learning automatic analysis by synthesis and comparing the F_0 contours generated with

the extracted parameters to those of natural utterances through numerical evaluation and perceptual testing. The F_0 contours generated by the qTA model with the learned parameters were very close to the natural contours in terms of root mean square error. More important, the generated speech quality was well perceived by human listeners.

According to Fujisaki and Xu models, to predict the form of F0 contours for words, both models follow through two main steps:

Step 1. Extracting the model parameters from given sample F_0 contours corresponding to the words by fitting methods.

Step 2. Predicting the model parameter to generate the F_0 contours of the words.

Generally until now, both methods have not provided a numerical solution to do step 1 automatically. Therefore, in this paper we propose a numerical method to solve the problem based on the approach qTA of Xu et al.

The rest of the paper is organized as follows. Section 2 provides brevity of Fujisaki model and qTA model. Section 3 presents two algorithms for fitting a given F_0 contour of a syllable or a two-syllable tone respectively. Experiment results are given in section 4. Conclusion is in section 5.

2. Related Work.

This issue requests algorithms to determine parameters for the representation of F_0 contours of word tones of the tonal languages such as Vietnamese, Mandarin or Thai and so on.

In the tonal languages, by distinguishing the meaning of a syllable and by tone sandhi in which the tones assigned to individual syllables change based on the pronunciation of adjacent syllables, one of the basic parameters of speech is the parameters generating the F_0 contour of the word.

2.1. Fujisaki model.

The Fujisaki model is a super positional model for representing F_0 contour of speech. According to the model, F_0 contour is generated as a result of the superposition of the outputs of two second order linear filters with a base frequency value. The second order linear filters are for generating the phrase and accent components of speech. The base frequency is the minimum frequency value of the speaker. In other words, F_0 contour is obtained by adding base frequency, phrase components and accent components.

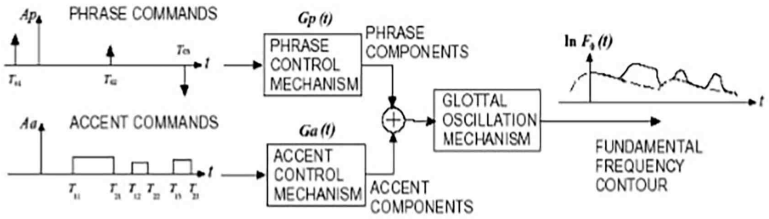


Figure 1. Fujisaki model.

Fujisaki model has many parameters which are described in the below formula, and currently, there is no numerical method to solve fitting problems when knowing a contour in advance.

$$\log F_0(t) = \log F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) + \sum_{j=1}^{J} A_{ai} \{G_a(t - T_{1j}) - G_a(t - T_{2j})\} \qquad (1)$$

, where

$$G_p = \begin{cases} \alpha^2 t.\exp(-\alpha), & for \;\; t \geq 0 \\ 0 \qquad\qquad\quad , & for \;\; t < 0 \end{cases}$$

$$G_a = \begin{cases} \min\left[1 - (1+\beta t).\exp(-\beta t)\right], & for \;\; t \geq 0 \\ 0 \qquad\qquad\qquad\qquad\quad , & for \;\; t < 0 \end{cases}$$

F_b : Baseline value of fundamental frequency
I : Number of phrase commandsJ : Number of accent commands
A_{pi} : Magnitude of i^{th} phrase command
T_{0i} : Timing of i^{th} phrase command
A_{aj} : Amplitude of j^{th} accent command
T_{1j} : Onset of j^{th} accent command
T_{2j} : Offset of j^{th} accent command
α : Natural angular frequency of the phrase control mechanism
β : Natural angular frequency of the accent control mechanism
γ : Relative ceiling level of accent components

In the Fujisaki model, as illustrated in Fig. 1, the shapes of local F_0 peaks and global F_0 trends are modeled as the on - and off -ramps of step and pulse responses of a second-order linear system. These responses are assumed to be associated with accent and phrase commands that are linguistically meaningful. Thus the commands, as the hypothetical underlying components of intonation, are different from the surface F_0 contour. In addition, the latter are the product the underlying commands generated by the articulatory mechanism implemented in the model. The surface F_0 contours are generated by a mechanism that compromises between maximum smoothness and full realization of the underlying tonal templates. Fujisaki model is also available for generating F_0 contours of any language such as Russian, English, Vietnamese and so on. For example, in [9, 10, 11] Hansjoerg Mixdorf et al already used the Fujisaki model to model Vietnamese F_0 contours of syllables in the phrase context.

In the Fujisaki model, F_0 contours are formed from the intonation contours and the stress contours. For tonal languages, this can lead to a change in the shape of the original tone in tones, such as flat tone being converted to another tone with the fundamental frequency value falling down due to the influence of the intonation contours. In addition, the Fujisaki model requires a lot of parameters to represent the F_0 contours. Therefore, it is not easy to calculate Fujisaki model parameters by fitting a given F_0 contour and until now there are no numerical computation methods to extract the parameters by fitting methods. The qTA model, which will be detailed in the next section, simulates F_0 contours as syllable-synchronized laryngeal movements toward underlying pitch targets that are either static or dynamic. Therefore, all these models are assumed that the surface F0 contours are the result of certain articulation mechanisms rather than from direct sound manipulation.

2.2. **qTA Model (Xu Model).** In the phrase context, by the occurrence of tone sandhi, the number of contour shapes of syllables increases many times over the isolated syllables. Therefore, it is not easy to model these variations.

In speech languages, for parameterizing F_0 contours of speech utterances, it is able to use the Xu model, namely qTA model [8]. This is a model that has been investigated and used by Xu et al to generate F_0 contours for tonal languages or non tonal ones such as Mandarin, Thai and English [12, 13].

Especially for tonal languages, tones can be analyzed into two components frequently combined: the pitch (the height of the base bar, referred to as the static characteristic)

and the tone (direction of the high-frequency change, called dynamic features) in the process of expression. Thus, each tone can be described as a combination of these two components. The static and dynamic characteristics can be modeled using the "pitch target" concept. Advantages of the model are simplicity, few parameters and can be learned statistically to generate the appropriate F_0 contours representation. More about recent results using the qTA representations can be found in [14, 15, 16, 17, 18, 19].

In details, to generate F_0 contours of tones, we will use qTA representations as follows:

$$F(t) \approx at + b + \alpha e^{-\lambda t} \tag{2}$$

or

$$F(t) \approx at + b + (ct^2 + dt + g)e^{-\lambda t} \tag{3}$$

The linear function $t \mapsto at + b$, called a "pitch target", reflects the tendency of the tone at the end of the F_0 contour.

The F_0 control is implemented through a third order critically damped linear system, in which the total response is the remain component given by formula (4), where the first linear term is the forced response of the system which is the pitch target and the second term is the natural response of the system. The transient coefficients c, d and g are calculated based on the initial F_0 dynamic state and the pitch target of the specified segment. The parameter λ represents the strength of the target approximation movement. The dynamic state is transferred from one syllable to the next at the syllable boundary to ensure continuity of F_0.

Compared to Mandarin and Thai tones, Vietnamese ones have more complex F_0 shapes [13, 20, 21], thus the representation formulas (2) and (3) should be replaced by one that can be a better model such complex tones. At the moment there are no numerical computation methods for estimating automatically the coefficients of each component of the model by fitting methods.

It is not easy to solve this problem because a suitable generated $t \mapsto F(t)$ contour must satisfy the two following conditions, given a sample of f_0 contour of the tone:

(C1) Pitch target constraint (PTC), $\{F(t) - (at + b)\} \approx 0$ with big enough time t.

(C2) Fitting constraint, $\{F(t) - f_0(t)\} \approx 0$ for any time t.

3. Proposed methods for estimating quantitative target approximation vectors.

In this section, we propose two algorithms for fitting a given F_0 contours of syllables or two-syllable tones based on qTA model respectively.

3.1. Representational vectors for a tone.

Definition 3.1. *(qTA vectors) Let tn be a given tone, vector* $v_{tn} = [a_{tn}, b_{tn}, k_{tn}, a_{P,0}, .., a_{P,Q-1}, a_{P,Q}]$ *is called a qTAV of tn if the function* F_{tn}*:*$t \mapsto F_{tn}(t) = a_{tn}t + b_{tn} + P_{tn,K}(t)k_{tn}^t$ *satisfies conditions (C1) and (C2) above, where* $P_{tn,K}(t) = \sum_{j=0}^{Q} a_{P,j}t^j$ *with given order* $Q \in \boldsymbol{N}$*,* $k_{tn} \in (0, 1)$*.*

The linear function $t \mapsto a_{tn}t + b_{tn}$ *will be a desired pitch target of tn. The following proposition confirms the soundness of qTA vector representation.*

Proposition 3.1. *Let* $t \mapsto f_0(t)$ *denote an* F_0 *contour of a given tone tn, where* $F_0(t)$ *is a continuous function on* $[T_1, T_2]$ *(*$0 < T_1 < T_2$*) with the pitch target* $t \mapsto at + b$*. Then, with some given* $k \in (0,1)$ *and any* $\epsilon > 0$*, there exists a polynomial P(t) such that* $\max_{t \in [T_1, T_2]} |f_0(t) - \{at + b + P(t)k^t\}| < \varepsilon$*.*

Proof.
Put $G(t) = \frac{F_0(t)-(a*t+b)}{k^t}$, G(t) is a continuous function on [T_1,T_2]. According to the Stone-Weierstrass theorem [21, 22], there exist a polynomial P(t) such that $\max_{t\in[T_1,T_2]} |G(t) - P(t)| < \varepsilon$, so $\forall t \in [T_1, T_2], |F_0(t) - at - b - P(t)k^t| < \varepsilon k^t < \varepsilon$ by 0<k<1, so that $\max_{t\in[T1,T2]} |F_0(t) - \{at + b + P(t)k^t\}| < \varepsilon$. □

For a given tone tn and an F_0 contour $t \mapsto f_0(t), 1 \le t \le T$ of tn, the objective function PTF is constructed to estimate a vector qTAV v_{tn} of tn and given as follows:

Denote $\Theta = \{Q, first, last, \beta, \gamma, k_{min}, k_{max} \mid 0 \le k_{min} < k_{max} \le 1, \beta > 0, \gamma > 0\}$ are experimental parameters,

$$\begin{aligned} &PTF_\Theta(v_{tn}) \stackrel{def}{=} \beta * PTC + \gamma * FC \to \min, \ where \\ &PTC_\Theta \stackrel{def}{=} \sum_{T-last+1\le t\le T\wedge f_0(t)\neq NaN} \{f_0(t) - (a_{tn}t + b_{tn})\}^2, \\ &FC_\Theta \stackrel{def}{=} \sum_{1\le t\le T\wedge f_0(t)\neq NaN} \left\{f_0(t) - \left(a_{tn}t + b_{tn} + P_{tn,K}(t)k_{tn}^t\right)\right\}^2, k_{min} \le k \le k_{max} \ . \end{aligned} \tag{4}$$

We propose a method with two stages to estimate qTAVs. For first stage, to determine the pitch target component, we will find the line connecting the beginning and the end points of the F_0 contour to obtain the initial values a and b. Then, using an algorithm that solves the objective function PTF (formula (4)) where k is variant in the narrower interval than [0,1] (e.g, $k_{min} = 0.05$, $k_{max} = 0.85$).

For second stage, by reinitializing the component k and the other components of $\mathbf{v}_{tn}$ and resolving the objective function PTF but k is variant in [0,1] ($k_{min} = 0$,$k_{max} = 1.0$), we already obtain the values of components of $\mathbf{v}_{tn}$.

Algorithm: SPTF V Estimating qTAV of a given tone tn:
$v_{tn} = SPTF_\Theta\left(\{f(t)\}_{\overline{t=1,T}}\right)$

Input: a given F_0 contour $\{f_0(t)\}_{1\le t\le T}$ of tn, F_0(t) = NaN if t-th frame belongs to the unvoiced part.
Parameters: Θ ={Q, first, last, β, γ, k_{min},k_{max} | Q $\in \mathbf{N}$, $0 < k_{min} < k_{max} < 1$, $\beta > 0$, $\gamma > 0$}.
Output: the qTAV v_{tn} which chosen to optimize the objective function PTF.
Step 1: Estimate a_1, b_1 such that:

$$\sum_{1\le t\le first\wedge f_0(t)\neq NaN} \{f_0(t) - (a_1t + b_1)\}^2 + \sum_{T-\text{last}\le t\le T\wedge f_0(t)\neq NaN} \{f_0(t) - (a_1t + b_1)\}^2 \to \min.$$

Step 2: Initialize the components of v_{tn}.
2.1: a = a_1, b = b_1 and k = k_{max}.
2.2: Initialize other components by small random values.
Step 3: Repeat to solve the PTF: $PTF_\Theta(v_{tn}) \to \min$.
Step 4: Estimate v_{tn} such that: $PTF_{\Theta'}(v_{tn}) \to \min$
, where Θ'={Q, first, last, β, γ, k'$_{min}$= 0, k'$_{max}$ = 1} .
4.1: Reinit the components of $\mathbf{v}_{tn}$ as follows:
Keep values of a and b, but reassign k =1.0, and initialize other parameters of by small random values.
4.2: Repeat to solve the PTF: $PTF_{\Theta'}(v_{tn}) \to \min$.
Step 5: Stop and return v_{tn}.

3.1.1. *Estimating qTAVs of two-syllable tones.* Given an F_0 contour of two-syllable tones tn_1 and tn_2. Without loss of generality, we may assume that the F_0 contour is represented by two functions $\{f_0^1(t)\}_{T_0\le t\le T_1}$ and $\{f_0^2(t)\}_{T_2\le t\le T_3}$ of tn_1 and tn_2 respectively, where $\mathrm{T}_1\le\mathrm{T}_2$, and

$$\begin{aligned} &\arg\min_t \left\{ f_0^1(t) \neq NaN \right\} = T_0 = 1, \\ &T_1 = \arg\max_t \left\{ f_0^1(t) \neq NaN \right\}, \\ &T_2 = \arg\min_t \left\{ f_0^2(t) \neq NaN \right\}, \\ &T_3 = \arg\max_t \left\{ f_0^2(t) \neq NaN \right\}. \end{aligned} \tag{5}$$

We need to estimate qTAVs $v_{tn_1} = [a_{tn_1}, b_{tn_1}, k_{tn_1}, a^1_{P,0}, .., a^1_{Q_1-1}, a^1_{Q_1}]$ of tn_1 and $v_{tn_2} = [a_{tn_2}, b_{tn_2}, k_{tn_2}, a^2_{P,0}, .., a^2_{Q_2-1}, a^2_{Q_2}]$ of tn_2 such that the conditions (C1) and (C2) above are satisfied simultaneously for the $\{f_0^1(t)\}_{1\le t\le T_1}$ and $\{f_0^2(t)\}_{T_2\le t\le T_3}$.

To solve this problem, we can use the algorithm SPTF again for each tone tn_1 and tn_2. But at first, we need to connect the two functions $\{f_0^1(t)\}_{1\le t\le T_1}$ and $\{f_0^2(t)\}_{T_2\le t\le T_3}$ to construct a continuous function $t \mapsto \overline{f_0^2}(t)$ on $[1,\ \mathrm{T}_3]$ as follows:

$$\begin{aligned} &\overline{f_0^2}(t) \overset{def}{=} f_0^1(t), \forall t \in [1, \mathrm{T}_1], \\ &\overline{f_0^2}(t) \overset{def}{=} (1 - \tanh(m * (t - T_1)))\, f_0^1(T_1) + \tanh(m * (t - T_1)) * f_0^2(T_2), \forall t \in (T_1, T_2), \\ &\overline{f_0^2}(t) \overset{def}{=} f_0^2(t), \forall t \in [\mathrm{T}_2, \mathrm{T}_3],\ \textit{where } m > 0. \end{aligned} \tag{6}$$

Then, the timing onset T_2 of tn_2 will be pulled back to T_c, where

$$T_c \overset{def}{=} 1 + [(T_1 + T_2)/2]\ \ if\ f_0^1(T_1) \neq f_0^2(T_2),\ else\ T_c \overset{def}{=} T_2. \tag{7}$$

Lastly, we have the following algorithm:

Algorithm: DSPTF - Estimating qTAVs of a given two - syllable tones.

Input: Given F_0 contours $\{f_0^1(t)\}_{1\le t\le T_1}$ and $\{f_0^2(t)\}_{T_2\le t\le T_3}$ of two-syllable tones tn_1 and tn_2 respectively.
Parameters: Θ ={m, Q_1, Q_2, first, last, β, γ, k_{min}, k_{max} | m >0, Q_1, $\mathrm{Q}_2 \in \mathbf{N}$, $0 <$ $\mathrm{k}_{min} <$ $\mathrm{k}_{max} < 1$, $\beta > 0$, $\gamma > 0$}.
Output: qTAVs v_{tn_1},v_{tn_2}, and a new F_0 generated $\left\{f_0^{2,new}(t)\right\}_{1\le t\le T_3}$.
Step 1: Calculates T_c by formula (7), $\overline{f_0^2}(t)$, $t = \overline{T_1, T_2}$ by formula (6).
Step 2: Calculates v_{tn_1}: $v_{tn_1} = SPTF_{\Theta_1}\left(\{f_0^1(t)\}_{\overline{t=1,T_1}}\right)$, where Θ_1 ={ Q_1, first, last, β, γ , k_{min}, k_{max}}.
Step 3: Calculates v_{tn_2}:

$$v_{tn_2} = SPTF_{\Theta_2}\left(\left\{\overline{f_0^2}(t + T_c - 1)\right\}_{\overline{t=1,T_3-T_c-1}}\right),$$

where Θ_2 ={ Q_2, first, last, β, γ , k_{min}, k_{max}}.
Step 4 (optional step): Calculates $\left\{f_0^{2,new}(t)\right\}_{1\le t\le T_3}$,

$$\begin{aligned} &f_0^1(t) \neq NaN : f_0^{1,new}(t) = a_{tn_1} * t + b_{tn_1} + (k_{tn_1})^t P_{K,tn_1}(t),\ t = \overline{1, T_1}, \\ &f_0^2(t) \neq NaN : t' = t - T_c + 1, f_0^{2,new}(t) = a_{tn_2} * t' + b_{tn_2} + (k_{tn_2})^{t'} P_{K,tn_2}(t'),\ t = \overline{T_2, T_3}. \end{aligned} \tag{8}$$

Return: v_{tn_1} ,v_{tn_2} , $\left\{f_0^{2,new}(t)\right\}_{1\le t\le T_3}$.

4. Experiment result.

4.1. **Experimental data.** The speech input for the experiment on our proposed algorithms is in Vietnamese, a monosyllabic and tonal language with six tones (see Tab. 1) that has the most complex lexical tones compared to other tonal languages.

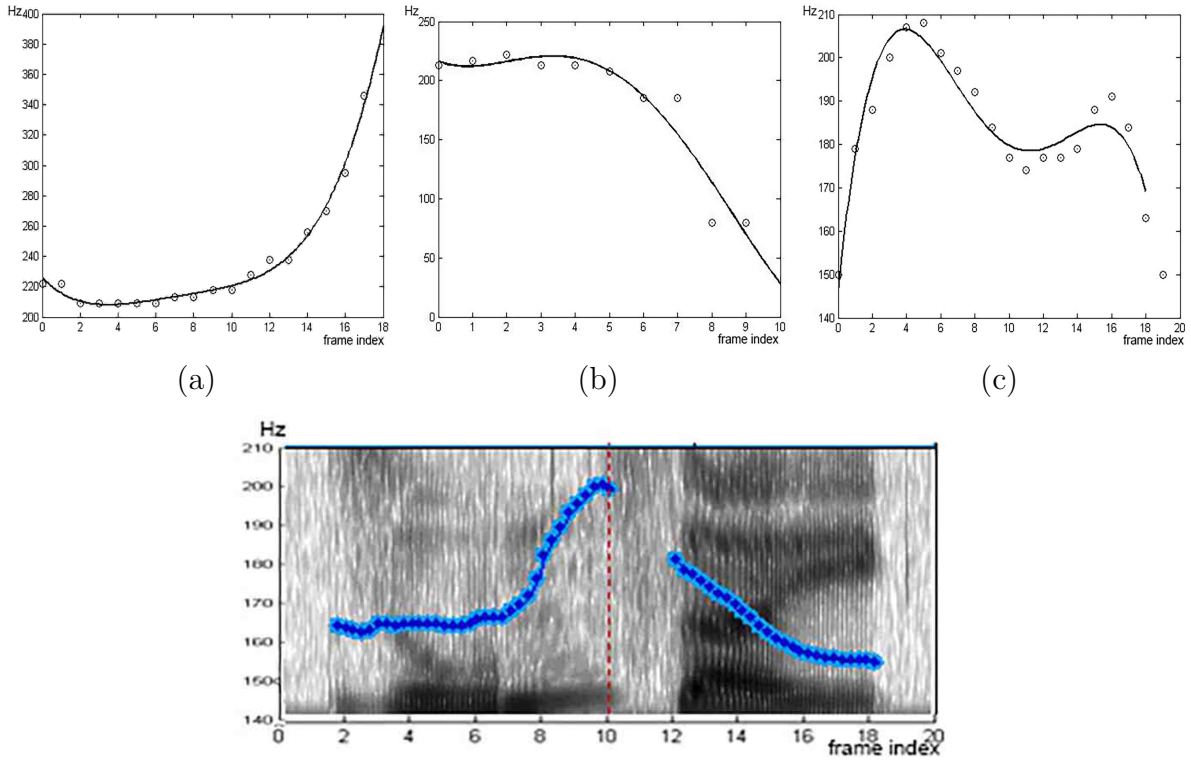


Figure 2. The typical F_0 contours shape of some tones of Vietnamese isolated syllables (a) Rising tone (b) Broken tone (c) Drop tone and (d) An F_0 contour of the word /dún/dảy/ (/zu↑/ zạj ↘┤/) with tone sandhi.

In order to test the algorithms, a single speaker story reading corpus was created, uttered by a female speaker of standard Vietnamese. The entire size of the corpus is 5 hours. There are 567 sentences, 7,724 syllables and 23,482 phonemes in the corpus. The recording was cut manually at the sentence level and then words with two syllables created [9].

4.2. **The experiment of calculating qTAVs.** For algorithm 1, we choose first = 2, last = 2, $k_{min} = 0.05$, $k_{max} = 0.85$, $\beta = 1.0$, $\gamma = 3.0$ and for the algorithm 3, the connection parameter m = 0.3. Specifically, the order polynomials Q for qTAVs of Vietnamese tones selected from Tab. 1 below.

In [23] the authors used PENTATrainer1 which is a Praat script for analyzing and synthesizing intonation based on the PENTA model.

The PENTATrainer1 requires manually, entering onset and offset timing for each tone, especially with the tones that change the direction of F_0 contours (such as, the Broken tone) we have to enter two couple of onset, offset timing for each tone. Also, the result depends on the accuracy of the onset, offset timings.

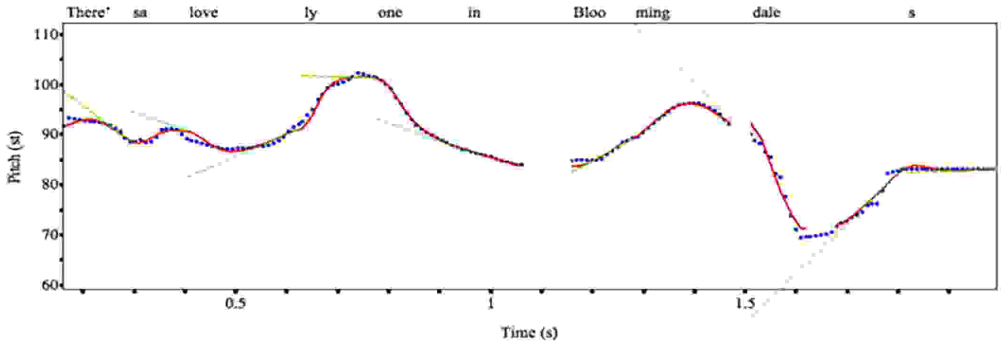
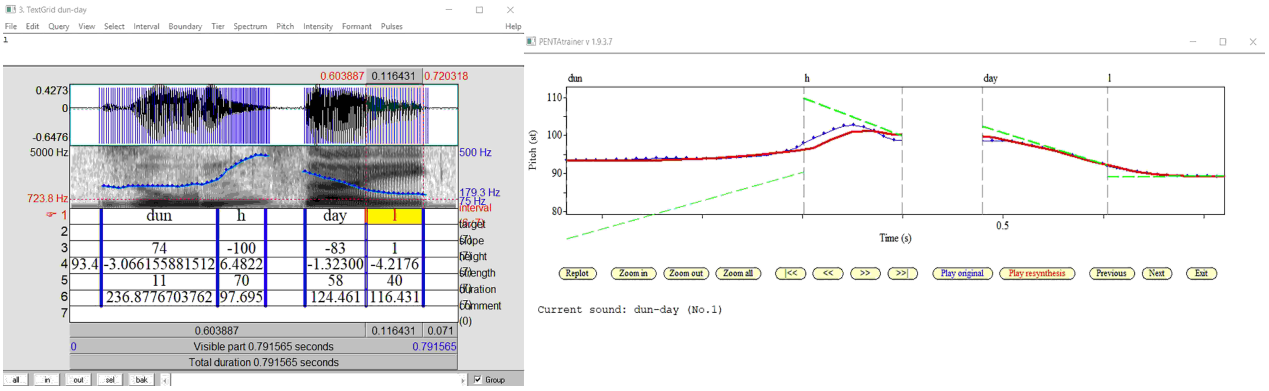
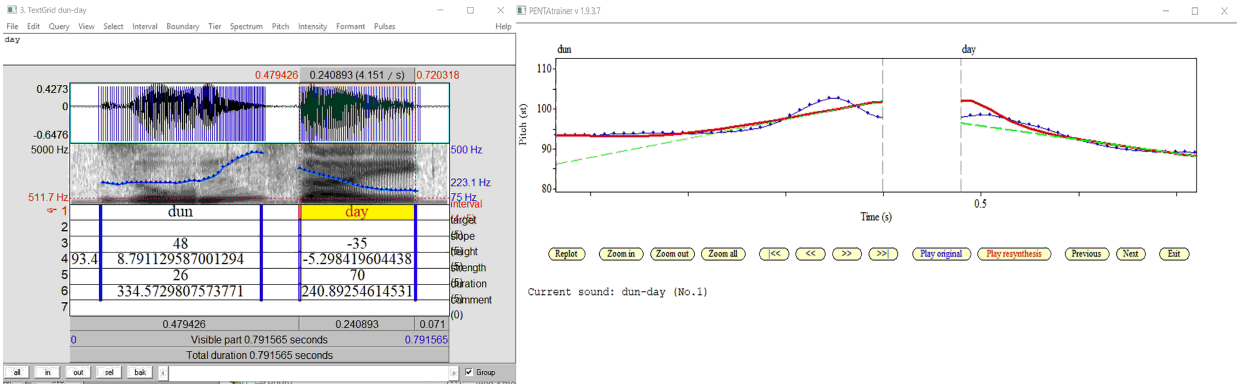


Figure 3. A generated F_0 contours by PENTATrainer1 [20].



(a)



(b)

Figure 4. Generated F_0 contours of two-syllable tones *dún*/*dẩy* (/zu˥/ za̰j ˧˩˧)by PENTA tool where onset and offset time are given by manual, (a) Correct onset and offset timings and (b) Incorrect onset and offset timings.

However, in contrast to PENTATrainer1, our DSPTF algorithm is automatic when it comes to extracting all parameters of a qTAV for each instance of a given tone. To model the Vietnamese tones, the input parameters for the DSPTF algorithm are given by Tab. 1.

Moreover, in order to select the optimal parameters in step 3 and step 4.2 of the SPTF algorithm, we can use Matlab's fminsearch procedure [13]. By using an auxiliary variable $\overline{k} \in R$, such that

$$k = k_{\min} + \frac{k_{\max} - k_{\min}}{1 + \overline{k}^2} \in [\mathrm{k}_{\min}, \mathrm{k}_{\max}] \tag{9}$$

Table 1. The order Q, a and k coefficients of qTAVs of Vietnamese tones with different F_0 contour shapes (Outline, trend/ duration/direction), 2' and 6' are 2 and 6 tones ending with stop consonants respectively [10, 11].

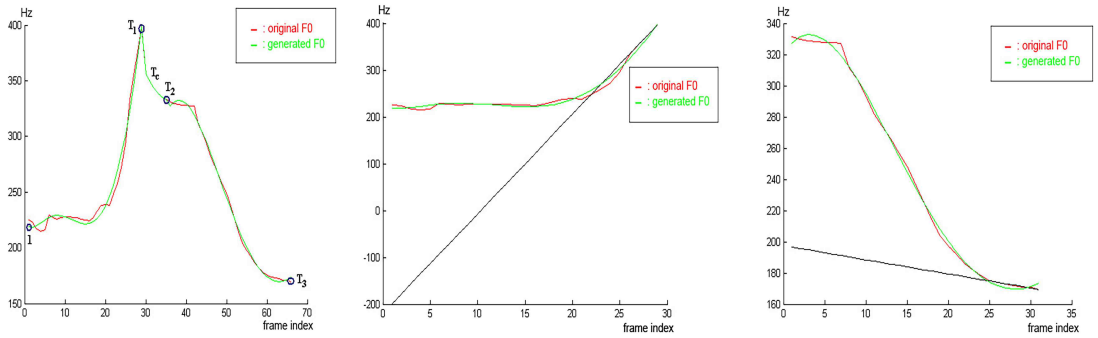
Tone index (IPA)	Example	F_0 shape	Q	a	k
1(˧˧)	Ba (kwa̰:˧˧, father)	Level/long/no change	0	No prior	0
2(˧˥)	Má (ma:˧˥, mother)	Rising/long/change	4	No prior	No prior
2'(˧˥)	Mắt (mat˧˥, eye)	Level/short/no change	0	0	0
3(ʔa˧˥)	Mã (maʔa˧˥, horse)	Broken/long/change	4	No prior	No prior
4(˨˩˧)	Quả (kwa̰:˨˩˧, fruit)	Curve/medium/no change	4	No prior	No prior
5(˨˩)	Mà (ma̤:˨˩, but)	Falling/medium/no change	0	No prior	0
6(ʔ˨˩)	Quạ (kwa̰ːʔ˨˩, crows)	Drop/short/change	4	No prior	No prior
6'(ʔ˨˩)	Mặt (ma̰ʔt˨˩, face)	Level/short/no change	0	0	0

, then (4) will be transformed into an optimal problem without any constraint for $\overline{k}$.

The iterative process of the objective function is convergent. This is achieved by the second fitting steps of the SPTF algorithm when selecting the appropriate starting point for the variable values before entering the target function optimization loop.

For demonstration the suitable results of qTAVs obtained by algorithm DSPTF, two-syllable tones with complex shape of F_0 contours (see Tab. 1) such as /tone 2/tone 4/, /tone 2/tone 3/ and /tone 2/tone 6/ are selected to analyze.

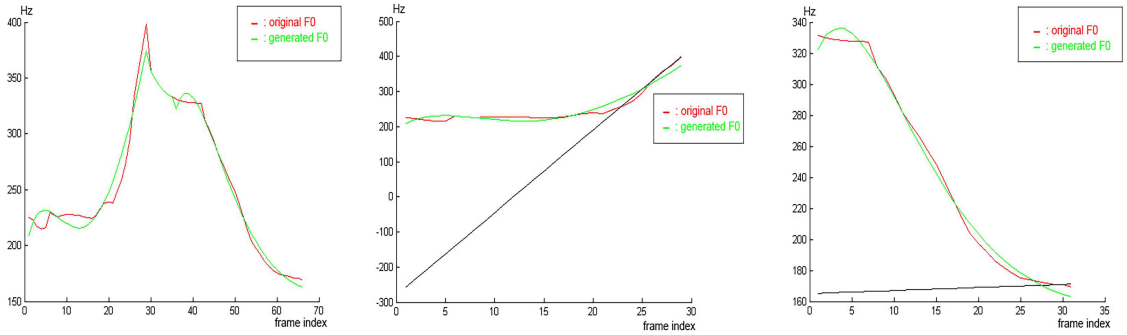
Figs. 5 and 6 show the results obtained by DSPTF algorithm with the same two-syllable tones *dún/dẩy* (/zu˧˥/ za̰j˨˩˧), (tone 2 and tone 4), where we use $Q_2 = Q_4 = 4$ for qTAVs (see Tab. 1 above) and $Q_2 = Q_4 = 2$ (see the formula (3), or [23]) respectively.



(a) (b) (c)

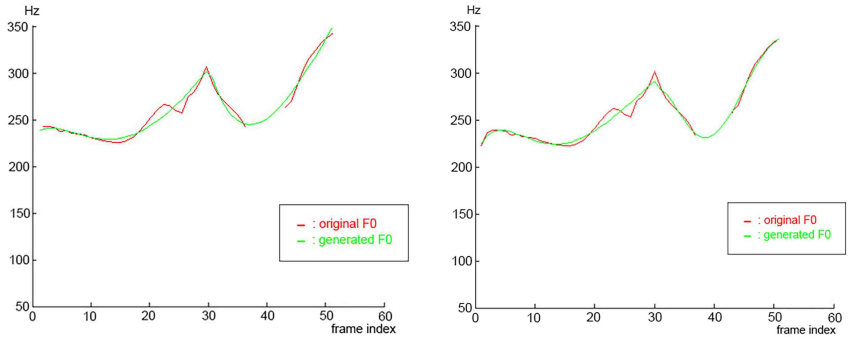
Figure 5. For $Q_2 = Q_4 = 4$ (a) Generated F_0 contours by DSPTF, (b) PT line's 2 and (c) PT line's tone 4.

Clearly, the PTs and F_0 contours generated by the DSPTF algorithm are sharper, and using the higher orders Q(Q>2, see Tab. 1) for the qTAVs will give better fitting PT lines and F_0 contours of two-syllable tones than Q = 2. Fig. 7 also illustrates the results with the same two-syllable tones *đấu/võ* (ɗəw˧˥/ vɔʔɔ˧˥), tone 2 and tone 3, *"fighting martial arts"*).



(a) (b) (c)

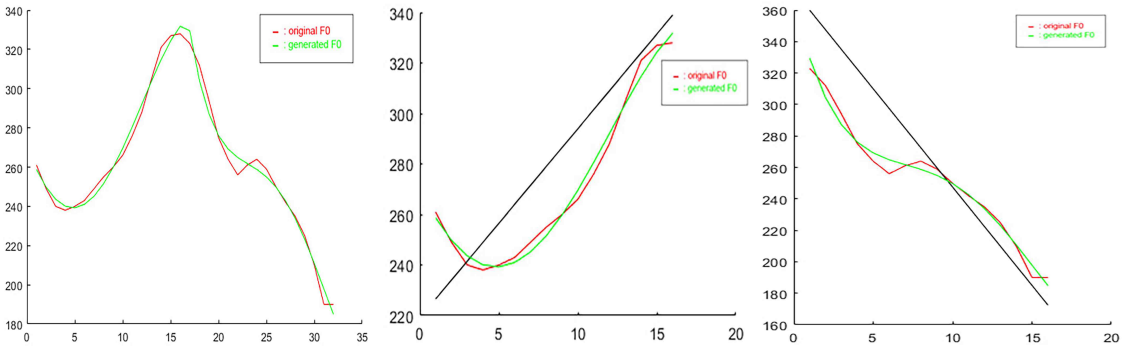
Figure 6. For $Q_2 = Q_4 = 2$ (a) Generated F_0 contours by DSPTF, (b) PT line's 2 and (c) PT line's tone 4.



(a) (b)

Figure 7. Generated F_0 contours by DSPTF, (a) For $Q_2 = Q_3 = 2$ and (b) For $Q_2 = Q_3 = 4$.

Figs. 8 shows the results obtained by DSPTF algorithm with two-syllable tones /tiếng/động/ (sound, IPA /tiəŋˊ ɗʂwŋˋˊ/, tone 2 followed by tone 6), where we use Q2 = Q4 = 4 for qTAVs.



(a) (b) (c)

Figure 8. (a) Generated F_0 contours by DSPTF, (b) PT line's 2 and (c) PT line's tone 6.

In short, the results illustrated in Figs. 5-8 already show that the parameter qTAVs are suitable for representing two-syllable tones. Pitch target of each tone is quite stable and the whole F_0 contours are fit well.

5. **Conclusion.** The qTA model is focused to consider the tones and pitch accents as abstract units called pitch targets. In tonal languages, pitch targets are able to be separated into static targets-[high] and [low], and dynamic ones-[rise] and [fall], which are associated with lexical tones respectively. This model proposed by Yi Xu et al gives a more accurate description of lexical tone variations in the syllable than the Fujisaki model. However, the qTA model needs labels on the onset and offset of the pitch target, and is not easy to implement on training speaker-dependent prosodic styles without automatic numerical methods.

In this paper, we have proposed a new computational method to determine quantiative target approximation vectors (qTAV) that generate the F_0 contours of two-syllable tones. Our method include a numerical solution. The first and second algorithms are explicit calculations to estimate the representational parameters of F_0 contours of one or two-syllable tones. Like qTA of previous Xu model, these parameters are used by qTAV including the linear target coefficients and the approximation polynomial coefficients. The experiments also show the effectiveness of the proposed methods when generating the F_0 contours of two-syllable tones with complex shapes as those in the Vietnamese language. The F_0 contours generating method by qTA vectors is highly generalized, so in our next studies we will expand the results to predict qTAV parameters and generate the F_0 contours of multi syllable tones.

REFERENCES

[1] M. D. Kovacs, M. Y. Cho, P. F. Burchett and M. Trambert, Benefits of integrated ris/pacs/reporting due to automatic population of templated reports, *Current Problems in Diagnostic Radiology*, vol. 48, no. 1, pp. 37-39, 2019.

[2] freedomscientific, *Job Access With Speech, screen reader.*

[3] Santiham Prom-on, F. Liu and Y. Xu, Functional modeling of tone focus and sentences type in mandarin Chinese, *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)* , pp. 1638-1641, 2011.

[4] G. Bailly and B. Holm, SFC: A trainable prosodic model, *Speech Communication*, vol. 46, no. 3-4, pp. 348-364, 2005.

[5] H. Fujisaki, Dynamic characteristics of voice fundamental frequency in speech and singing, *The Production of Speech*, pp. 39-55, edited by P.F. MacNeilage Springer-Verlag, New York, 1983.

[6] G. Kochanski and C. Shih, Prosody modeling with soft templates, *Speech Communication*, vol. 39, no. 3-4, pp. 311-352, 2003. 39, 2003.

[7] H. Fujisaki and K. Hirose, Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *Journal of the Acoustical Society of Japan (E)*, vol. 5, no. 4, pp. 233-242, 1984.

[8] Y. Xu and Q. E. Wang, Pitch targets and their realization: Evidence from Mandarin Chinese, *Speech Communication*, vol. 33, no. 4, pp. 319-337, 2001.

[9] H. Mixdorf, N. T. Dung, L. C. Mai, N. H. Huy, and V. K. Bang, Toward integrating the Fujisaki model into Vietnamese TTS, *Proceeding of the International Conference on Spoken Language*, pp. 177-180, Processing, Korea, 2004. Processing, Korea, 2004.

[10] H. H. N. B. K. V. H. M. Dung Nguyen, Chi Mai Luong, Fujisaki model based F0 contours in vietnamese TTS, *INTERSPEECH 2004-ICSLP, 8th International Conference on Spoken Language Processing*, Jeju Island, Korea, October 4-8, 2004.

[11] H. Mixdorff, N. H. Bach, H. Fujisaki and M. C. Luong, Quantitative Analysis and Synthesis of Syllabic Tones in Vietnamese, *Eurospeech 2003 - Interspeech 2003, 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 1-4, 2003.

[12] Y. Xu and S. Prom-on, Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning, *Speech Communication*, vol. 57, pp. 181-208, 2014.

[13] Mathworks, Optimizing Nonlinear Functions - MATLAB and Simulink.

[14] S. Prom-on and Y. X., The qTA Toolkit for Prosody: Learning underlying parameters of communicative functions through modeling, *Proc. 5th International Conference of Speech Prosody*, Chicago, USA, 2010.

[15] Y. Li, J. Tao, W. Lai and X. Xu, Quantitative intonation modeling of interrogative sentences for Mandarin Speech Synthesis, *Speech Communication*, vol. 89, pp. 92-102, 2017.
[16] L. Jiao and Y. Xu, Whispered mandarin has no production-enhanced cues for tone and intonation, *Lingua*, vol. 218, pp. 24-37, 2019.
[17] B. Wang, Y. Xu and Q. Ding, Interactive prosodic marking of focus, boundary and newness in Mandarin, vol. 75, no. 1, pp. 24-56, 2018.
[18] T. Y. Ta, H. V. Nguyen, T. V. Dao, H. Ngo and A. Sergey, An effective algorithm for determining pitch markers of vietnamese speech sentences, *Advances in Neural Networks-ISNN 2018*, pp. 628-636, 2018.
[19] M. Brookes, VOICEBOX: Speech Processing Toolbox for MATLAB.
[20] Y. Xu and S. Prom-on, A Praat script for automatic analysis and synthesis of intonation based on the PENTA model, working on individual sound files.
[21] M. H. Stone, The generalized weierstrass approximation theorem, *Mathematics Magazine*, vol. 21, no. 5, pp. 237-254, 1948.
[22] J. L. P. Henryk Gzyl, The weierstrass approximation theorem and large deviations, *The American Mathematical Monthly*, vol. 104, no. 7, pp. 650-653, 1997.
[23] Y. Xu and S. Prom-on, Articulatory-functional modeling of speech prosody: A review, *11th Annual Conference of the International Speech Communication Association*, vol. 1, pp. 46-49, 2010.