УДК 681.3

СИНТЕЗ РЕЧИ ТОНАЛЬНЫХ ЯЗЫКОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ НЕПРЯМЫХ МАРКЕРОВ И КОЛИЧЕСТВЕННОГО ПРИБЛИЖЕНИЯ ЦЕЛИ

***Т. Й. ТХАЙ*[1)], *Х. Н. ХУИ*[2)], *Д. В. ТУИЕТ*[3), 4)],**
***С. В. АБЛАМЕЙКО*[3)], *Н. В. ХУНГ*[5)], *Д. В. ХОА*[5)]**

[1)]*Ханойский университет бизнеса и технологий,*
ул. Вин Туи, 29А, Вин Туи, Хай Ба Трунг, г. Ханой, Вьетнам
[2)]*Университет электроэнергии Министерства промышленности и торговли Вьетнама,*
ул. Хоанг Куок Вьет, 235, 129823, Ко Нуэ, Ту Лием, г. Ханой, Вьетнам
[3)]*Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Беларусь*
[4)]*Университет Бинь Зыонг, пр. Бин Дуонг, 504, 820000, г. Тху Дау Мот, район Бин Дуонг, Вьетнам*
[5)]*Военный институт науки и техники, ул. Хоанг Сэм, 17, Нгхиа Ду, Кау Гиау, г. Ханой, Вьетнам*

Синтезирующие тоны играют важную роль в системах преобразования текста в речь тональных языков. Для этого необходимо выполнить два важных шага: определить маркеры высоты тона голосовых высказываний и синтезировать траектории F_0 для лексических тонов. В этой статье мы предлагаем два эффективных алгоритма, один из которых заключается в расположении маркеров высоты тона на пиках кумулятивного сигнала каждой озвученной части входного высказывания, а другой – в генерации F_0-траекторий тонов с количественными параметрами приближения цели (qTA). Эксперимент показал, что предложенные алгоритмы представляют маркеры высоты звука с высокой точностью, что позволило нам генерировать тоны со сложной формой.

Ключевые слова: маркеры основного тона; кумулятивный сигнал; модель Сюй; qTA; полиномиальное приближение.

Образец цитирования:
Тхай ТЙ, Хуи ХН, Туиет ДВ, Абламейко СВ, Хунг НВ, Хоа ДВ. Синтез речи тональных языков с использованием методов непрямых маркеров и количественного приближения цели. *Журнал Белорусского государственного университета. Математика. Информатика.* 2019;3:105–121 (на англ.).
https://doi.org/10.33581/2520-6508-2019-3-105-121

For citation:
Thai TY, Huy HN, Tuyet DV, Ablameyko SV, Hung NV, Hoa DV. Tonal languages speech synthesis using an indirect pitch markers and the quantitative target approximation methods. *Journal of the Belarusian State University. Mathematics and Informatics.* 2019;3:105–121.
https://doi.org/10.33581/2520-6508-2019-3-105-121

Авторы:
Та Йен Тхай – лектор на факультете информатики.
Хоан Нго Хуи – кандидат наук (информатика); заместитель декана факультета информатики.
Дао Ван Туиет – старший исследователь Центра биомедицинской информатики[4)]; аспирант кафедры веб-технологий и компьютерного регулирования механико-математического факультета[3)]. Научный руководитель – С. В. Абламейко.
Сергей Владимирович Абламейко – академик НАН Беларуси, доктор технических наук, профессор; профессор кафедры веб-технологий и компьютерного регулирования механико-математического факультета.
Нгуен Ван Хунг – кандидат наук (информатика); лектор на факультете информатики.
Доан Ван Хоа – кандидат наук (информатика); лектор на факультете информатики.

Authors:
Ta Yen Thai, lecturer at the faculty of informatics.
thaity@hubt.edu.vn
Hoang Ngo Huy, PhD (informatics); vice dean of the faculty informatics.
huynh@epu.edu.vn
Dao Van Tuyet, senior researcher at the Biomedical Informatics Center[d] and postgraduate student at the department of web-technologies and computer simulation, faculty of mechanics and mathematics[c].
daovt@bsu.by
Sergey V. Ablameyko, academician of the National Academy of Sciences of Belarus, doctor of science (engineering), full professor; professor at the department of web-technologies and computer simulation, faculty of mechanics and mathematics.
ablameyko@bsu.by
Nguyen Van Hung, PhD (informatics); lecturer at the faculty of informatics.
nvhnt73@gmail.com
Doan Van Hoa, PhD (informatics); lecturer at the faculty of informatics.
doanvanhoa@gmail.com

Журнал Белорусского государственного университета. Математика. Информатика. 2019;3:105–121
Journal of the Belarusian State University. Mathematics and Informatics. 2019;3:105–121

TONAL LANGUAGES SPEECH SYNTHESIS USING AN INDIRECT PITCH MARKERS AND THE QUANTITATIVE TARGET APPROXIMATION METHODS

***T. Y. THAI*[a], *H. N. HUY*[b], *D. V. TUYET*[c, d],**
***S. V. ABLAMEYKO*[c], *N. V. HUNG*[e], *D. V. HOA*[e]**

[a]*Hanoi University of Business and Technology, 29A Vinh Tuy Street, Vinh Tuy Ward, Hai Ba Trung Dist, Hanoi, Vietnam*
[b]*Electric Power University, Vietnam Ministry of Industry and Trade, 235 Hoang Quoc Viet Street, Co Nhue, Tu Liem, Hanoi 129823, Vietnam*
[c]*Belarusian State University, 4 Niezaliežnasci Avenue, Minsk 220030, Belarus*
[d]*Binh Duong University, 504 Binh Duong Avenue, Thu Dau Mot Town 820000, Binh Duong Province, Vietnam*
[e]*Military Institute of Science and Technology, 17 Hoang Sam Street, Nghia Do Ward, Cau Giay District, Hanoi, Vietnam*
Corresponding author: D. V. Tuyet (daovt@bsu.by)

Synthesizing tones plays an important role in text-to-speech systems of tonal languages. To accomplish this, the two important steps are to determine the pitch markers of voice utterances and synthesize F_0 trajectories for lexical tones. In this paper, we propose two efficient algorithms, one of them is to locate the pitch markers at the peaks of the cumulative signal of each voiced part of the input utterance and the other is to generate F_0 trajectories of tones with quantitative target approximation (qTA) parameters of Xu model. The experimentation has shown that the proposed algorithms present pitch markers with high accuracy which has enabled us to generate tones with complex shapes.

Keywords: pitch markers; cumulative signal; Xu model; qTA; polynomial approximation.

Introduction

Nowadays, text to speech (TTS) systems and speech to text (STT) systems are increasingly used by the radiologist to create radiology study reports. Besides, TTS systems and STT systems can help people with disabilities integrate into the community by using computer easier. With the integration of the laboratory information system (LIS) and radiological information systems (RIS) patient identifiers and examination information can automatically map into examination reports. There are many potential benefits of report automation to radiologists including improvements in efficiency, accuracy, and fatigue [1].

Besides, TTS systems can help people with disabilities integrate into the community by using computer easier. For example the JAWS (job access with speech) software, 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. JAWS users around the world sent us videos about the impact JAWS has made on their lives [2].

Due to the application needs, the research on speech representation has been increasingly developed, the issues of research on estimation and modeling of fundamental frequency trajectories is still open research issues until now.

Frequency relates to the individual pulsations produced by vocal cord vibrations for a unit of time. The rate of vibration depends on the length, thickness, and tension of the vocal cords, and thus is different for child, adult male and female speech. A speech sound contains an important type of frequencies namely fundamental frequency (F_0) which relates to vocal cord function and reflects the rate of vocal cord vibration during phonation (pitch).

Pitch markers (PM) play a central role in phonetics signal analytic because pitch is a big part of hearing music, we can be tricky sounds without clear F_0. In addition PM is also useful for coding or representation for extracting information of speech for telephony and communication.

The fundamental frequency, pitch markers and hearing quality. The fundamental frequency is the primary element of speech signal and because the pitch marker indicates the beginning of each cycle of the waveform, PM plays a very important role in generating and recognizing speech sentences. However, pitch is an inherently subjective quantity and cannot be directly measured from the speech signal. It is a nonlinear function of the signal's spectral and temporal energy distribution. Therefore, PM estimation is an unsolved and challenging problem. It is one of the key technologies that determines the performance of speech processing.

Autocorrelation method or average magnitude difference function (AMDF) is commonly used. In addition, modified autocorrelation method [3] is also commonly used to compute the auto correlation instead of speech signal [2]. However, these methods suffer from error estimation in noisy environment. Robust algorithm for pitch tracking (RAPT) is well-known and widely used F_0 estimation method since it does offer low delay, low computational amount and robust against noise [4].

The YIN [5] algorithm uses a difference function based on the autocorrelation function as the candidate generator in conjunction with a number of optimization steps. Named after the oriental yin-yang principle of duality, it aims to balance between the autocorrelation and the cancelation that it involves.

The dynamic programming projected phase-slope algorithms (DYPSA) [6] was originally designed for automatic estimation of glottal closure instants (GCIs) in voiced speech but as a consequence also gives pitch information. The algorithm is based on an enhancement of the group delay algorithm [7] by R. Smits and B. Yegnanarayana, which is used as the primary candidate generator. DYPSA uses dynamic programming to identify the best GCI candidates by minimizing some cost functions. The DYPSA algorithm operates on the speech signal alone and does not require an electroglottography reference signal. The pitch estimate is derived from the inter GCI duration and mapped into frames.

F_0 trajectory representation and analysis-by-synthesis. From a modeling perspective, a model is of little use 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 tone, focus and sentence type and their interactions are directly behind the complex surface F_0 trajectories in Mandarin, 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 surface F_0 trajectories 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* [8].

Parametric representation of speech often implies F_0 trajectory 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 [9–12]. These approaches can be divided into two general categories, namely, those that model F_0 trajectories 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 trajectories, with minimal consideration about the articulatory process of F_0 production.

The Fujisaki model is an effective model for approximating the trajectory of the fundamental frequency precisely for the source model of speech synthesis, representing the coarticulation of spectral frequencies making an equation for a target model of speech perception and so on [10–12].

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. Thus for improving our understanding of human speech, quantitative modeling is also indispensable. In the present paper we report the results of an attempt to simulate tone, stress, and focus in Mandarin and English with a quantitative model that generates surface F_0 trajectories through the process of target approximation TA [13]. qTA model for generating F_0 trajectories of speech. 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 communicative functions directly modeled are lexical tone in Mandarin and lexical stress in English and focus in both 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 trajectories generated with the extracted parameters to those of natural utterances through numerical evaluation and perceptual testing. The F_0 trajectories generated by the qTA model with the learned parameters were very close to the natural trajectories in terms of root mean square error, rate of human identification of tone, and focus and judgment of naturalness by human listeners.

qTA and improving for generating F_0 trajectories of words with complex shape. In the detail, to generate F_0 trajectories of tones, we are able to use Xu model, which has been widely used for Mandarin [14; 15] to model the F_0 trajectories in the context:

$$f_0(t) \approx at + b + (ct^2 + dt + g)e^{-\lambda t}. \quad (1)$$

The linear function $t \mapsto a_{tn}t + b_{tn}$, called a «pitch target», reflects the tendency of the tone at the end of the F_0 trajectory.

The computational model used in the present study is the quantitative target approximation (qTA) model. This model simulates the production of tone and intonation as a process of syllable-synchronized sequential target approximation [15; 16]. Figure 2 illustrates the basic idea of target approximation [15]. The qTA model represents F_0 as the surface response of the target approximation process which is driven by pitch targets. A pitch target is a forcing function representing the joint force of the laryngeal muscles that control vocal fold tension. It is represented by a simple linear equation $x(t) = a^*t + b$ given by the formula (1).

Compared to Mandarin, Thai and Vietnamese tones have more complex F_0 shapes [17–20], thus the representation formula (1) should be replaced with one that can better model such complex tones. In [21], the authors present a Thai tone model based on qTA method. However, the result of the authors still has some limitations, namely:

(L1) there are no numerical computation methods for estimating automatically the coefficients of each component of the model by fitting methods.

Besides, lack of mathematical foundation to explain the use of second order polynomials in the qTA model. It is not easy to solve (L1) above because a suitable trajectory must satisfy following two conditions, given a sample of fundamental frequency trajectory of the tone:

(C1) pitch target constraint (PTC), with big enough time *t*;

(C2) fitting constraint, for any time *t*.

In this paper, we propose new computational methods to determine the pitch markers of the original speech signal based on its cumulative signal and quantitative target approximation vectors namely qTA that generate the fundamental frequency trajectories of two-syllable tones. Our methods include three numerical solutions. For the first solution, we determine the pitch markers of the original speech signal in a time domain based on its cumulative signal. The second ones is proposed to calculate the qTA parameters by fitting a given F_0 trajectory of a speech syllable. This numerical solution is a tool for determining qTA parameters by fitting a given F_0 trajectory of a speech syllable and of a multi-syllable word. The third ones calculates qTA parameters by fitting a given F_0 trajectory of a multi-syllable word with the first step is the concatenating each F_0 trajectory of each speech syllable of the given speech two-syllable word to archieve a continuous F_0 trajectory and the second step is according to each syllable, calculating qTA parameters of the its part of the F_0 trajectory by applying the second solution.

By using polynominals for the approximation component of qTA model, qTA parameters obtained by the second solution already generates a F_0 trajectory with fitting a given complex shape F_0 trajectory of a mutli syllable word is better than the results published in [16; 20]. The target and plynominal's coefficients are namely qTA vector parameters or qTA representation. By the well-known Weierstrass approximation theorem, any given F_0 trajectory of word is fitting by synthesized trajectories based on qTA parameters. In addition, it should be emphasized that qTA's parameter calculation is completely automated.

The rest of the paper is organized as follows. Section 2 presents about RAPT framework, Fujíaki model and qTA model. Section 3 presents an algorithm to determine the pitch markers of the original voice signal in a time domain based on the cumulative signal. This section also presents two algorithms to solve (L1) at one and two-tone levels respectively. Experimental results are given in section 4. Conclusions and future research direction are in section 5.

Theoretical basis

RAPT framework and instantaneous pitch estimation. This issue requests to develop algorithms in determining parameters for representing fundamental frequency trajectories 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 PM and the parameters generating the fundamental frequency trajectory of the word.

For determining PMs and calculating the fundamental frequencies of speech samples, there are many algorithms in the literal, such as the results published in [4; 6; 22–25].

Most of these results follow an approach with three main steps: (i) divide a voiced segment into short segments (frames), (ii) estimate the fundamental frequency value at each frame and (iii) use dynamic programming algorithms to determine the PMs taken from the peak, or valley points of the speech signal and so on.

In details, RAPT estimates overall periodicity of the analysis frame using normalized cross-correlation function (NCCF). Let $s(m)$ be a speech signal, z – step size in samples and n – window size. The NCCF $\varphi(x, k)$ of K samples length at lag k and analysis frame x is defined as [4]:

$$\varphi(x,k)=\frac{\sum_{i=m}^{m+n-1}S(i)S(i+k)}{\sqrt{e_m e_{m+k}}},$$

$$k=0,K-1;\ m=xz;\ x=0,M-1,$$

where $e_i=\sum_{l=i}^{i+n-1}s_l^2$.

The Fujisaki model is a super positional model for representing F_0 trajectory of speech. According to the model, F_0 trajectory 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 trajectory is obtained by adding base frequency, phrase components and accent components.

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

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

$$G_p=\begin{Bmatrix}\alpha^2 t\exp(-\alpha)\text{ for }t\ge 0\\ 0\text{ for }t<0\end{Bmatrix},$$

$$G_a=\begin{Bmatrix}\min\left[1-(1+\beta t)\exp(-\beta t)\right]\text{ for }t\ge 0\\ 0\text{ for }t<0\end{Bmatrix},$$

where F_b – baseline value of fundamental frequency; I – number of phrase commands; J – number of accent commands; A_{pi} – magnitude of i phrase command; T_{0i} – timing of i phrase command; A_{aj} – amplitude of j accent command; T_{1j} – onset of j accent command; T_{2j} – offset of j 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.

The NCCF is the most computationally expensive operation in RAPT and so the algorithm performs the NCCF in a two pass process. A down-sampled version of the input signal issued to estimate the first set of candidate peaks, followed by a high resolution (full sample rate) NCCF around the candidates of interest.

The algorithm is summarized below:

• periodically compute the NCCF of the down sampled signal for all lags in the range of pitch. Location so flocal maxima in this 1st pass of the NCCF are recorded;

• compute the high resolution NCCF (signal at original sampling frequency) only around the peak locations recorded in previous step;

• search for local maxima in the high resolution NCCF to obtain improved peak locations and amplitude estimates;

• dynamic programming is used to select the set of NCCF peaks or unvoiced hypothesis across all frames.

Fujisaki model. In the Fujisaki model, as illustrated in the 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 trajectories. And the latter are the product the underlying commands generated by the articulatory mechanism implemented in the model. The surface F_0 trajectories 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 intonation trajectories of any language such as Russian, English, Vietnamese and so on. However, it is a complex model with a lot of parameters.

The qTA model is presented on the fig. 2, which will be detailed in the next section, simulates F_0 trajectories as syllable-synchronized laryngeal movements toward underlying pitch targets that are either static or dynamic.

Thus all these models assume that surface F_0 trajectories result from certain articulatory mechanisms rather than from direct acoustic manipulations.

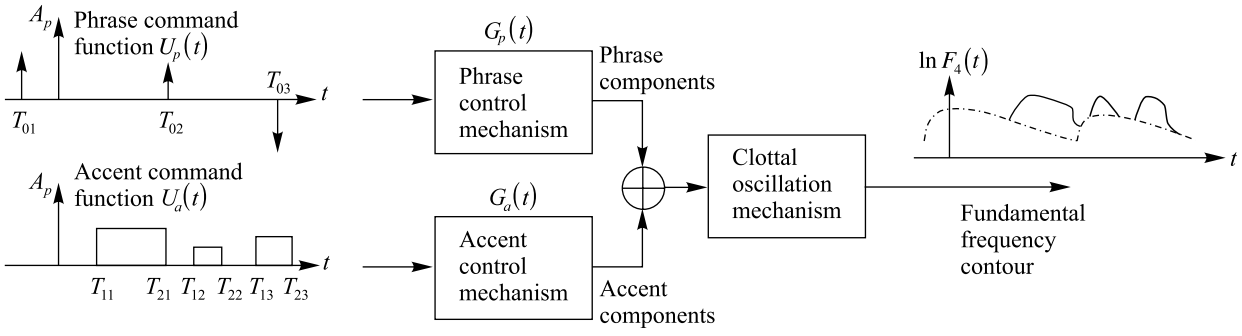

Fig. 1. Fujisaki model

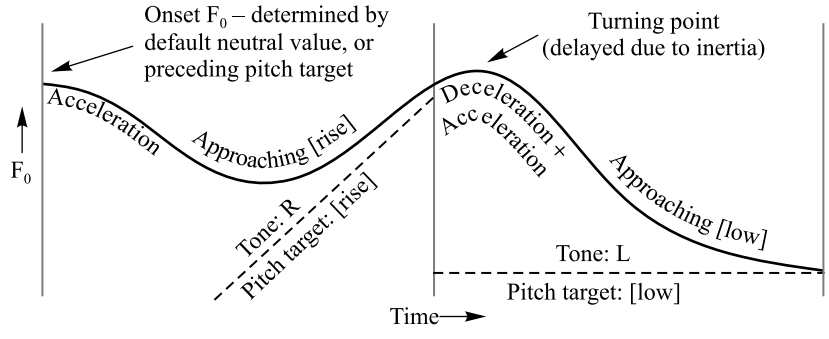

Fig. 2. The qTA model

qTA model (Xu model). In the phrase context, by the tone sandhi occuring, the number of trajectory shapes of syllables is increasing many times over the isolated syllables. Therefore, it is not easy to model these variations.

In the tonal languages, for parameterizing fundamental frequency trajectories of speech utterances, it is usual to use the Fujisaki or Xu models. For example, in [20] Hansjoerg Mixdorf and his colleagues already used the Fujisaki model to model Vietnamese fundamental frequency trajectories of syllables in the phrase context.

In the Fujisaki model, fundamental frequency trajectories are formed from the intonation trajectories and the stress trajectories. 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 trajectories. In addition, the Fujisaki model requires a lot of parameters to represent the fundamental frequency trajectories. Therefore, it is not easy to calculate Fujisaki model parameters by fitting the given fundamental frequency trajectory and until now there are no numerical computation methods to extrace the parameters by fitting methods.

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 the two.

The static and dynamic characteristics can be modeled using the «pitch target» concept of the Xu model [6]. This is a model that has been investigated and used by Xu and his colleagues to generate fundamental frequency trajectories for tonal languages such as Mandarin and Thai, for example Prom-on and Yi Xu [24; 26]. Advantages of the model are simple, less parameters and can be learned statistically to generate the appropriate fundamental frequency trajectories representation. About recent results using qTA representations of Xu model can be read in [21; 27; 28].

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 (1), where the first term $x(t)$ 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. In qTA, the initial F_0 dynamic state consists of initial F_0 level, $f_0(0)$, velocity $f_0'(0)$, and acceleration $f_0''(0)$. The dynamic state is transferred from one syllable to the next at the syllable boundary to ensure continuity of F_0. The three transient coefficients are computed with the formula presented on the fig. 2.

Proposed method for determining PMs and qTA representation

In this section, we propose a pitch mark detection algorithm for utterances and F_0 trajectories generation algorithms for tones in a tonal language.

PMs with cumulative signals. Let $x = \{x_j\}_{1 \le j \le N}$ be a voiced segment, without loss of generality, we assumed that the signal x is sampled from an interval $[-a, a]$ with some $a > 0$. The cumulative signal $s = \{s_j\}_{1 \le j \le N}$ of x can be defined by

$$s_1 = x_1, \ \forall j = \overline{2, N}, \ s_j \overset{\text{def}}{=} s_{j-1} + x_j = \int_1^j x_t dt.$$

Example 1. Consider the following utterance extracted in the Vietnamese book «Adventures of a Cricket», where PMs (marked by small circles) of the original speech are located at the signal points changing from positive to negative, as the peaks of the cumulative signal respectively, this case is described by the fig. 3.

As we can see, there is a relationship between signal points changing from positive to negative and the peaks of the correspondent cumulative signal by the following fig. 4.

For the following utterance, it is divided automatically into 7 voiced segments by using a method for locating the silience/voiced/unvoiced part (see [21]), each segment is shown in a pair of red-blue dashed lines, this case described by the fig. 5.

The peaks of the cumulative signal are more visible than the peaks of the original voice signal as illustrated in fig. 6, *b*.

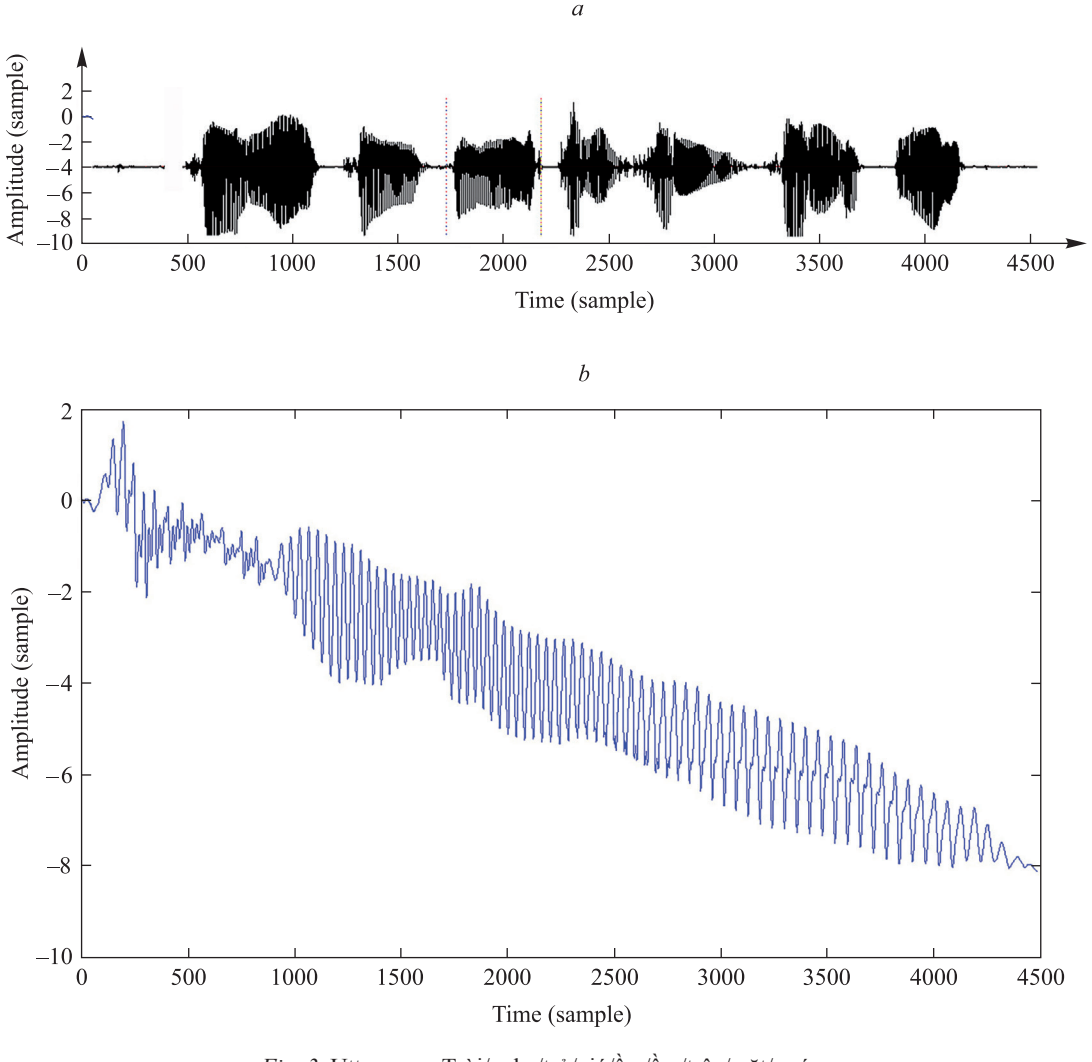

Fig. 3. Utterance «Trời/nghe/trở/gió/ầm/ầm/trên/mặt/nước»
(IPA transcription: «tɕəː˨˩ ŋɛ˧˧ tɕəː˧ zɔ˧˩ ə̰m˨ ə̰m˨ tɕen˧˧ mə̰ʔt˨ niək˧˩»;
translation: «God make the rumbling wind on the water») (*a*)
and the cumulative signal of the voiced part /gió/(/zɔ˧˩/, /wind/) (*b*)

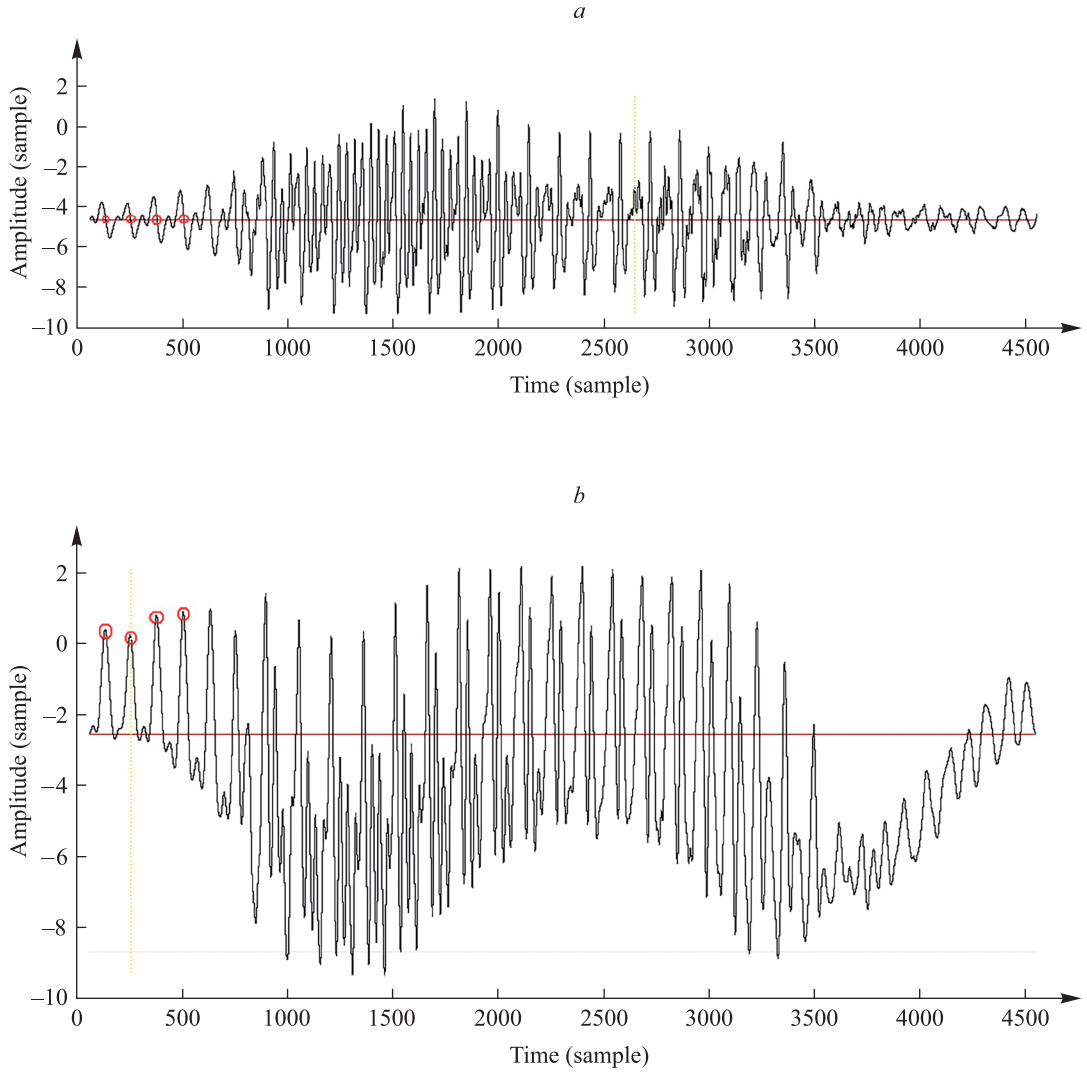

Fig. 4. PMs are located at the points
in which the voice signal changes from positive to negative (*a*)
and corresponding peaks of the cumulative signal (*b*)

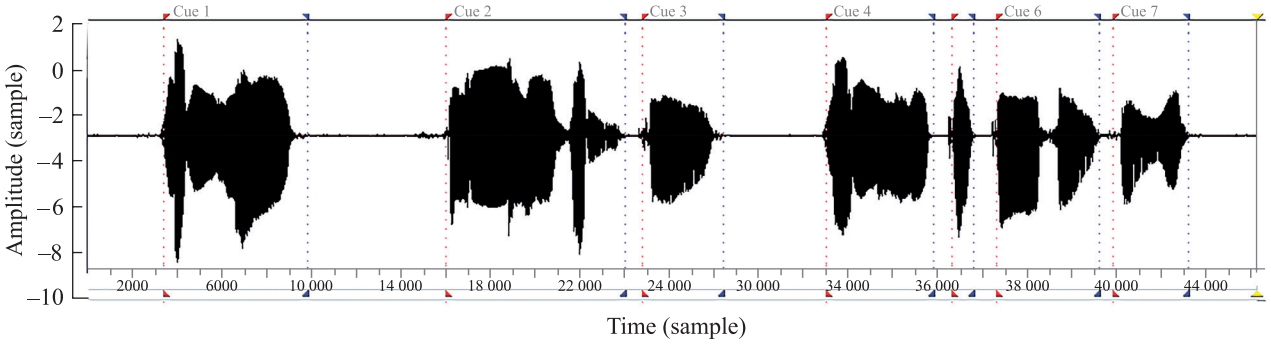

Fig. 5. «Đừng lo, xem mây vận trời đêm nay có cơ đổi gió»
(IPA transcription: «ɗi̤ŋ˨ lɔ˧˧ sɛm˧˧ məj˧˧ və̰ʔn˨ tɕə̰ːj˨ ɗem˧˧ naj˧˧ kɔ˧ kəː˧˧ ɗo̰j˨ zɔ˧»;
translation: «Do not worry, looking at the clouds, the wind may change direction tonight»)

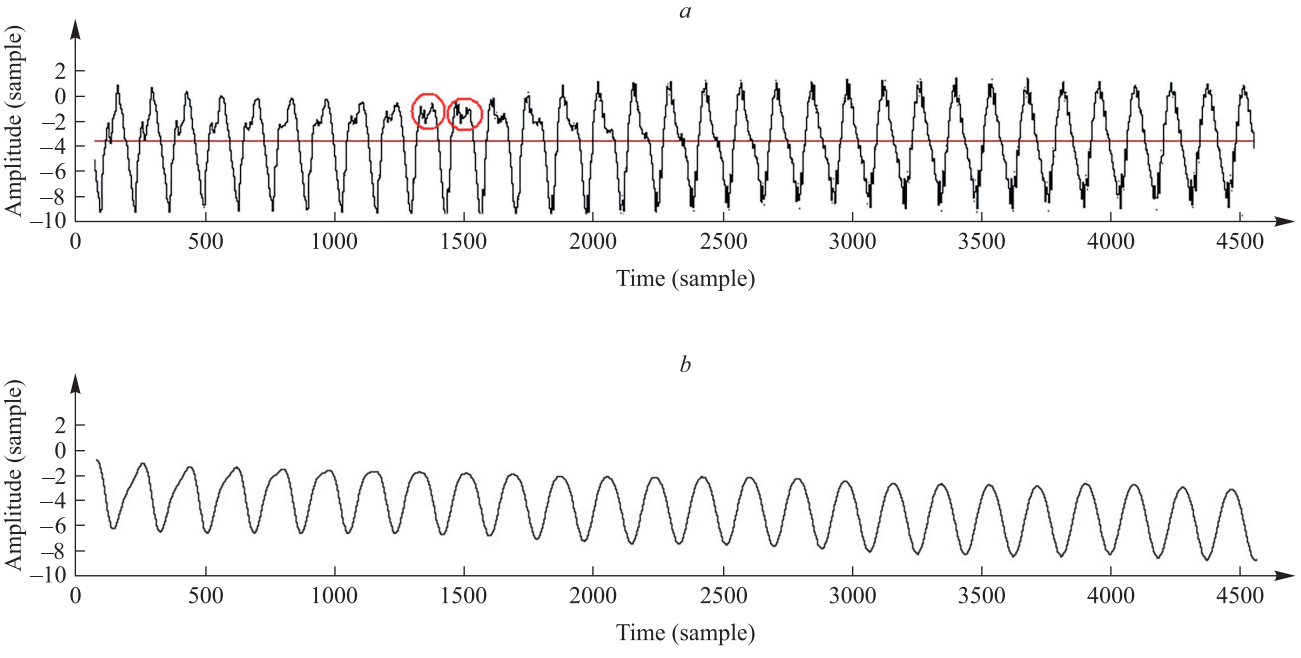

Fig. 6. A partial signal of the second segment (*a*) and corresponding cumulative signal (*b*)

The PMs located at peaks of the original voice signal are indistinguishable in amplitude from points surrounding them. In the details, the peaks having higher amplitude than the surrounding ones, usually also are PMs of the cumulative signal.

The peaks of the cumulative signal are related to the time points at which the original signal changes from positive to negative. This is the principle that if the PMs of the cumulative signal are well positioned, we will successfully locate PMs from the peak or valley points of the original voiced signal.

First of all, we will give some definitions and prove some simple properties derived from them.

Definition 1. (The sets of time points at which the original signal changes from positive to negative and vice versa.)

For $x = \{x_j\}_{1 \le j \le N}$, we let z_x^+ and z_x^- denote two sets of time points as the following:

$$z_x^+ \stackrel{\text{def}}{=} \{i | x_i > 0 \wedge x_{i+1} < 0\},\ z_x^- \stackrel{\text{def}}{=} \{i | x_i < 0 \wedge x_{i+1} > 0\}.$$

In addition, we also denote

$$x^+ \stackrel{\text{def}}{=} \{i | x_i > 0\}$$

and

$$x^- \stackrel{\text{def}}{=} \{i | x_i < 0\}$$

and $\text{peak}(\mathbf{x})$ denote the peak set of $\mathbf{x}$ (to get $\text{peak}(\mathbf{x})$, see [28]).

Proposition 1. *(i)* $\text{peak}(s) \subset z_x^+$ *and* $\text{peak}(-s) \subset z_x^-$, *where s is the cumulative signal of x.*

(ii) If $\overline{i, j} \subset x^+$ *then* $\{s_i, s_{i+1}, \ldots, s_j\}$ *is a monotonic increasing sequence, and if* $\overline{i, j} \subset x^-$ *then* $\{s_i, s_{i+1}, \ldots, s_j\}$ *is a monotonic decreasing sequence.*

Proof.

(i) $\forall i \in \text{peak}(s) \Rightarrow s_i > s_{i-1} \wedge s_i > s_{i+1} \Rightarrow (x_i = s_i - s_{i-1} > 0) \wedge (x_{i+1} = s_{i+1} - s_i < 0) \Rightarrow x_i > 0 \wedge x_{i+1} < 0 \Rightarrow$

$\Rightarrow i \in z_x^+$. So $\text{peak}(s) \subset z_x^+$. Similarly, we have $\text{peak}(-s) \subset z_x^-$.

(ii) $\overline{i, j} \subset x^+ \Rightarrow \forall k = \overline{i, j-1},\ x_k > 0 \Rightarrow s_{k+1} - s_k = x_{k+1} > 0 \Rightarrow s_{k+1} > s_k$.

Moreover, $\overline{i, j} \subset x^- \Rightarrow \forall k = \overline{i, j-1},\ x_k < 0 \Rightarrow s_{k+1} - s_k = x_{k+1} < 0 \Rightarrow s_{k+1} < s_k$.

From here, we propose a new approach, instead of locating PMs based on the original speech wave, we determine PMs in the timing of peaks of the cumulative signal of the speech. From the PMs of the cumulative signal we will locate the other PMs, such as the PMs located from the peaks or valleys of the speech signal.

Definition 2. Let $x = \left\{x_j\right\}_{1 \le j \le N}$ be a voiced segment and $s = \left\{s_j\right\}_{1 \le j \le N}$ the cumulative signal of **x**. Let denote pitch marker zeros (PMZ), $\mathrm{PMZ}_x^+ = \left\{pmz_j^+\right\}$ as the given PMs which located from the peaks of **s**. We let PMZ_x^-, PM_x^+ and PM_x^- denote three PM sets derived from PMZ_x^+ (find in each range of two consecutive PMs of PMZ_x^+) as the following:

$$\mathrm{PMZ}_x^- \stackrel{\text{def}}{=} \left\{k/\exists j : k = \min\left\{l/l \in \mathrm{peak}(-s),\ pmz_{j-1}^+ \le l \le pmz_j^+\right\}\right\},$$

$$\mathrm{PM}_x^+ \stackrel{\text{def}}{=} \left\{k/\exists j : k = \min\left\{l/l \in \mathrm{peak}(x),\ pmz_{j-1}^+ \le l \le pmz_j^+\right\}\right\},$$

$$\mathrm{PM}_x^- \stackrel{\text{def}}{=} \left\{k/\exists j : k = \min\left\{l/l \in \mathrm{peak}(-x),\ pmz_{j-1}^+ \le l \le pmz_j^+\right\}\right\}.$$

Let s_k be the cumulative signal of x_k, where x_k is the k voiced segment of the utterance. To determine PMZ_k^+ of s_k, we can see that the PMs are chosen based on the following two criteria:

(i) the dependencies of the distances between consecutive PMs;

(ii) with two adjacent peaks of s_k, the peak with a greater amplitude is preferred over the other.

The process of selecting the appropriate peaks of s_k is a looping, multi-step process, consisting of appends, deletions, insertions and modifications to ensure that the criteria (i) and (ii) described above do not create redundancy and lost of PMs. With that said, we propose a simple and intuitive R1–R6 rules, to determine $\mathrm{PMZ}_{k,x}^+ = \left\{p_{k,n} \middle| p_{k,n} \in \mathrm{peak}(s_k)\right\}$ for s_k.

R1. (Appending the first PM.)

$\mathrm{PMZ}_{k,x}^+ = \left\{p_{k,1}\right\}$, where $\mathrm{mean}_k \stackrel{\text{def}}{=} \dfrac{\sum\limits_{n \in \mathrm{peak}\{s_k\}} \left|s_{k,n}\right|}{\#\mathrm{peak}\left\{s_k\right\}}$, $p_{k,1} = \underset{n \in \mathrm{peak}\{s_k\}}{\arg\min}\left\{\left|s_{k,n}\right| \ge \mathrm{mean}_k\right\}$ (the first PM $p_{k,1}$ of s_k is the first n peak whose amplitude $s_{k,n}$ is over the threshold mean_k).

R2. (Appending the next temporary PM.)

If there exist some $m \in \mathrm{peak}(s_k)$ and $m - p_{k,j} \in \left[f_s/f_{0,\max},\ f_s/f_{0,\min}\right]$, where $p_{k,j} = \max\left\{\mathrm{PMZ}_{k,x}^+\right\}$ then $\mathrm{PMZ}_{k,x}^+ = \mathrm{PMZ}_{k,x}^+ \cup \left\{m\right\}$.

R3. (Delete a temporary PM.)

If there exist some two consecutive temporary PMs, $p_{k,j-1}, p_{k,j} \in \mathrm{PMZ}_{k,x}^+$ such that $p_{k,j} - p_{k,j-1} \notin \left[f_s/f_{0,\max},\ f_s/f_{0,\min}\right]$, then $\mathrm{PMZ}_{k,x}^+ = \mathrm{PMZ}_{k,x}^+ \setminus \left\{p_{k,j}\right\}$.

R4. (Delete a temporary PM.)

If there exist some three consecutive temporary PMs $p_{k,j-1}, p_{k,j}, p_{k,j+1} \in \mathrm{PMZ}_{k,x}^+$ such that $s_{k,p_{k,j}} < \min\left\{s_{k,p_{k,j-1}}, s_{k,p_{k,j+1}}\right\} \wedge \min\left\{p_{k,j} - p_{k,j-1},\ p_{k,j+1} - p_{k,j}\right\} < 0.5^* \max\left\{p_{k,j} - p_{k,j-1},\ p_{k,j+1} - p_{k,j}\right\}$, then $\mathrm{PMZ}_{k,x}^+ = \mathrm{PMZ}_{k,x}^+ \setminus \left\{p_{k,j}\right\}$.

R5. (Insert a peak into the temporary PM set.)

If there exist some three consecutive temporary PMs $p_{k,j-1}, p_{k,j}, p_{k,j+1} \in \mathrm{PMZ}_{k,x}^+$ such that $p_{k,j+1} - p_{k,j} > \alpha^*\left(p_{k,j} - p_{k,j-1}\right)$ then $\mathrm{PMZ}_{k,x}^+ = \mathrm{PMZ}_{k,x}^+ \cup \left\{m\right\}$, where m is $m \in \mathrm{peak}(S_k)$: $p_{k,j} < m < p_{k,j+1}$, $\left|m - \left(p_{k,j} + p_{k,j+1}\right)/2\right| \to \min$ and α is an experimental parameter, $\alpha > 1$.

R6. (Replace value of a temporary PM.)

If there exist some three consecutive temporary PMs $p_{k,j-1}, p_{k,j}, p_{k,j+1} \in \mathrm{PMZ}_{k,x}^+$ such that $\exists m \in \mathrm{peak}\left(S_{\left[p_{k,j} - T/2,\ p_{k,j} + T/2\right]}\right) \wedge \left(\forall t \in \left[p_{k,j} - T/2,\ p_{k,j} + T/2\right] \Rightarrow s_{k,t} \le s_{k,m}\right)$ then reassign $p_{k,j} = m$, where $T \stackrel{\text{def}}{=} \stackrel{\text{def}}{=} \min\left\{p_{k,j} - p_{k,j-1},\ p_{k,j+1} - p_{k,j}\right\}$.

Using R1–R6 rules, the proposed algorithm determining the PMs based on the cumulative signals includes some simple main steps, it is described by the fig. 7.

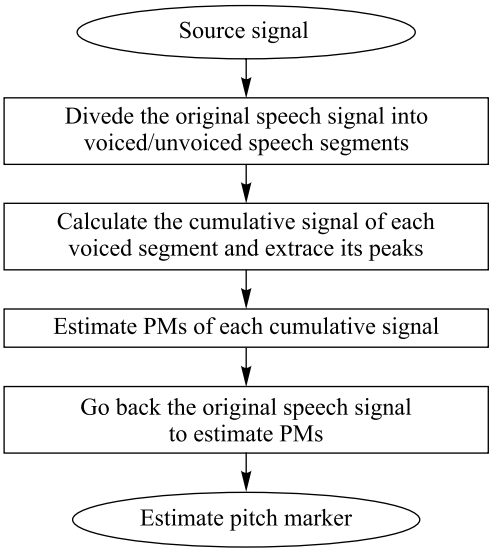

Fig. 7. Scheme of estimate PMs of a speech utterance

The algorithm of EPM is given as follows.

Algorithm 1. EPM (Estimating the PMs of speech waves.)
Input: speech signal $\{x_m\}_{1 \le m \le N}$ in time domain.

Sampling frequency value: f_s, $\left[f_{0,\min}, f_{0,\max}\right]$ is the range of F_0 values.

Output: number of voiced segments K, PMs according to four types

$$\left\{pm^+_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^+_k},\ \left\{pm^-_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^-_k},\ \left\{p^-_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^-_{k,2}},\ \left\{p^+_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^+_{k,2}},$$

where $\left\{pm^+_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^+_k}$, $\left\{pm^-_{k,j}\right\}_{1 \le k \le K,\, 1 \le j \le n^-_k}$ are the two traditional PMs.

Step 1: segment the signal $\{x_m\}_{1 \le m \le N}$ into K voiced segments, $\{x_m\}_{N_{k,1} \le m \le N_{k,2}}$ and other ones (a simple method, see [14]).

Step 2: $T_{\min} = f_s / f_{0,\max}$, $T_{\max} = f_s / f_{0,\min}$.

Step 3: repeat, on each voiced segment $\{x_m\}_{N_{k,1} \le m \le N_{k,2}}$, $k = \overline{1, K}$ to determine $\mathrm{PMZ}^+_{x,k}$:

3.1: calculates the cumulative signal $s_k = \{s_m\}_{N_{k,1} \le m \le N_{k,2}}$, $k = \overline{1, K}$ following the formula (1).

3.2: determine the peak of s_k, compute the average amplitude at the peak of s_k:

$$\mathrm{mean}_k = \sum_{n \in \mathrm{peak}\{S_k\}} \left|s_{k,n}\right| / \#\mathrm{peak}\left\{s_{k,n}\right\}.$$

3.3: determine the first PM of $\mathrm{PMZ}^+_{x,k}$ by using rule R1.
3.4: repeat the substeps 3.5–3.8 when at least one of the conditions of the rules R2 – R6 is true.
3.5: using the rule R2 to extend $\mathrm{PMZ}^+_{x,k}$.
3.6: using the rules R3 and R4 to reduct $\mathrm{PMZ}^+_{x,k}$.
3.7: using the rule R5 to extend $\mathrm{PMZ}^+_{x,k}$.
3.8: using the rule R6 to change the element values of $\mathrm{PMZ}^+_{x,k}$.
3.9: stop and obtain $\mathrm{PMZ}^+_{x,k}$.
3.10: determine the PMs according to the local maximum point criterion of k segment $\{x_m\}_{N_{k,1} \le m \le N_{k,2}}$.

For each range of two consecutive PMs of PMZ^+_x, find $pm^+_{k,j} = \max\left\{\mathrm{peak}\left\{x_n\right\}_{p_{k,j} \le n \le p_{k,j+1}}\right\}$ and obtain $\mathrm{PM}^+_{x,k} = \left\{pm^+_{k,j}\right\}$.

3.11: determine PMs according to the local minimum point criterion of k segment $\{x_m\}_{N_{k,1} \le m \le N_{k,2}}$.

For each range of two consecutive PMs of PMZ_x^+, find $pm_{k,j}^- = \min\left\{\mathrm{peak}\left\{-x_n\right\}_{p_{k,j} \le n \le p_{k,j+1}}\right\}$ and obtain $\mathrm{PM}_{x,k}^- = \left\{pm_{k,j}^-\right\}$.

Step 4: determine PMs like pulse points for Praat type [24], the same as step 3 above, but taking the local minimum point of the cumulative signal, obtain $\mathrm{PMZ}_{x,k}^- = \left\{p_{k,j}^-\right\}$ on k segment $\left\{x_m\right\}_{N_{k,1} \le m \le N_{k,2}}$, $k = \overline{1, K}$.

Step 5: put $\mathrm{PMZ}_x^+ = \bigcup\limits_{1 \le k \le K} \mathrm{PMZ}_{x,k}^+$, $\mathrm{PMZ}_x^- = \bigcup\limits_{1 \le k \le K} \mathrm{PMZ}_{x,k}^-$, $\mathrm{PM}_x^+ = \bigcup\limits_{1 \le k \le K} \mathrm{PM}_{x,k}^+$ and $\mathrm{PM}_x^- = \bigcup\limits_{1 \le k \le K} \mathrm{PM}_{x,k}^-$.

Return: number of voiced segments K, PMZ_x^+, PMZ_x^-, PM_x^+ and PM_x^-.

After obtaining PMs of tonal word speech signals, the next step is to stylize F_0 trajectory of the tones and finally use an algorithm such as the PSOLA [29] to create the desired speech word from multiple input syllables.

The following proposed algorithms will focus on generating F_0 trajectories of tones by using the pitch target model.

Generating F_0 trajectories of Vietnamese isolated syllables. We will apply the method to identify PMs to synthesize tones of Vietnamese isolated syllables. To stylized tones, we use Xu model, which has been widely used for Mandarin [30] to model F_0 contours of the tones (for tonal languages). $F(t) \approx \alpha^* e^{-\lambda t} + a^* t + b$ such that a F_0 contour is created from the combination of the two components: the linear approximation $\alpha^* t + b$ and the non-linear approximation $\alpha^* e^{-\lambda t}$.

The computing of the coefficients of the model, given trend-line F_0 value also uses the least squares method, instead of finding the coefficients a, b, α, λ we determine a, b, k ($k = e^{-\lambda}$) by minimize the objective function:

$$\sum_{i=1}^{n-1}\left(F_{0,i+1} - a^*(i+1) - b - k^*\left(F_{0,i} - a^* i - b\right)\right)^2 \to \min, \tag{2}$$

where n is the number of speech frames, $\left\{F_{0,i}\right\}_{i=1}^{n}$ is a F_0 sequence of each frame corresponding. The stylized method using Xu model is built as follows.

Step 1: select syllables with level tone, drop tone with syllables ending *p-t-c*/*ch*, determine F_0 trajectory of them.

Step 2: determine the PMs of this wave of tone by algorithm 1.

Step 3: using least squares method to fit Xu model's parameters as a, b, k. Generate target F_0 trajectory by the Xu model.

Step 4: using PSOLA algorithm to synthesize a syllable with the target tone.

The algorithm of synthesis of tones is given as follows.

Algorithm 2. (Synthesis of tone for a Vietnamese syllable signal.)

Input: voice signal x_{in} in time domain of a Vietnamese syllable with any given tone {level, falling, raising, drop, curve, broken}. Sampling frequency value f_s.

Parameters $[a_{tn}, b_{tn}, c_{tn}, d_{tn}, g_{tn}, k_{tn}]$ represent the target tone *tn* belong to {level, falling, raising, drop, curve, broken}. Need to synthesize in the form of qTA in formula (2), $0 < k_{tn} < 1$.

$\Delta > 0$ is the width parameter of the frame with measure units of milliseconds, N, M are the length of input syllable length and synthesis syllable, the calculating unit is milliseconds.

Output: x_{out}, the sound wave has the tone *tn*.

Step 1: use the value of f_s, convert N, M, Δ to the number units of sample.

Step 2: determine the set $\mathrm{PM}_{\mathrm{in}}^+$ (starting assign $\mathrm{PM}_{\mathrm{in}}^+(0) = 0$) of input sound waves using the proposed algorithm 1. Notice that on the unsound we assign:

$$\mathrm{PM}_{\mathrm{in}}^+(k) = \mathrm{PM}_{\mathrm{in}}^+(k-1) + \Delta,\ k \in (1, N_{\mathrm{PM}}).$$

Step 3: generating F_0 trajectory of target syllables using formula (2), concretely calculated as follows:

$$f_{0,\mathrm{out}}(t) = a_{tn}^* t + b_{tn} + (k_{tn})^{t^*}\left(c_{tn} t^2 + d_{tn} t + g_{tn}\right),\ t = \overline{1, T_{\mathrm{out}}},$$

where $T_{\mathrm{out}} = [M/\Delta]$.

Step 4: determine the set $\mathrm{PM}_{\mathrm{out}}$ as the following formula:

$$PM_{\mathrm{out}}(0) = 0,\ PM_{\mathrm{out}}(k) = PM_{\mathrm{out}}(k-1) + f_s / f_{0,\mathrm{out}}(k/\Delta),\ k = \overline{1, N_{\mathrm{out}}},$$

where $N_{\mathrm{out}} = \max\left\{k : PM_{\mathrm{out}}(k) \le M\right\}$.

Step 5: use the algorithm PSOLA [29] getting:

$$x_{\text{out}} = \text{PSOLA}\left(x_{\text{in}}, PM^{+}_{\text{in}}, PM_{\text{out}}\right).$$

Output: wave signal x_{out} syllable has new tone is *tn*.

Experiment

In order to experiment with proposed algorithms, we use Vietnamese voice data to illustrate. The Vietnamese language is a monosyllabic and tonal language with six tones (see table) and is the most complex lexical tone in tonal languages.

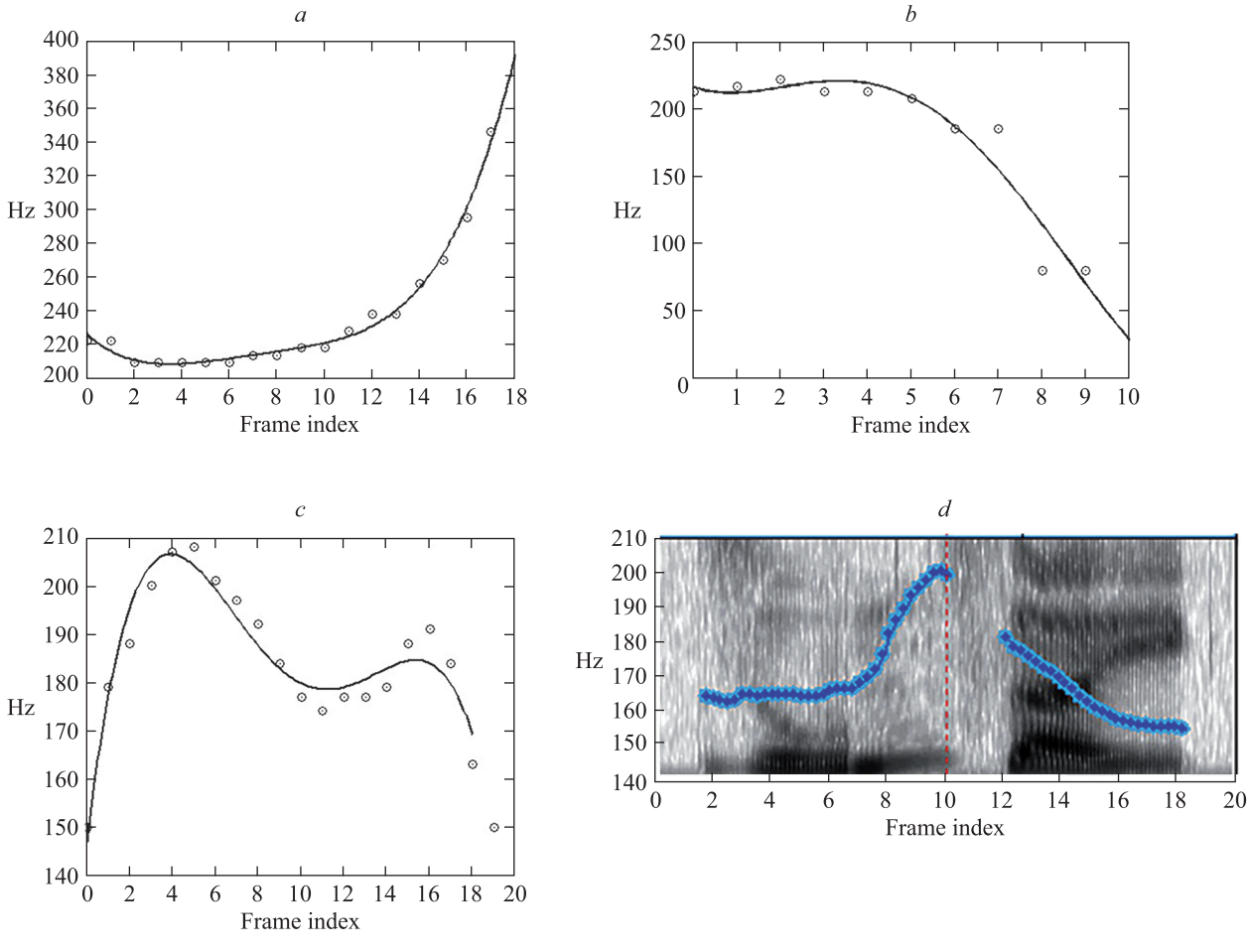

Fig. 8. The typical F_0 trajectories shape of some tones of Vietnamese isolated syllables rising tone (*a*), broken tone (*b*), drop tone (*c*) and A F_0 trajectory (*d*) of the word /*dún*/*dây*/ (z unˀ z ajˋ) with tone sandhi

Experimental data. In order to experiment the algorithms, a single speaker story reading corpus was created, uttered by a female speaker of standard Vietnamese voice. Sentences are extracted in the Vietnamese book «Adventures of a Cricket».

Experiment to extract the PM points. The formulas show that algorithm 1 has a smaller computational complexity than dynamic programming-type algorithms [4] because it does not require the steps to segment the whole speech utterance into short time frames and choose a suitable time point of each short time frame that gives high autocorrelation value. For the reliability of algorithm 1, we will compare algorithm 1 with the Talkin-type algorithm implemented in software Praat [23]. The parameter $f_{0,\,\min} = 50$ Hz, $f_{0,\,\max} = 550$ Hz and $a = 1.6$ for the R5 rule.

To compare the similarity between the two PM sequences of the same voiced segment, we give the following objective indexes that is based on the edit-distance (about a related work, see the algorithm for alignment of the reference epochs (EGG epochs) to the test epochs [18]).

Firstly, let $\text{PM}_I = \left\{pm_i\right\}_{i=1}^{m}$ and $\text{PM}_J = \left\{pm'_j\right\}_{j=1}^{n}$, then we define the measured value $D_{\text{PM}}\left(\text{PM}_I, \text{PM}_J\right)$ by:

$$D_{\mathrm{PM}}\left(\mathrm{PM}_I, \mathrm{PM}_J\right) \stackrel{\text{def}}{=} D_{m,n}\left(\left\{pm_i\right\}_{i=1}^{m}, \left\{pm'_j\right\}_{j=1}^{n}\right)/\min\{m, n\},$$

where $D_{1,1} = \left|pm_1 - pm'_1\right|$ and $D_{i,j} = \left|pm_i - pm'_j\right| + \min\left\{D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}\right\} \forall i,\ j \geq 2.$

Secondly, over the whole utterance, we get the average of the D_{PM} values calculated from the same voiced segment sequence of the utterance. The average $\overline{ED}$ is defined as follows:

$$\overline{ED} = \sum_{k=1}^{K} D_{\mathrm{PM}}\left(\mathrm{PM}_{I,k}, \mathrm{PM}_{J,k}\right)/K,$$

where $\left\{\mathrm{PM}_{I,k}\right\}_{k=1}^{K}$ and $\left\{\mathrm{PM}_{J,k}\right\}_{k=1}^{K}$ are PM sequences of k voiced segment of the utterance that have K voice segments total.

To compare with another PM estimation method such as Praat [23], we use the algorithm 1 to obtain the PMs PM_x^- with valley type for each voiced segment received by Praat, then we calculate $\overline{ED}$ values. Table below shows the similarity between the estimated PM_x^- and PMs (called pulse points, PPs) of the Praat type.

Measuring the similarity between PM_x^- and PPs of Praat

Utterance	Content	K	$\overline{ED}$, ms
#1	*«Đừng lo xem mây vận trời đêm nay có cơ đổi gió»* «ɗiŋ˧˩ lɔ˧˥ sɛm˧˥ məj˧˥ vəʔn˧˩ tɕəːj˧˩ ɗɛm˧˥ naj˧˥ kɔ˧˩ kəː˧˥ ɗoj˧˩˧ zɔ˧˩» «Do not worry, looking at the clouds, the wind may change direction tonight»	7	0.502 2
#2	*«Từ chỗ này muốn qua chỗ khác chúng tôi chỉ lách nhích từng tẹo»* «tḭ˧˩ tɕoʔo˧˩˩ʔnaj˧˩˧amuən˧˩˧ləkwaː˧˥ tɕoʔo˧˩˩ʔxaːk˧˩˧atɕuŋ˧˩˧ŋtoj˧˥ tɕḭ˧˩˧ lajk˧˩ ɲïk˧˩˧ktiŋ˧˩˧ʈɛʔw˧˩˧ʔ» «To move from one place to another, we have to move little by little»	13	0.323 4
#3	*«Chũi bảo chũi không nhìn thấy»* «tɕuʔuj˧˩˩ʔɓaːʔujthtɕuʔuj˧˩˩ʔxəwŋ˧˥wɲiɲ˧˩˧ʈʰəj˧˩˧ləj» «Chui claimed he could not see anything»	5	0.367 1
#4	*«Trời nghe trở gió ầm ầm trên mặt nước»* «tɕəːj˧˩ ŋɛ˧˥ tɕəː˧˩˧ zɔ˧˩ əm˧˩ əm˧˩ tɕen˧˥ maʔt˧˩ niək˧˩» «God makes the rumbling wind on the water»	4	0.275 1
#5	*«Thì ra bè chúng tôi từ lúc nào đã trôi vào gần một bờ cỏ»* «tʰḭ˧˩ zaː˧˥ ɓɛ˧˩ tɕuŋ˧˩ toj˧˥ tḭ˧˩ luk˧˩ naː˧˩ ɗaʔa˧˩ tɕoj˧˥ vaː˧˩ ɣən˧˩ moʔt˧˩ ɓəː˧˩ kɔ˧˩˧» «*Turns out our boat has drifted toward the grasslands*»	13	0.229 2
#6	*«Ấy vậy mà lúc đó chén ngon đáo để»* «əj˧˩ vəʔj˧˩ maː˧˩ luk˧˩ ɗɔ˧˩ tɕɛn˧˩ ŋɔn˧˥ ɗaːw˧˩ ɗḛ˧˩˧» «*The food was surprisingly yummy to me though*»	7	0.221 7

As we can see, the PMs determined by the algorithm 1 are more noticeable than the result of Praat when directly observing by eyes the speech signals as illustrated in fig. 9, *a*, and fig. 9, *b*, below.

However, Praat can ignore some pulse points, this case is described by the fig. 10, *a*, and fig. 10, *b* (whereas algorithm 1 does not).

Conclusion

In this paper, we propose two algorithms to determine the pitch markers of the original voice signal based on the cumulative signal and generate F_0 trajectories of tones.

The first algorithm is effective, with no need to divide a voiced segment into short segments (frames) as other methods, yet still achieving high accuracy. With the Vietnamese speech data of the lexical tones and phonetics tested (the full coverage of the Vietnamese phonetics was included), the results of calculating the pitch markers according to the new approach proved to be correct. The second algorithm used for generating F_0 trajectories of tones with qTA parameters of Xu model.

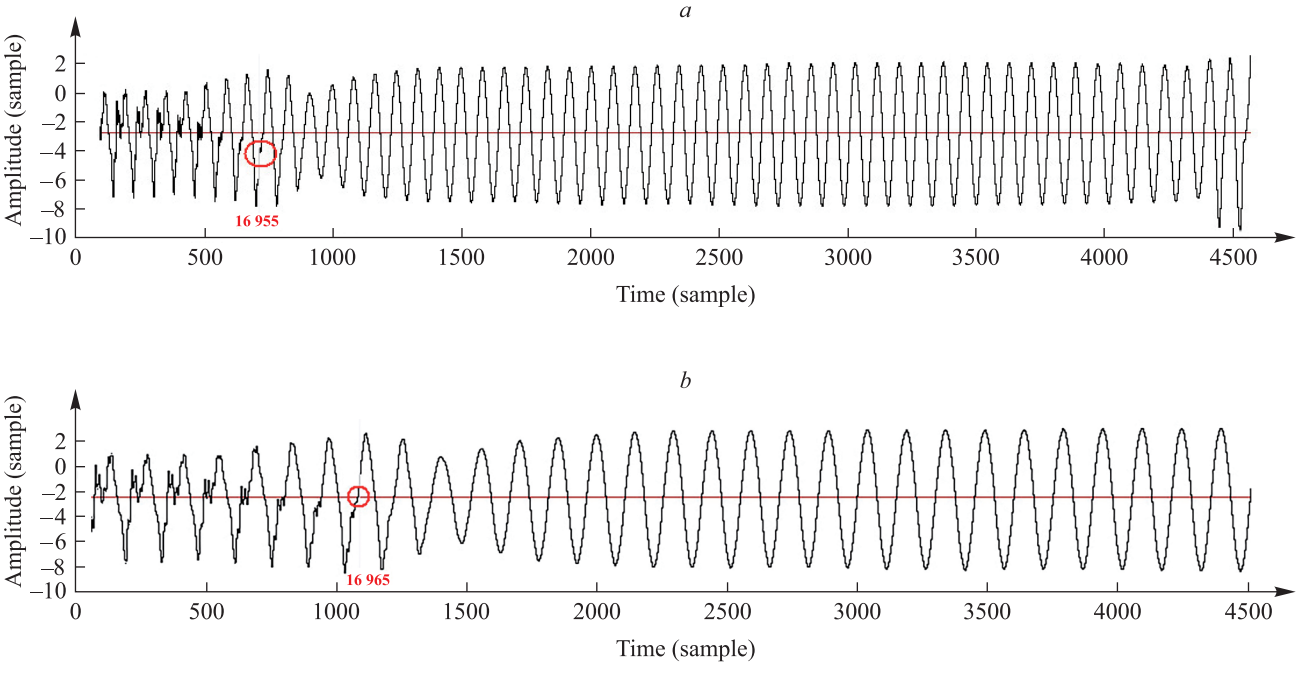

Fig. 9. One PM is determined by Praat (*a*); one PM is determined by the algorithm 1 (*b*).
Source: [8]

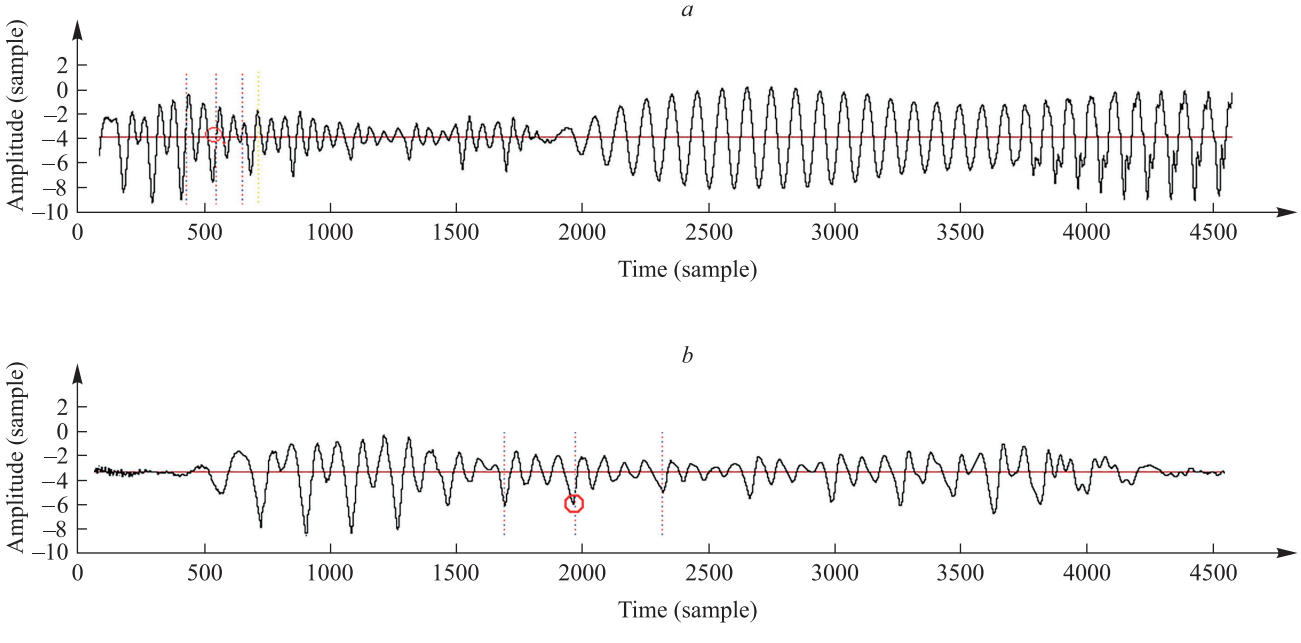

Fig. 10. With the utterance #2 (see table)
in the second voice segment, mising a PP around 355th sample (*a*),
and in the five voice segment, mising a PP around 497th sample (*b*)

Yi Xu has focused on how lexical tones of Mandarin were produced and perceived in continuous speech and has proposed the qTA model which considers the segmental phonemes, tones, and pitch accents as abstract units called pitch targets. In Mandarin, pitch targets are separated into static targets-[high] and targets-[low], and dynamic ones-[rise] and ones-[fall], which are associated with the four lexical tones respectively. This model 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 difficult to implement on training speaker dependent prosodic styles. 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 related to pause, 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) as the same) is the most characteristic.

References

1. Kovacs MD, Cho MY, Burchett PF, Trambert M. Benefits of integrated RIS/PACS/Reporting due to automatic population of templated reports. *Current Problems in Diagnostic Radiology.* 2019;48(1):37–39. DOI: 10.1067/j.cpradiol.2017.12.002.

2. Plonkowski M, Urbanovich P. The use of pitch in large-vocabulary continuous speech recognition system. *Przeglad Elektrotechniczny.* 2016;92(8):78–81.

3. Wang D, Hansen JHL. F_0 estimation for noisy speech by exploring temporal harmonic structures in local time frequency spectrum segment. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2016 March 20–25; Shanghai, China.* [S. l.]: IEEE; 2016. p. 6510–6514. DOI: 10.1109/ICASSP.2016.7472931.

4. Talkin D. A Robust Algorithm for Pitch Tracking (RAPT). In: Kleijn WB, Paliwal KK, editors. *Speech Coding & Synthesis.* [S. l.]: Elsevier Science B. V.; 1995. p. 495–518.

5. Xu Yi, Prom-on S. Articulatory-functional modeling of speech prosody: a review. In: Kobayashi T, Hirose K, Nakamura S. *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH-2010); 2010 September 26–30; Makuhari, Chiba, Japan.* [S. l.]: International Speech Communication Association; 2010. p. 46–49.

6. Kounoudes A, Naylor PA, Brookes M. The DYPSA algorithm for estimation of glottal closure instants in voiced speech. In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (CASSP'02); 2002 May 13–17; Orlando, FL, USA.* [S. l.]: IEEE; 2002. p. I349–I352. DOI: 10.1109/ICASSP.2002.5743726.

7. Smits R, Yegnanarayana B. Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing.* 1995; 3(5):325–333. DOI: 10.1109/89.466662.

8. Prom-on S, Liu F, Xu Y. Functional modeling of tone, focus and sentence type in mandarin Chinese. *Proceedings of the 17th International Congress of Phonetic Sciences; 2011 August 17–21; Hong Kong, China.* Hong Kong: City University of Hong Kong; 2011. p. 1638–1641.

9. Bailly G, Holm B. SFC: a trainable prosodic model. *Speech Communication.* 2005;46(3–4):348–364.

10. Fujisaki H. dynamic characteristics of voice fundamental frequency in speech and singing. In: MacNeilage PF, editor. *The Production of Speech.* New York: Springer; 1983. p. 39–55. DOI: 10.1007/978-1-4613-8202-7_3.

11. Kochanski G, Shih C. Prosody modeling with soft templates. *Speech Communication.* 2003;39(3–4):311–352. DOI: 10.1016/S0167-6393(02)00047-X.

12. Fujisaki H, Hirose K. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan.* 1984;5(4):233–242.

13. Xu Y, Wang QE. Pitch targets and their realization: evidence from Mandarin. *Speech Communication.* 2001;33(4):319–337. DOI: 10.1016/S0167-6393(00)00063-7.

14. Thai TY, Hung NV, Tuyet DV, Huy NHo, Ablameyko S. An effective algorithm for determining pitch markers of Vietnamese speech sentences. In: Huang T, Lv J, Sun C, Tuzikov A, editors. *Advances in Neural Networks – ISNN'2018. Proceedings of the 15th International Symposium on Neural Networks, ISNN'2018; 2018 June 25–28; Minsk, Belarus.* Cham: Springer; 2018. p. 628–636. (Lecture Notes in Computer Science; volume 10878).

15. Brookes M. Voicebox: speech processing toolbox for MATLAB [Internet; cited 2019 April 24]. Available from: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

16. Xu Y, Prom-on S. Toward invariant functional representations of variable surface fundamental frequency trajectories: synthesizing speech melody via model-based stochastic learning. *Speech Communication.* 2014;57:181–208. DOI: 10.1016/j.specom.2013.09.013.

17. Weierstrass K. *Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen einer reellen Veränderlichen Sitzungsberichteder.* Berlin: Königlich Preussischen Akademie der Wissenschaften zu Berlin; 1885. p. 633–639.

18. Cabral JP, Kane J, Gobl C, Carson-Berndsen J. Evaluation of glottal epoch detection algorithms on different voice types. In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (INTERSPEECH-2011); 2011 August 27–31; Florence, Italy.* [S. l.]: International Speech Communication Association; 2011. p. 1989–1992.

19. Optimizing Nonlinear Functions – MATLAB and Simulink [Internet; cited 2019 April 20]. Available from: https://www.mathworks.com/help/matlab/math/optimizing-nonlinear-functions.html.

20. Xu Y, Prom-on S. What is PENTAtrainer2? [Internet; cited 2019 April 20]. Available from: http://www.homepages.ucl.ac.uk/~uclyyix/PENTAtrainer2/.

21. Prom-on S, Xu Yi. The qTA toolkit for prosody: learning underlying parameters of communicative functions through modeling. In: Hasegawa-Johnson M, editor. *Proceedings of Speech Prosody 2010.* 2010;100034:1–4.

22. Chen JH, Kao YA. Pitch marking based on an adaptable filter and a peak-valley estimation method. *Computational Linguistics and Chinese Language Processing.* 2001;6(2):31–42.

23. Boersma P, Weenink D. Praat: Doing phonetics by computer [Internet; cited 2019 May 3]. Available from: http://www.fon.hum.uva.nl/praat/.

24. Babacan O, Drugman T, d'Alessandro N, Henrich N, Dutoit T. A comparative study of pitch extraction algorithms on a large variety of singing sounds. *Proceedings of International Conference on Acoustics, Speech and Signal Processing (CASSP'13); 2013 May 26–31; Vancouver, BC, Canada.* [S. l.]: IEEE; 2013. p. 7815–7819. DOI: 10.1109/ICASSP.2013.6639185.

25. Yin pitch estimator [Internet]. 2012 November 27 [cited 2019 August 28]. Available from: http://audition.ens.fr/adc/sw/yin.zip.

26. Prom-on S, Xu Yi. Discovering underlying tonal representations by computational modeling: a case study of thai. *Phonology Journal.* 2015;32(3):505–535.

27. Li Y, Tao J, Lai W, Xu X. Quantitative intonation modeling of interrogative sentences for Mandarin speech synthesis. *Speech Communication.* 2017;89:92–102. DOI: 10.1016/j.specom.2017.03.002.

28. Wang B, Xu Y, Ding Q. Interactive prosodic marking of focus, boundary and newness in Mandarin. *Phonetica.* 2018;75(1): 24–56. DOI: 10.1159/00045308.

29. Charpentier F, Stella M. Diphone synthesis using an overlap-add technique for speech waveforms concatenation. In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'86); 1986 April 7–11; Tokyo, Japan.* [S. l.]: IEEE; 1986. p. 2015–2018. DOI: 10.1109/ICASSP.1986.1168657.

30. Ching XXu, Yi Xu, Li-Shi Luo. A pitch target approximation model for F_0 trajectories in Mandarin. In: Ohala JJ, editor. *Proceedings of the 14th International Congress of Phonetic Sciences (ICPHS'99).* San Francisco: University of California; 1999. p. 2359–2362.

Received by editorial board 04.09.2019.