ISBN: 978-93-86878-17-5
22nd IIE International Conference on Computer,Electrical,Electronics & Communication Engineering (CEECE-2018)
Pattaya (Thailand) May 31-June 2, 2018

A Modified Teager Energy Algorithm for Tracking Fundamental Frequency of Speech

Siripong Potisuk

Department of Electrical & Computer Engineering, The Citadel School of Engineering, 171 Moultrie Street, Charleston, SC 29409 USA

Abstract: *A solution to the problem of tone classification in a Thai speech recognition system requires a fast and reliable method for pitch detection, estimation and tracking because each of the five Thai tones has a rather definite fundamental frequency (F0) contour height and shape. A new method for tracking of fundamental frequency speech based on a Modified Teager energy algorithm was proposed and successfully applied to F0 extraction of the five Thai tones spoken in isolation. The advantages of the proposed method include reduced computational complexities and improved resolution due to the availability of pitch estimates at every sample location. In this paper, the same algorithm has been extended and applied to F0 extraction of the five Thai tones in continuous speech. Preliminary results suggest comparable or better performance than the typical frame-based autocorrelation method.*

Keywords: *modified Teager energy algorithm, F0 extraction, instantaneous pitch frequency.*

1. Introduction

In the realm of intelligent speech signal processing, fundamental frequency or pitch extraction and tracking is considered to be one of the fundamental problems that continue to warrant serious investigation within the speech community. Intrinsic to any speech signal is its fundamental frequency (F_0) of voiced sound segments and is known perceptually as pitch frequency. In other words, F_0 is the acoustic correlate of pitch, and the ability of humans to perceive pitch is associated with this frequency that impinges upon the ears. F_0 is estimated as the reciprocal of the fundamental period (T_0) of a voiced sound segment, which is defined as the elapsed time between two successive laryngeal pulses generated by the vibration of vocal folds as air is pushed up from the lungs during phonation. The pitch frequency is time-varying in nature and its reliable estimation is considered one of the most difficult tasks in acoustic processing of speech, especially in the presence of environmental noise. Despite the fact that speech generation is a highly variable and convoluted process and F_0 extraction is no trivial task, high performance pitch detection, estimation, and tracking are currently being pursued by speech scientists and engineers nonetheless.

Pitch frequency determination process can roughly be classified into two broad types: frame-based and instantaneous methods. In frame-based processing, F_0 is computed using an analysis window of a certain interval of speech samples called frame and advancing across the input speech with or without overlapping adjacent frames. The underlying assumption is that the speech signal within a given frame is locally stationary. Depending on speaker's gender, a typical frame length is 15 to 25 ms with a frame step of 10 ms resulting in roughly 50 % overlapping of adjacent frames. On the other hand, the instantaneous frequency can be computed for every sample location of the input signal. Oversimplifications in terms of linearity of speech production and local stationarity are not made. Consequently, this make the instantaneous pitch extraction process more accurate than the traditional frame-based method because of the availability of pitch estimates at every sample location. This means that the accuracy of the instantaneous pitch extraction is increased because the resolution

https://doi.org/10.17758/DIRPUB3.IE0518002

between pitch estimates is reduced from the typical 10 ms down to the sampling step of 0.045 ms for a sampling rate of 22050 Hz.

Several approaches for pitch frequency determination process have been developed and reported in the literature with varying degree of success since the early 1970s. For the frame-based method, many algorithms were proposed such as short-time average magnitude difference function (AMDF) method [1], autocorrelation method [2], cepstrum method [3], simplified inverse filter tracking (SIFT) method [4], and sub-harmonic summation (SHS) method [5]. For the instantaneous method, recent advances include utilization of B-spline expansion [6], the Hilbert-Huang transform [7], glottal closure instants [8], ensemble empirical mode decomposition (EEMD) [9], the wavelet transform [10], and variational mode decomposition (VMD) [11]. Although these algorithms show improved performance in terms of accuracy and resolution, they suffer from increased complexities and, as a result, high computational cost. They certainly do not lend themselves to real-time or online F_0 extraction for pitch detection, estimation, and tracking.

2. Research Motivation

Pitch information from speech signals in terms of F_0 contours is useful for a wide range of applications including speech recognition/understanding, speaker verification, speech-based emotion classification, language identification, voice transformation/morphing, singing, music, and pathological voice processing. In this paper, a previously proposed method for instantaneous F_0 extraction from the speech signal based on modified Teager energy algorithm [12] is extended and applied to connected speech. The impetus for this research arose during an investigation on the problem of tone classification in a Thai speech recognition system and the need for a fast (i.e., real-time) and reliable method for pitch tracking.

It is well known that F_0 variations in speech contribute to prosody and segmental qualities in any languages. This is particularly significant in tone language (e.g., Chinese, Thai, Vietnamese, etc.) in which tone is a suprasegmental feature indicated by contrasting variations in F_0 at the syllable level. It signals differences in lexical meaning and is considered an important part of a speech recognition/understanding system. Since Thai is the main focus of this paper, it is imperative that Thai tonal system be described as follows. Thai has five contrasting lexical tones traditionally labelled *mid* (M), *low* (L), *falling* (F), *high* (H), and *rising* (R). The following examples illustrate the effect that tone has on meaning.

TABLE I: Five different Thai words with the same segmental sequence but carrying different tones

Tone	Phonemic transcription with diacritic symbol	Meaning
Mid (M)	/ k^{h}aa /	‘To get stuck’
Low (L)	/ k^{h}àa /	‘Galangal’
Falling (F)	/ k^{h}âa /	‘To kill’
High (H)	/ k^{h}áa /	‘To engage in trade’
Rising (R)	/ k^{h}ǎa /	‘leg’

The height and shape of average F_0 contours of the five Thai tones produced in isolation can be found in [13] and will not be presented in this paper because the focus is on tones in connected speech. However, it has been shown by perceptual investigations that F_0 height and shape carry sufficient information for high intelligibility of Thai tones [14].

A Thai speech recognition system cannot be successful without tone classification because tone affects the lexical identification of words. The problem of tone classification in connected Thai speech can be simply stated as finding the best sequence of tones given an input speech signal. Since tone is a property of the syllable, each tone is associated with a syllable of the utterance. Because the primary acoustic correlate of tone is F_0 and Thai has five distinct F_0 contours, the problem is to find the best possible combination of F_0 tonal contour patterns that closely match the input F_0 contour.

The design of a tone classifier involves F_0 contours extraction and pattern matching. This pattern-matching process is relatively easy for isolated words because the tones produced have rather definite F_0 contours height and shape. However, the tones produced on words in connected speech are much more difficult to identify. This is because there are several interacting factors affecting F_0 realization of tones: syllable structure, tonal assimilation, stress, and intonation. Detailed explanations for each factor can be found in [15] and is not the primary concern of this paper. Rather, the main focus is on the former, which is the process of automatically and reliably extracting the F_0 contours from the input speech signal. Moreover, real-time implementation issue must be taken into account as well.

As previously mentioned, the main focus of this research is on an application of a previously proposed method for instantaneous F_0 extraction from the speech signal based on modified Teager energy algorithm [12]. The rest of the paper is organized as follows. The previously proposed F_0 extraction algorithm is presented in the next section for easy referencing. The following section discusses performance and compares the results with those obtained from the autocorrelation method via the Microsoft freeware “Speech Analyzer”. Finally, conclusions and future works end the paper.

3. The Proposed F_0 Extraction Algorithm

This section describes a previously proposed method for instantaneous F_0 extraction from the speech signal based on modified Teager energy algorithm. The proposed pitch determination system is illustrated in Figure 1 below.

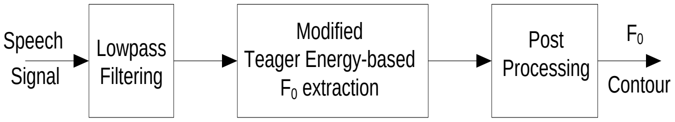

Fig. 1: The proposed F_0 extraction algorithm

As shown, the algorithm consists of three block operations: lowpass filtering, modified Teager energy-based F_0 extraction, and post processing.

3.1. Low-pass Filtering

Since speech generation is a highly variable and convoluted process, the input speech signal is first low-pass filtered to weaken the effect of speech resonances called formants. The digital low-pass filter is designed using the windowing method based on the Blackman-Harris window. It is of the finite-impulse-response (FIR) type of filter with sharp cut-off frequency at 100 Hz and of order 200. The gain characteristics of this low-pass filter are plotted in Figure 2 assuming the sampling frequency of 22050 Hz. Note that the cut-off frequency of 100 Hz is chosen based on the fact that the pitch frequency typically ranges from 60 to 200 Hz for male voice and 200 to 300 Hz for female voice. The non-ideal characteristic (i.e., a gradual roll-off) of the passband is also taken into consideration for the cut-off frequency selection.

3.2. Modified Teager Energy-based F_0 Extraction

The Teager energy operator (TEO), famously presented by Kaiser in [16], is a simple algorithm to obtain a measure of the energy of a simple (i.e., single component) sinusoidal oscillation and defined by the following relation:

$$E(n) = x^2(n) - x(n+1)x(n-1) = A^2 \sin^2(\Omega) \tag{1}$$

Note that $x(n) = A\cos(\Omega n + \phi)$ is the n^{th} sample of the signal representing the motion of an oscillatory body where Ω is the digital frequency in radians/sample given by $\Omega = 2\pi F/F_s$, F is the analog frequency, F_s is the sampling frequency, and ϕ is the arbitrary initial phase in radians. The parameters A, Ω, and ϕ are essentially constant. This non-linear energy-tracking operator has been modified and applied to the so-called AM-FM signals with time-varying amplitude envelope and instantaneous frequency. In [17], Maragos, Kaiser, and Quatieri proposed three discrete-time energy separation algorithms (DESA) based on the TEO, namely the DESA-1, DESA-1a, and DESA-2. Such AM-FM signals are very frequently used in communication systems. They also investigated its use in speech analysis to model time-varying speech resonances, particularly formant frequencies estimation and tracking.

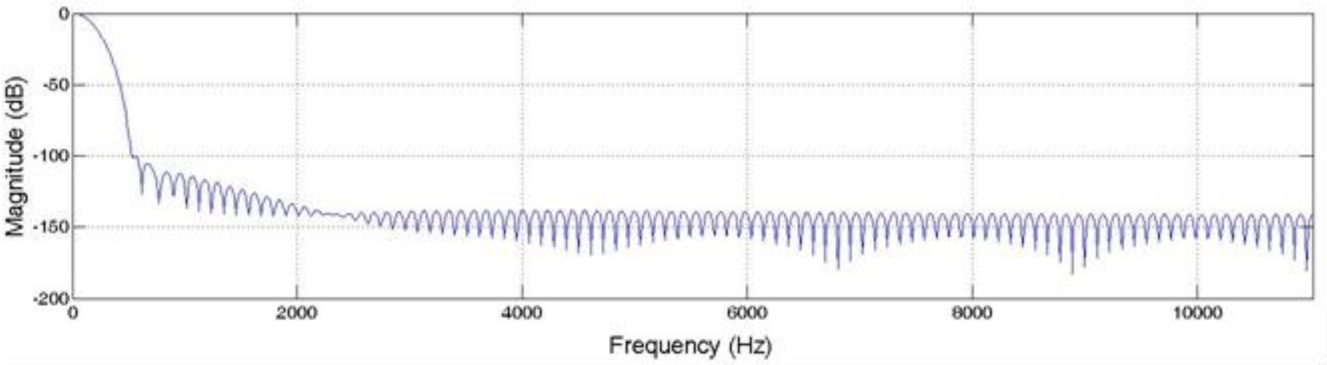

Fig. 2: The gain characteristics of the low-pass FIR filter applied to the input speech to weaken the effects of formants

In this paper, a new modification to the TEO is proposed for tracking the fundamental frequency of speech signals. Detailed derivation is given below by combining various shifted versions of the signal and their outputs from TEO to obtain a set of equations whose solution yields estimates of the amplitude and frequency signals. Starting from the cosine $x(n) = A\cos(\Omega n + \phi)$ with constant amplitude and frequency, it can be shown that

$$x(n+1) + x(n-1) = 2A\cos(\Omega)\cos(\Omega n + \phi) = 2\cos(\Omega)x(n).$$

$$\text{Or,} \qquad x(n+1) = 2\cos(\Omega)x(n) - x(n-1). \tag{2}$$

Substituting (2) into (1) yields

$$\begin{aligned} E(n) = A^2\sin^2(\Omega) &= x^2(n) - \{2\cos(\Omega)x(n) - x(n-1)\}\,x(n-1) \\ &= x^2(n) - 2\cos(\Omega)\,x(n)\,x(n-1) + x^2(n-1) \end{aligned} \tag{3}$$

Since the fundamental frequency of speech is below 500 Hz, $\Omega << \pi/4$, and thus, the following approximations: $\sin(\Omega) \approx \Omega$ and $\cos(\Omega) \approx 1 - \frac{\Omega^2}{2}$ can be used in (3) resulting in

$$\begin{aligned} E(n) = A^2\Omega^2 &= x^2(n) - 2\left(1 - \frac{\Omega^2}{2}\right)x(n)\,x(n-1) + x^2(n-1) \\ &= x^2(n) - 2\,x(n)\,x(n-1) + x^2(n-1) + \Omega^2 x(n)\,x(n-1) \\ &= [x(n) - x(n-1)]^2 + \Omega^2 x(n)\,x(n-1) \end{aligned} \tag{4}$$

Solving (4) for Ω and *A* in terms of instantaneous parameters by using the definition of *E* (*n*) from (1) yields

$$\begin{aligned} |\Omega(n)| &= \sqrt{\frac{E(n) - [x(n) - x(n-1)]^2}{x(n)\,x(n-1)}} \\ &= \sqrt{\frac{x^2(n) - x(n+1)x(n-1) - [x(n) - x(n-1)]^2}{x(n)\,x(n-1)}}, \end{aligned} \tag{5}$$

$$|A(n)| = \sqrt{\frac{E(n)}{\Omega^2(n)}} = \sqrt{\frac{x^2(n) - x(n+1)x(n-1)}{\Omega^2(n)}}. \tag{6}$$

Equations (5) and (6) can be used to extract the FM signal (instantaneous frequency) and the AM signal (amplitude envelope), respectively from the low-pass filtered input speech signal.

3.3. Post-processing Operation

After equation (5) is used in extracting the instantaneous F_0 from the low-pass filtered input speech signal, the resulting F_0 contour is very choppy containing several spurious minima and maxima. By taking the range of possible speech F_0 into consideration, several criteria from [7] are adopted to eliminate out-of-range F_0 values. They are: (i) the instantaneous frequencies outside the range of 60-500 Hz are set to zero; (ii) the instantaneous frequencies with a variation larger than 100 Hz within 5 ms are set to zero; and (iii) the instantaneous frequencies with corresponding amplitude less than ten percent of the maximum are also set to zero. In addition, the resulting F_0 contour is further smoothed by moving average filtering. To avoid too much smoothing, the window length is chosen to be 221 samples, 110 samples on either side of the current value of the F_0 contour. This window length represents a 10-ms length of samples for the sampling rate of 22050 Hz. Finally, the portions of the contour corresponding to the unvoiced sound segments in the input speech are set to zero. The voiced/unvoiced (V/UV) detection is carried out on the original input speech signal using the RMS energy and zero-crossing contours.

4. Experiment and Results

4.1. Speech Materials

The description of the experiment carried out to test the viability of this proposed approach to F_0 extraction in Thai connected speech emphasizes its preliminary nature, yet with promising results. The speech corpus contains 21 sentences (listed in the appendix) designed to highlight the occurrences of all 21 possible leading consonants in various syllable structures. The algorithm was tested on 84 (21 sentences × 4 speakers) connected speech utterances read by 2 male and 2 female speakers in the 22–35 age range. All subjects were mono-dialectal speakers of Standard Thai. They were free of any speech or hearing disorders by self-report based on a screening interview and as later judged by the investigator during the recording session.

Recordings were made in a quiet office using the recording feature of Microsoft freeware "Speech Analyzer" version 3.1.0 installed on a Dell Latitude laptop computer. The digitization is at a sampling rate of 22050 Hz by means of a 16-bit mono A/D converter. Speakers were seated and wore a regular Logitech computer headset with microphone maintained at a distance of 5 cm from the lips. Each speaker was asked to read a total of 21 sentences at their conversational speaking rate. Before the recording session began, the speakers were allowed to familiarize themselves with the sentences. Each session lasted about 20 minutes.

4.2. Analysis Results

Figure 3 shows three rows of plots resulting from the application of the method to the speech of one of the male subjects. The plots represent the original input speech (first row), the low-pass filtered input speech (second row) superimposed with the extracted amplitude envelop from equation (6), and the smoothed instantaneous F_0 contour in solid line corresponding to the voiced segments (i.e., vowels and voiced consonants) of the input speech (last row). The third plot indicates the results of F_0 extraction for utterance no. 3 of the speech corpus, which contains all voiced segments except for the initial consonant, k^h, of the second word. Visual inspection reveals that the shape of F_0 contours of every syllable closely resemble those of the tones spoken in isolation. However, the F_0 contours of unstressed syllables are shorter in length and cannot reach the pitch target during the final 30% of the contour. Nevertheless, the plot clearly shows the correct sequence of tones (FMRFMMMRL) extracted from the input speech.

It is important to note again that the horizontal axis represents the sample index because the calculations are done at every sample of the input speech (i.e., a step of one sample) with a window of length three samples as previously mentioned in section III. This is considered an advantage over the traditional frame-based method because the availability of pitch estimates at every sample location allows an increase in accuracy and resolution.

To assess performance of the proposed method, the smoothed F_0 contour (in + marker) resulting from the frame-based autocorrelation method via the Microsoft freeware "Speech Analyzer" was superimposed onto the plot of the smoothed F_0 contours from the algorithm (as seen in the last row of figure 3) to allow for visual comparison. A relatively close match of the overall pattern in terms of height and shape was observed for the utterance. No statistical analysis was performed to quantify the similarities of the contours in terms of the Pearson correlation coefficient because the two methods differ in terms of how the smoothed F_0 contours were obtained, i.e., frame-based vs. sample-based.

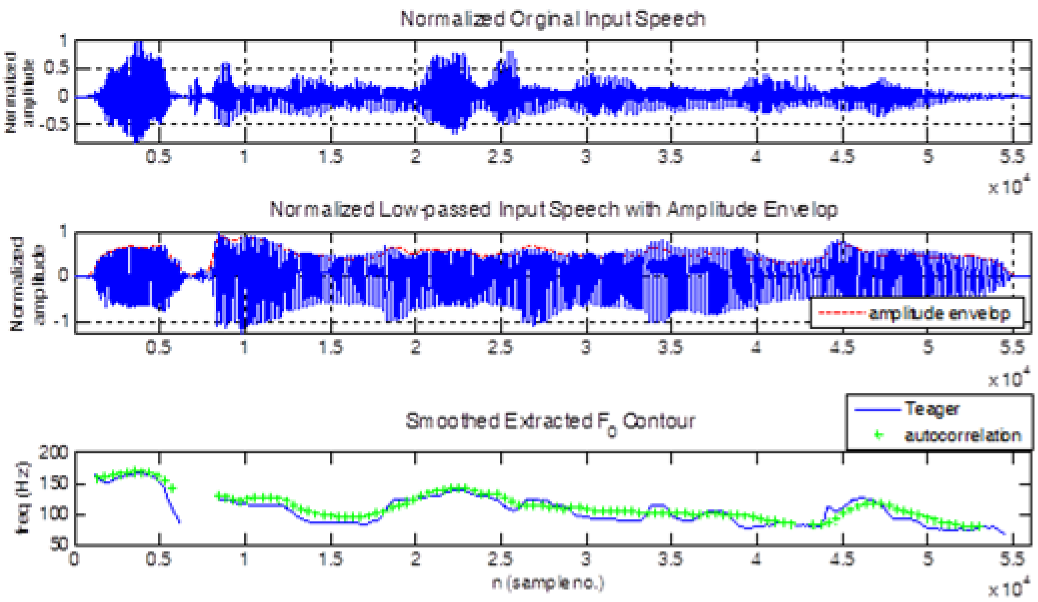

Fig. 3 Results from applying the proposed F_0 extraction algorithm to utterance no. 3 of the appendix. (first row) The original speech signal; (second row) The low-pass filtered input speech with amplitude envelop; (third row) The smoothed F_0 contour corresponding to the voiced segments of the input utterance superimposed with that from the autocorrelation method for visual comparison.

5. Conclusions and Future Works

This paper has presented an extension and application of a previously proposed method for instantaneous F_0 extraction from the speech signal. The method is based on Modified Teager Energy Algorithm. Preliminary results on its application to Thai connected speech suggest comparable, if not better, performance when compared with the autocorrelation method. However, the advantage of the proposed method lies in the fact that it is easy to implement and can lend itself to online or real-time implementation, which is crucial in the development of a tone classifier in speech recognition systems for tone language. In addition, it is superior in terms of being able to provide pitch estimates at every sample location rather than at every block or frame of data resulting in increased resolution and accuracy. However, this study is still in an early stage of investigation and implementation. A meaningful quantitative performance measure to allow for further comprehensive statistical analysis and thorough performance evaluation is needed to ascertain the usefulness of the method. Moreover, although recordings of carefully read speech, which may not truly represent continuous speech, were used to evaluate performance of the algorithm in this paper, the results obtained were very promising. As a result, the goal is to attempt to extend the method to spontaneous Thai speech in order for the method to be useful for real-time implementation. On-going experiment is being planned and soon conducted.

6. Appendix: Speech Stimuli

The following is a list of utterances comprising the speech materials used in the experiment. Phonemic transcription and English translation accompany each of the Thai utterance.

1. สมศรีอยากสักลายเสือที่แขนข้างขวา
   sǒmsǐı jàɑk sàk lɑɑı sɯ̌ɑ tʰîı kʰɛ̌ɛn kʰâɑŋ kʰwǎɑ
   ‘SomSri wants to tattooed a tiger pattern on her right arm.’
2. อรอนงค์ออกอาการอึดอัดเมื่ออ๊อตออดอ้อน
   ʔɔɔnʔɑnoŋ ʔɔɔ̀k ʔɑɑkɑɑn ʔɯ̀tʔàt mɯ̂ɑ ʔɔɔ́t ʔɔɔ̀tʔɔ̂ɔn
   Orn-anong felt uncomfortable when Aut pleaded with her.
3. แว่วคำหวานว่าวันวานยังหวานอยู่

wɛ̂ɛʊ kʰam wǎan wâa wan waan jaŋ wǎan jʊ̀ʊ
‘(I) vaguely hear the sweet talk that things are still as sweet as yesterday.’

4. กู้เกียรติเป็นคนกร้าวแกร่งกว่าที่คิด
kʊ̂ʊkìat pen kʰon kraâʊ krɛ̀ŋ kwàa tʰîɪ kʰít
‘Kookiat is tougher than I thought.’

5. ทหารถอดพระที่ห้อยคอไว้บนหิ้ง
tʰahǎan tʰɔ̀ɔt pʰrá tʰîɪ hɔ̂ɪ kʰɔɔ wáɪ bon hîŋ
The soldier took off a neck chain of amulets and put it on the shelf.’

6. ชาญชัยช่างเชื่องช้าเฉกเช่นชายชื่อชวน
cʰaancʰaɪ cʰâŋ cʰɯ̂aŋcʰ́aa cʰèek cʰên cʰaaɪ cʰɯ̂ɯ
cʰʊʊan
‘ChanChai was as slow as the man named Chuan.’

7. ยายเมี้ยนย่างเนื้อวัวไว้ยำเย็นนี้
jaaɪ mían jâaŋ nɯ́a wáɪ jaam jen níɪ
‘Grandma Mian is roasting beef for making salad tonight.’

8. ท่องเที่ยวทั่วไทยไปกับบริษัททัวร์สยาม
tʰɔ̂ŋtʰîaʊ tʰʊ̂a tʰaɪ paɪ kàp tʰʊa sajǎam
‘Travel all over Thailand with Siam Tour company.’

9. สถานบริบาลเด็กเล็กอยู่แถวบ้านที่บางบัว
s̀atʰǎanbarɪbaan dèklék jʊ̀ʊ tʰɛ̌ʊ bâan tʰîɪ baaŋbʊa
‘The daycare center is located in the vicinity of our BangBua home.’

10. รายการ คุยคุ้ยข่าว คนไม่ค่อยชอบดู
raaɪkaan kʰʊɪkʰʊ́ɪkʰàaɪ kʰon mâɪ kʰɔ̂ɪ cʰɔ̂ɔp dʊʊ
‘People don’t like to watch the *KhuiKhuiKhaw* show.’

11. เฟื่องฟ้าชอบแดดจัดแต่ไม่ชอบฝน
fɯ̂aŋfáa cʰɔ̂ɔp dɛ̀ɛt càt tɛ̀ɛ mâɪ cʰɔ̂ɔp fǒn
Bougainvillea likes full sun, but not rain.’

12. จันจิราเป็นคนจู้จี้จุกจิกเอาแต่ใจ
cancìraa pen kʰon cʊ̂ʊcîɪ cʊ̀kcìk ʔaʊ tɛ̀ɛ caɪ
‘Chanchira is a self-centered and nitpicking person.’

13. เด็กๆ ดึงดันจะดูแลเด่นดนัยด้วยกัน
dèkdèk dɯŋdan cà dʊʊlɛɛ dèndànaɪ dʊâɪ kan
‘Children insists on taking care of Dendanai together.’

14. ตฤณตรอมใจเพราะมองว่าตัวเองต่ำต้อย
trɪn trɔɔm caɪ pʰrɔ́ mɔɔŋ wâa tʊa ʔeeŋ tàmtɔ̂ɪ
Trin is depressed from his inferiority complex.’

15. ป้าแป้นถูกส่งไปประจำที่ด่านปอยเปต
pâa pɛ̂n tʰʊ̀ʊk sòŋ paɪ pràcam tʰîɪ dàan pɔɔɪpèet
‘Aunt Pan is assigned a post at the Poipet border control.’

16. ลุงหาญชอบล้อลลิตาเรื่องลุกลี้ลุกลน
lʊŋ hǎan cʰɔ̂ɔp lɔ́ɔ lálítaa rɯ̂aŋ lʊ́klíɪlʊ́klon
‘Uncle Hahn likes to tease Lolita’s clumsiness.’

17. เรารู้สึกรันทดหดหู่เมื่อได้ยินเรื่องราวร้ายๆ
raʊ rʊ́ʊsɯ̀k rantʰót hòthʊ̀ʊ mɯ̂a dâɪ jɪn rɯ̂aŋraaʊ ráɪ
ráɪ
‘We feel depressed and dejected when hearing terrible news.’

18. แพรวพรรณขึ้นทอดผ้าที่งานศพพวงพยอม
pʰrɛɛʊpʰan kʰɯ̂n tʰɔ̂ɔt pʰâa tʰîɪ ŋaan sòp pʰʊʊaŋ
pʰajɔɔm
‘Phraewphan laid a yellow monk robe at Phuangphayom’s funeral.

19. นลินีชอบทั้งขนมจีนน้ำยาและน้ำเงี้ยว

nálínɪɪ cʰɔ̂ɔp tʰáŋ kʰànǒmcɪɪn náamjaa lɛ́ náamŋɪáʊ
‘Nalinee likes both Namya and Nam-ngiaw rice noodle dish.’

20. เธองอแงเพราะหงุดหงิดที่ไม่มีใครงอนง้อ

tʰəə ŋɔɔŋɛɛ pʰrɔ́ ŋʊ̀tŋìt tʰîɪ mâɪ mɪɪ kʰraɪ ŋɔɔnŋɔ́ɔ
She made a big fuss because everyone ignores her.’

21. เขาซื้อมะม่วงมันมามากมายเมื่อตอนปีใหม่

kʰǎʊ sɯ́ɯ mámʊ̂aŋman maa mâakmaaɪ mɯ̂a tɔɔn pɪɪ màɪ
‘He bought a lot of nutty-flavored mangoes during the New Year’s celebration.

7. Acknowledgment

This research is supported in part by a faculty research grant from the Citadel Foundation. The author wishes to acknowledge the assistance and support from Mrs. Suratana Trinratana, Vice President & Chief Operation Officer, and her staff of the Toyo-Thai Corporation Public Company Limited, Bangkok Thailand, during the speech data collection process. The author would also like to thank the Citadel Foundation for its financial support in the form of a presentation grant.

8. References

[1] M. Ross, H. Shaffe, A. Cohen, R. Freudberg, et al., “Average magnitude difference function pitch extractor,” *IEEE Trans. Acoustics Speech Signal Processing*, vol. 22, no. 5, pp. 353–362, Oct. 1974.

https://doi.org/10.1109/TASSP.1974.1162598

[2] L.R. Rabiner, “On the use of autocorrelation analysis for pitch determination,” *IEEE Trans. Acoustics Speech Signal Processing*, vol. 25, no. 1, pp. 24–33, Feb. 1977.

https://doi.org/10.1109/TASSP.1977.1162905

[3] A. M. Noll, “Cepstrum pitch determination,” *Journal of Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

https://doi.org/10.1121/1.1910339

[4] J. Markel, “The SIFT algorithm for fundamental frequency estimation,” *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 5, pp. 367–377, Dec. 1972.

https://doi.org/10.1109/TAU.1972.1162410

[5] D.J. Hermes, “Measurement of pitch by subharmonic summation,” *The Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 257–264, Jan. 1988.

https://doi.org/10.1121/1.396427

[6] B. Resch, M. Nilsson, A. Ekman, and W.B. Kleijn, “Estimation of the instantaneous pitch of speech,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 813–822, Mar. 2007.

https://doi.org/10.1109/TASL.2006.885242

[7] H. Huang and J. Pan, “Speech pitch determination based on Hilbert-Huang transform,” Signal Processing, vol. 86, no.4, pp. 792–803, Apr. 2006.

https://doi.org/10.1016/j.sigpro.2005.06.011

[8] P.A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol.15, no.1, pp. 34–43, 2007.

https://doi.org/10.1109/TASL.2006.876878

[9] G. Schlotthauer, M.E. Torres, and H.L. Rufiner, “A new algorithm for instantaneous F_0 speech extraction based on ensemble empirical mode decomposition,” in *Proc. 17th EURASIP Signal Processing Conference*, Glasgow, Scotland, UK, 2009, pp. 2347–2351.

[10] Y. Li, B. Xue, H. Hong, and X. Zhu, “Instantaneous pitch estimation based on empirical wavelet transform,” in *19th International Conference on Digital Signal Processing*. IEEE, 2014, pp. 250–253.

https://doi.org/10.1109/ICDSP.2014.6900838

[11] A. Upadhyay and R. B. Pachori, “A new method for determination of instantaneous pitch frequency from speech signals.” in *Proc. IEEE Signal Processing and Signal Processing Education Workshop (SP/SPE)*, 2015, pp. 28–33.

https://doi.org/10.1109/DSP-SPE.2015.7369574

[12] S. Potisuk, “A new method for instantaneous F0 speech extraction based on modified Teager energy algorithm,” in Proc. of the 2016 International Conference on Electronics, Computer and Information Technology (ECIT-16), 8-9 August 2016, Pattaya, Thailand.

[13] A. S. Abramson, “The vowels and tones of standard Thai: acoustical measurements and experiments,” *International Journal of American Linguist*, vol. 28-2, Part III, no.20, 1962.

[14] J. T. Gandour, “Tone perception in Far Eastern languages,” *Journal of Phonetics*, vol. 11, pp. 149–175, 1983.

[15] S. Potisuk, “Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 7, no. 1, pp. 95–102, Jan. 1999.

https://doi.org/10.1109/89.736336

[16] J. F. Kaiser, “On a simple algorithm to calculate the ‘energy’ of a signal,” in *Proc. IEEE International Conference on Acoustic Speech and Signal Processing*, 1990, pp. 381–384.

https://doi.org/10.1109/ICASSP.1990.115702

[17] P. Maragos, J. F. Kaiser, and T. F. Quatieri, “Energy separation in signal modulations with applications to speech analysis,” IEEE Trans. on Signal Processing, vol. 41, no.10, pp. 3024–3051, Oct. 1993.

https://doi.org/10.1109/78.277799