A New Method for Instantaneous F_0 Speech Extraction Based on Modified Teager Energy Algorithm

Siripong Potisuk

***Abstract*—**A new method for instantaneous F_0 extraction from the speech signal based on Modified Teager energy algorithm is presented. Applications to Thai tone classification are considered. The need for a fast and reliable method for pitch detection, estimation and tracking is crucial for an investigation on the problem of tone classification in a Thai speech recognition system. The advantages of the proposed method include reduced computational complexities and improved resolution due to the availability of pitch estimates at every sample location. Preliminary results on F_0 extraction of the five Thai tones spoken in isolation suggest comparable or better performance than the typical frame-based autocorrelation method.

***Keywords*—**modified Teager energy algorithm, F_0 extraction, instantaneous pitch frequency.

I. Introduction

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 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.

II. 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 new method for instantaneous F_0 extraction from the speech signal based on modified Teager energy algorithm is proposed. 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.

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

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 in detail 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)	/ kʰaa /	‘To get stuck’
Low (L)	/ kʰàa /	‘Galangal’
Falling (F)	/ kʰâa /	‘To kill’
High (H)	/ kʰáa /	‘To engage in trade’
Rising (R)	/ kʰǎa /	‘leg’

Adapted from [12], average F_0 contours of the five Thai tones produced in isolation are shown in figure 1 below. Perceptual investigations have revealed that F_0 height and shape carry sufficient information for high intelligibility of Thai tones [13].

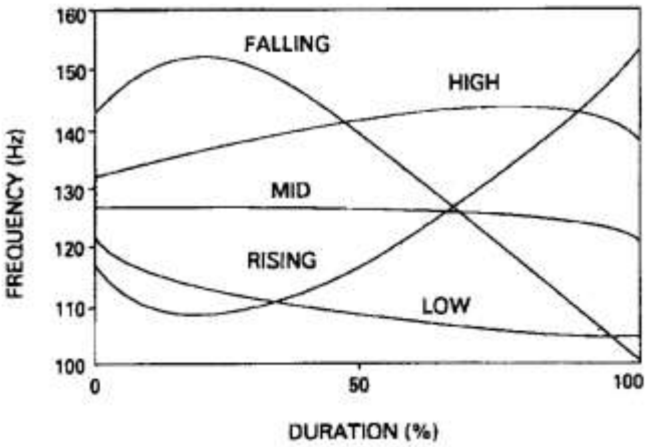

Fig. 1: Average F_0 contours of the five Thai tones produced in isolation by a male speaker

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 are very similar to those in figure 1. However, the tones produced on words in continuous 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 [14] 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 a proposed new method for instantaneous F_0 extraction from the speech signal based on modified Teager energy algorithm. The rest of the paper is organized as follows. The proposed F_0 extraction algorithm is presented in the next section. 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.

III. THE PROPOSED F_0 EXTRACTION ALGORITHM

This section describes a new 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 2 below.

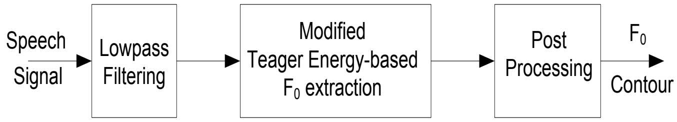

Fig. 2: The proposed F_0 extraction algorithm

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

A. 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 3 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.

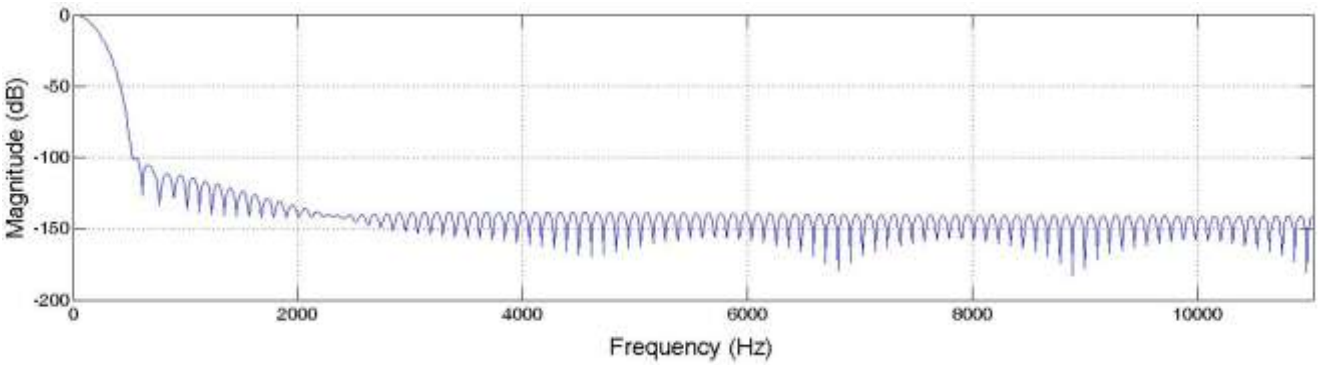

Fig. 3: The gain characteristics of the low-pass FIR filter applied to the input speech to weaken the effects of formants

B. Modified Teager energy-based F_0 extraction

The Teager energy operator (TEO), famously presented by Kaiser in [15], 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:

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

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 [16], 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.

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

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

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

Substituting (2) into (1) yields

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

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

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

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

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

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

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.

C. 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.

IV. Experiment and Results

A. Speech materials

The description of the experiment carried out to test the viability of this proposed approach to F_0 extraction emphasizes its preliminary nature, yet with promising results. The speech corpus contains five monosyllabic words of the same phonetic sequence ‘khaa’ with five possible tones. The algorithm was tested on 50 (5 words × 5 tokens × 2 subjects) monosyllabic words spoken in isolation without any carrier frame by one male and one female speaker in the 22-35 age range. Both subjects are 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 25 monosyllabic words at their conversational speaking rate. Each session lasted about 5 minutes.

B. Analysis Results

Figure 4 shows four rows and five columns of plots resulting from the application of the proposed method to the speech of the male subject. The four plots in each column represent the original input speech (first row), the low-pass filtered input speech (second row), the extracted instantaneous F_0 contour (third row), and the smoothed F_0 contour corresponding to the voiced segment or vowel of the input speech (last row). The five columns of plots indicate the results of F_0 extraction for each of the five tones starting from the ‘Mid’ tone on the left to the ‘Rising’ tone on the right.

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 3. 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 measure performance of the proposed method, visual comparison of the smoothed F_0 contours for each of the five tones (as seen in the last row of figure 4) was performed against the resulting plots of F_0 contour from the autocorrelation method via the Microsoft freeware “Speech Analyzer”. A relatively close match of the overall pattern in terms of height and shape was observed for all five tones. No statistical analysis was performed to quantify the similarities of the contours in terms of the Pearson correlation coefficient.

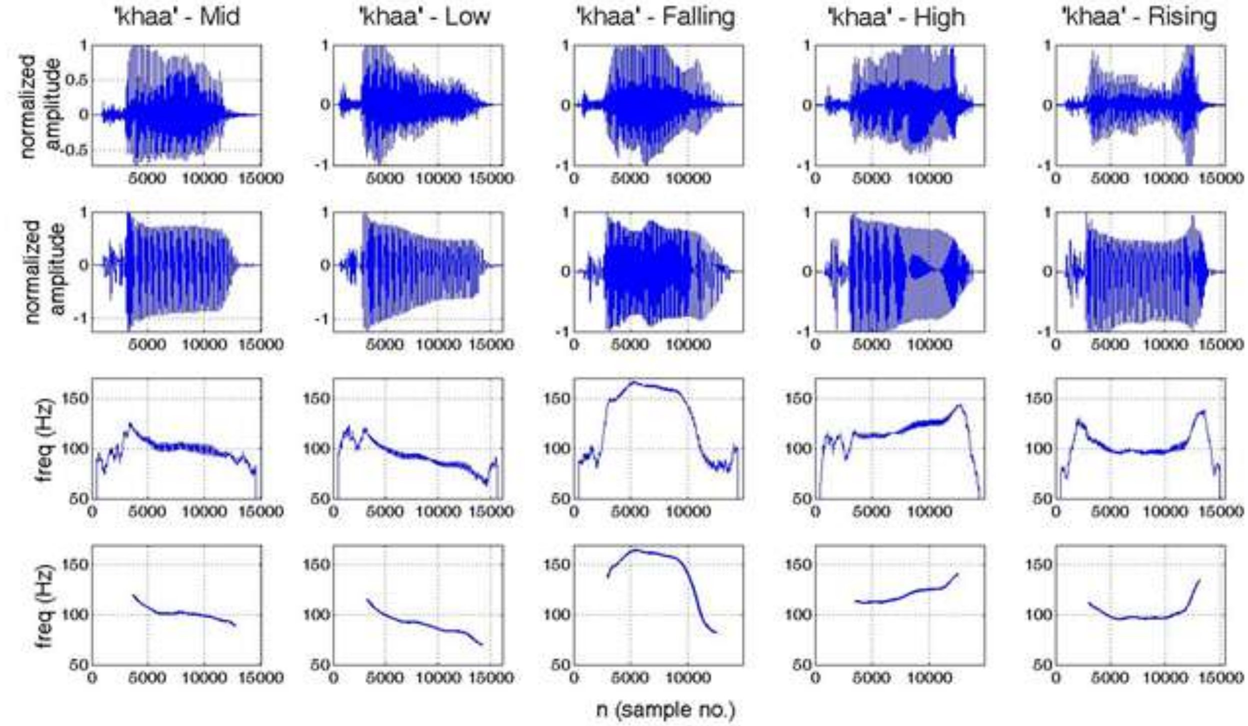

Fig. 4 Results from applying the proposed F_0 extraction algorithm to the five Thai monosyllabic words with the same phonetic sequence representing all five tones. (first row) The original speech signal; (second row) The low-pass filtered speech signal; (third row) The resulting instantaneous F_0 contours; (last row) The smoothed F_0 contours corresponding to the voiced segment or the vowel sound ‘aa’.

V. Conclusions and Future Works

This paper has presented a new 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 the five tones of Thai monosyllabic words spoken in isolation suggests 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. Further comprehensive statistical analysis and thorough performance evaluation is needed to ascertain its usefulness. Although the focus of this paper is on the application of the algorithm to isolated speech, the goal is to attempt to extend the method to Thai continuous speech. On-going experiment is being planned and soon conducted.

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 sup-port in the form of a research presentation grant.

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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.