# CHAPTER 3 THEORY AND SOLUTION OF LOW CALCULAION COMPLEXITY TONE RECOGNITION

In this chapter, an overview of reduced complexity tone recognition, theory, and solution used to recognize tones are presented.

This thesis aims to reduce the power consumption of tone recognition systems. In order to decrease power consumption, the complexity of the tone recognition has to be reduced.

We propose to reduce tone recognition complexity by extracting fundamental frequency ($F_0$) from a new magnitude difference function, called vowel-MDF. The proposed tone recognition method is described as follows:

## 3.1 Overview of Reduced Complexity Tone Recognition

The reduced complexity tone classifier consists of three processes: feature extraction, $F_0$ estimation, and tone decision as shown in Figure 3.1. Each process is described as follows: firstly, in the feature extraction process, word segmentation separates each syllable using the zero-crossing method, then the vowel signals are detected from the energy of the input signals, after which the pitch period is estimated from the first local minimum of magnitude difference function (MDF); secondly, the $F_0$ contour of each syllable is extracted in $F_0$ estimation process; finally, the tone result classifies tone by a decision tree algorithm which is regulated from the $F_0$ contour characteristics.

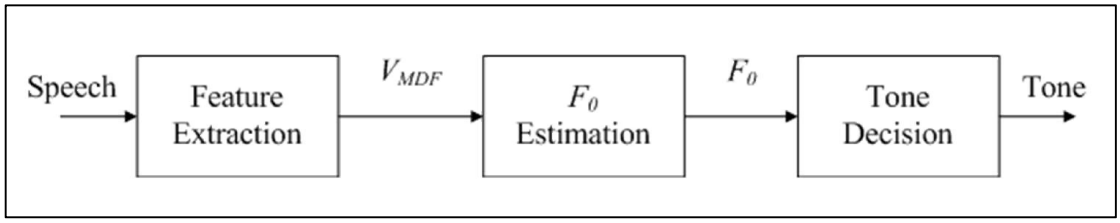


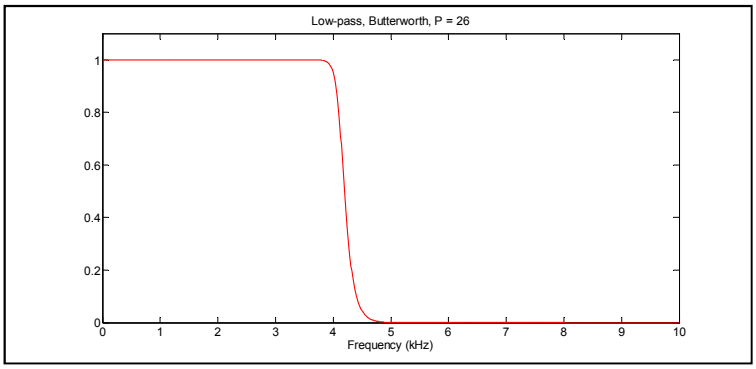

**Figure 3.1** An overview of reduced complexity tone recognition

The mathematical methods calculated in the feature extraction, $F_0$ estimation, and tone decision are described in the rest of this chapter.

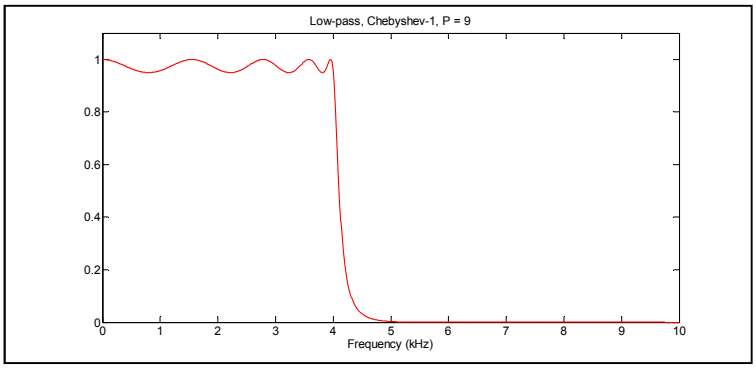
## 3.2 Feature Extraction

In the feature extraction process, the vowel parts and vowel-MDF ($V_{MDF}$) are estimated. We design this process by simple algorithms and thus its calculation costs are low.

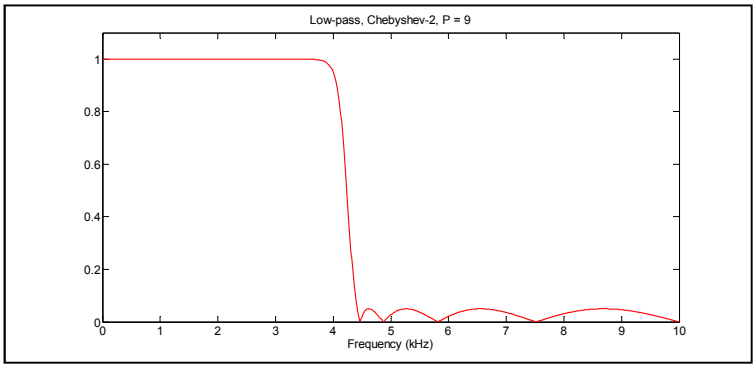
Voiced signals indicate more energy in lower frequency whereas unvoiced signals, such as noise, are not much energy in higher frequency [57]. Since this thesis aims to detect feature for $F_0$, we therefore apply a band-pass filter to reduce environmental noise and also filter input signals including the possible pitch frequency range.



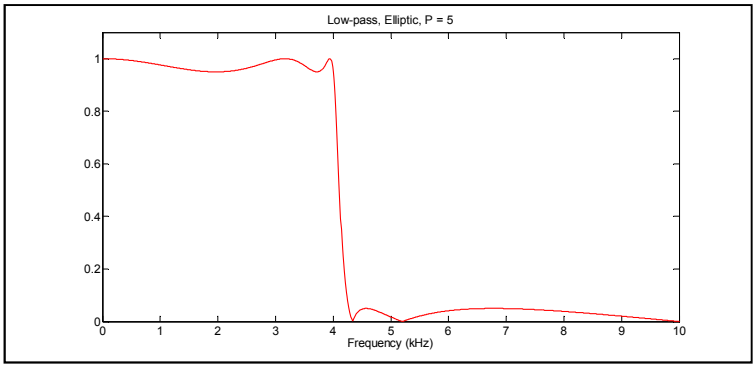

(a)




(b)




(c)




(d)

**Figure 3.2** Digital filter low-pass filters (a) Butterworth low-pass filter with $P = 26$, (b) Chebyshev type 1 low-pass filter with $P = 9$, (c) Chebyshev type 2 low pass filter with $P = 9$, (d) Elliptic low-pass filter with $P = 5$

The recursive filter transfer function in z-domain [58] is defined as:

$$H(z) = \frac{\sum_{i=0}^{P} b_i z^{-i}}{1 + \sum_{i=0}^{P} a_i z^{-i}} \quad ; a_i \neq 0 \tag{1}$$

where $P$ is filter orders, $b_i$ is feedforward filter coefficients, and $a_i$ feedback filter coefficients.

Figure 3.2 indicates digital low-pass filters of 4 different prototypes (a) Butterworth, (b) Chebyshev type 1, (c) Chebyshev type 2, and (d) elliptic filters. In figure 3.2, the elliptic filter provides the smallest filter order for the same specification, and also shows the narrowest transition width for the same filter order, compared with Butterworth and Chebyshev filters [59-60]. We therefore filter the input speech out by using elliptic filter. This results in computation time reduction.

In telecommunications applications, the input signals including the frequency range of the human voice are generally limited to between 300 and 3,400 Hz [61]. We therefore filter the input signals by using a band-pass filter where $P$ is equal to 4. The design specification for the digital band-pass filter is shown in Figure 3.3. Due to the frequency range of the human voice, we design the pass-band frequencies [$f_{p1}$,$f_{p2}$] between 300 and 3,400 Hz while the stop-band frequencies [$f_{s1}$,$f_{s2}$] are between 200 and 3500 Hz. $f_s$ is the sampling frequency which is set to 11.025 kHz. $A_p$ and $A_s$ are the ripple in the pass-band and stop-band which are set to 0.001 and 30 dB. Figure 3.4 indicates the frequency response of the proposed digital band-pass filter generated from the specifications.

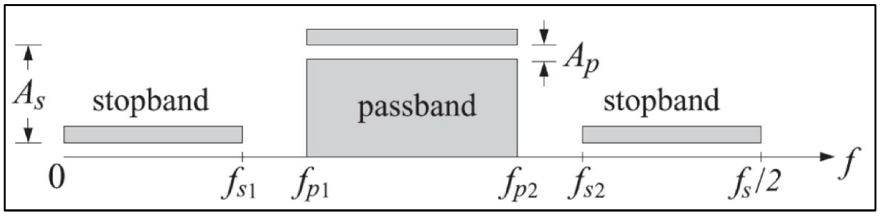



**Figure 3.3** Specifications of digital band-pass filter [62]

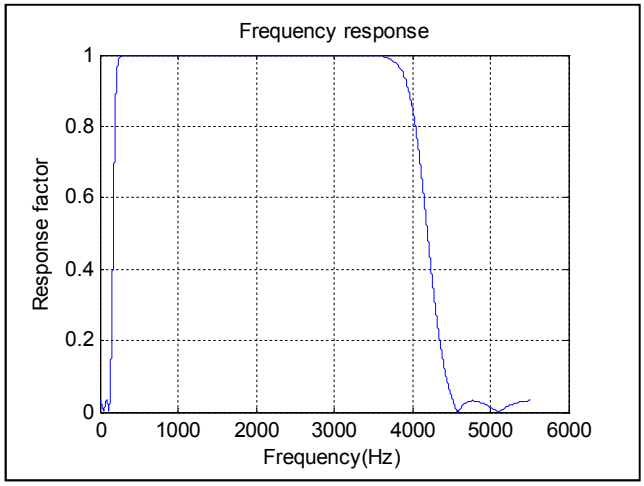



**Figure 3.4** Frequency response of digital band-pass filter

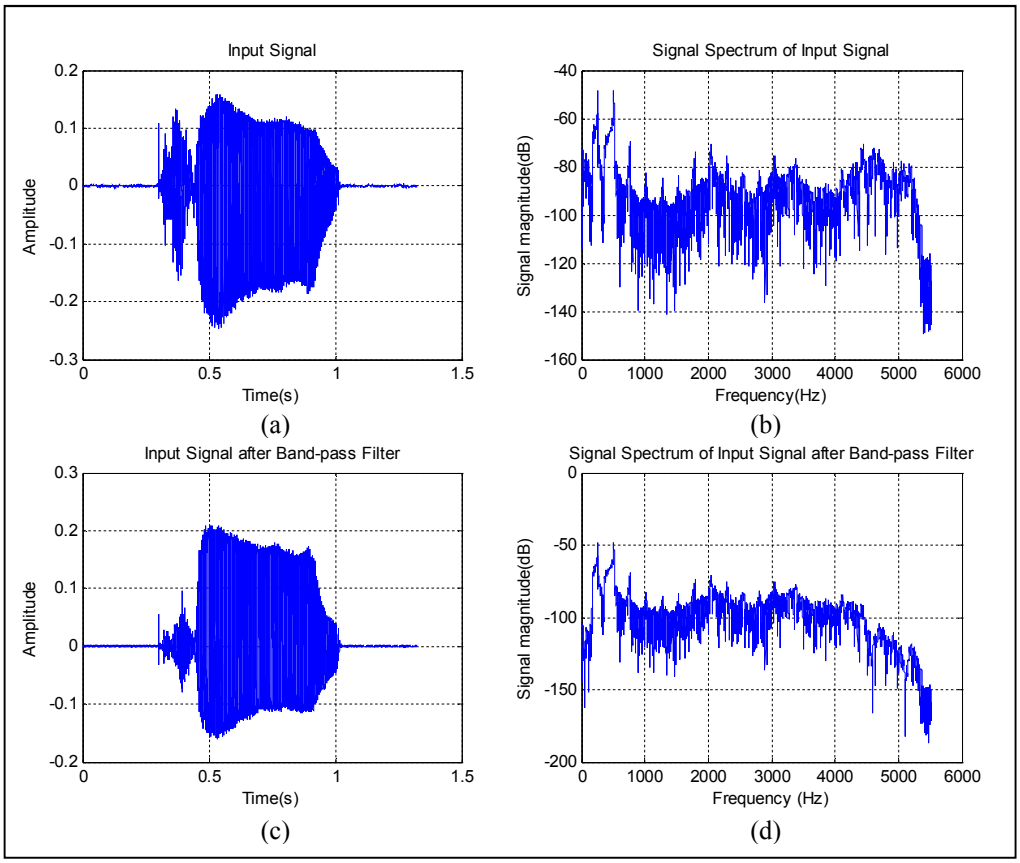



**Figure 3.5** The word "*chûu*" before (top) and after (bottom) elliptic low-pass filter

Figure 3.5 (a) shows the input signal of the word "*chûu*". A signal spectrum of the input signal is shown in (b). After we apply the digital band-pass filter in Figure 3.4 for the input signal, we then get a smoother input signal in (c). This also results in the lower signal magnitude of signal spectrum shown in (d).

Next, we detect start and end points of the voiced speech. The zero crossing method [63] and the energy of an input speech are used to detect the voiced speech. Subsequently, the frame including vowel is determined from the frame which has higher energy than the energy threshold. In this research, $j$ is the index of the $j$-th frame where $j$ = 1, 2, 3, ..., $k$ and $k$ is the frame ending. The frame length is $N$ = 110 (10 ms) and the frame shift is $M$ =55 (5 ms). Every frame is applied with the Hamming window. The zero-crossing rate method is defined by

$$Z(j) = \sum_{i=2}^{2M} \left| \mathrm{sgn}\left(\tilde{s}[x+i]\right) - \mathrm{sgn}\left(\tilde{s}[x+i-1]\right) \right| \tag{2}$$

where $x = (j\text{-}1)M$ and the function of $\mathrm{sgn}(\tilde{s}[n])$ is defined as

$$\mathrm{sgn}(\tilde{s}[i]) = \begin{cases} 1, s(i) \geq 0 \\ 0, s(i) < 0. \end{cases}$$

It is considered that voiced speech has a low $Z(j)$ compared with unvoiced speech. Accordingly, when the calculated $Z(j)$ is lower than the zero-crossing threshold ($T_z$),

this method determines that its frame includes voiced speech. In this research, $T_z$ is calculated from the minimum value between 40 and a silence threshold ( $T_{sil}$ ). The value of $T_z$ and $T_{sil}$ are given as:

$$T_z = \min(40, T_{sil}) \tag{3}$$

$$T_{sil} = \overline{sil} + 2 \cdot \sqrt{\frac{1}{Q}\sum_{i-1}^{Q}\left(Z(i) - \overline{sil}\right)^2} \tag{4}$$

$$\overline{sil} = \frac{1}{Q}\sum_{i=1}^{Q} Z(i) \tag{5}$$

where $Q$ is the number of silence frames. In this thesis, $Q$ is equal to 19.

In addition, the energy of a vowel signal is several orders of magnitude higher than that of a consonant signal. Therefore, a frame including vowels is determined from this relationship. The energy of the input speech is defined as:

$$E(j) = \sum_{i=1}^{2M} |s(i)| \tag{6}$$

where $M$ is 55 (5 ms), $x = (j-1)M$, and $s(i)$ is an observed input speech signal in each word at the $i$-th sample.

Figure 3.6 presents the example of the word "*thōo / rā / sa`p* ". Figure 3.6 (top) indicates the input signal, zero crossing (middle), and energy (bottom). Unvoiced signal regions provide a high zero crossing rate and low energy while voiced signal regions provide a low zero crossing rate and high energy. In figure 3.6, the unvoiced regions are shown in the rectangles.

In continuous speech, when we pronounce words successively, their energy waveform appears as the same word. This results in syllable segmentation errors. We therefore apply modified energy $\widetilde{E}(j)$ to this problem, since it allows the amplitude of the input signal to be observed more easily. We modify energy $\widetilde{E}(j)$ by dividing it by its maximum values shown below:

$$\widetilde{E}(j) = \frac{E(j)}{\max(E(j))}. \tag{7}$$

In each input frame, if the local minimum energy is lower than 0.38, which are 38% of the maximum of the modified amplitude, it is established as the end point and/or the start point candidate. If the difference of the fourth preceding and the successive frames are also more than 0.38, the local minimum point is determined to be the start/end point.

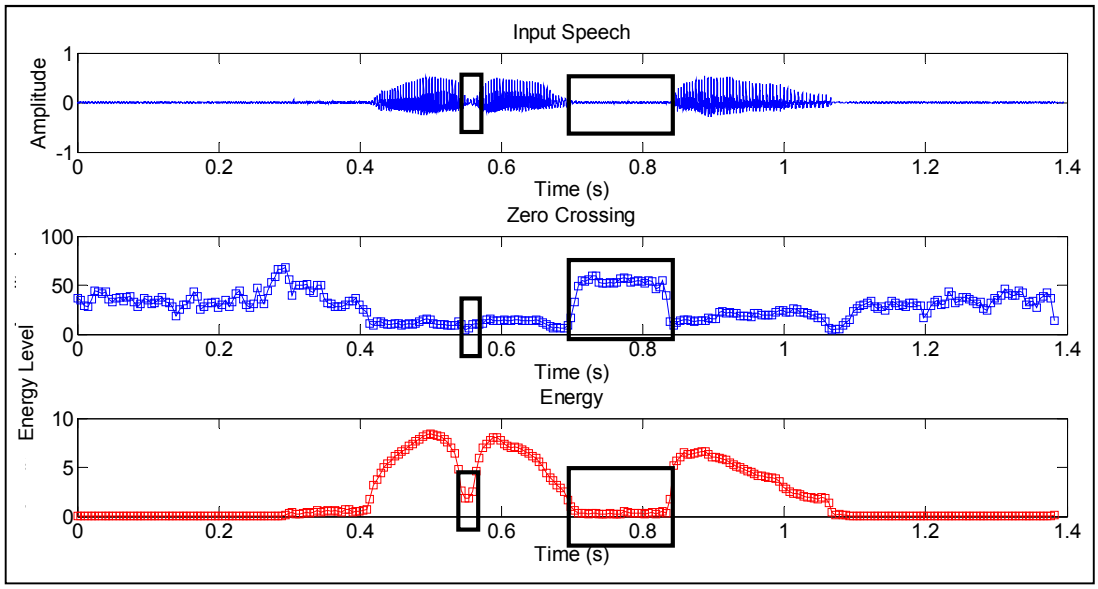



**Figure 3.6** Example of the word "*thōo / rā / sa`p*" with input speech (top), zero crossing (middle), speech energy (bottom)

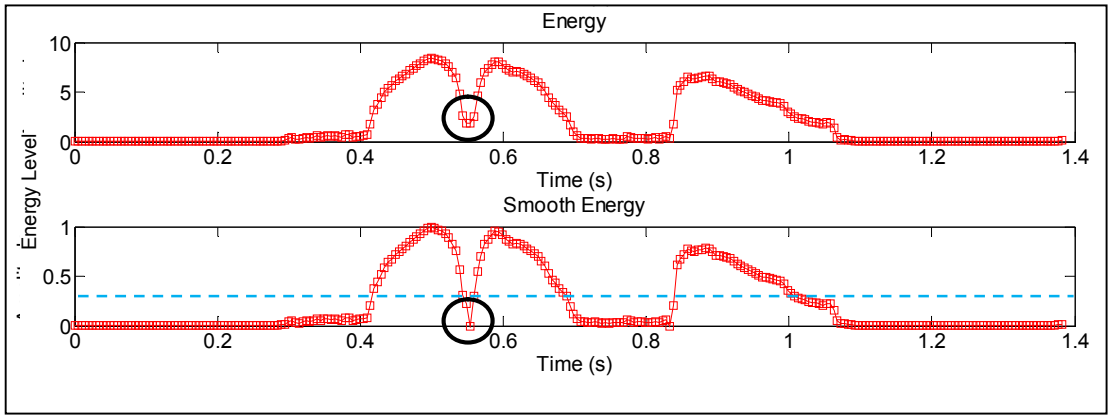



**Figure 3.7** Example of the word "*thōo / rā / sa`p*" syllable segmentation

Figure 3.7 illustrates an example of the word "*thōo / rā / sa`p*" syllable segmentation. The original energy waveform is shown in Figure 3.7 (top) while the modified energy is shown in the bottom waveform. The dashed line in the bottom waveform indicates the 0.38 value. The local minimum point and the start/end points are shown in the circles of Figure 3.7 (top) and (bottom), respectively.

After we determine start and end points, we segment the vowel signals. Since the energy of a vowel signal is several orders of magnitude higher than that of a consonant signal, we segment the vowel signals from the frame which has an energy value that is greater than or equal to a vowel threshold. The vowel segmentation provides high accuracy when we find the vowel threshold from the average energy of the entire syllable.

However, the average energy threshold calculation starts after the end of the utterance. This results in a processing time delay response. We therefore find the vowel threshold by adapting the average energy value. We assume that the maximum energy value

occurs in the middle frame and the threshold is half of the energy value of this frame. Since vowel signals can be estimated by time-synchronous calculation, the tone model can be processed in real-time processing. Due to the proposed vowel segmentation, the tone model provides greater precision compared with using the fixed threshold value [64] and also can be appropriately applied to a wide variety of input speech.

In this thesis, we assume that the maximum energy value in each frame occurs in the middle frame ( $j_v$ ). $j_v$ is set to the frame at $\left\lceil \frac{k}{2} \right\rceil$ where $k$ is the frame ending and $\lceil \ \rceil$ is the ceiling. The threshold of the vowel energy $T_e(j)$ [65] is given as:

$$T_e(j) = \frac{1}{2} E(j_v). \tag{8}$$

Figure 3.8 shows vowel segmentation of the word "*lûak*". Figure 3.8 (top) indicates the energy of input signal before the proposed $T_e(j)$ is applied, which is the energy of the entire syllable. The dashed line gives the threshold energy $T_e(j)$ which is 0.47 in this word. The frame with a value greater than $T_e(j)$ is determined as the vowel signal. Figure 3.8 (bottom) indicates the energy after we carry out $T_e(j)$. This results in a reduction in the number of frames.

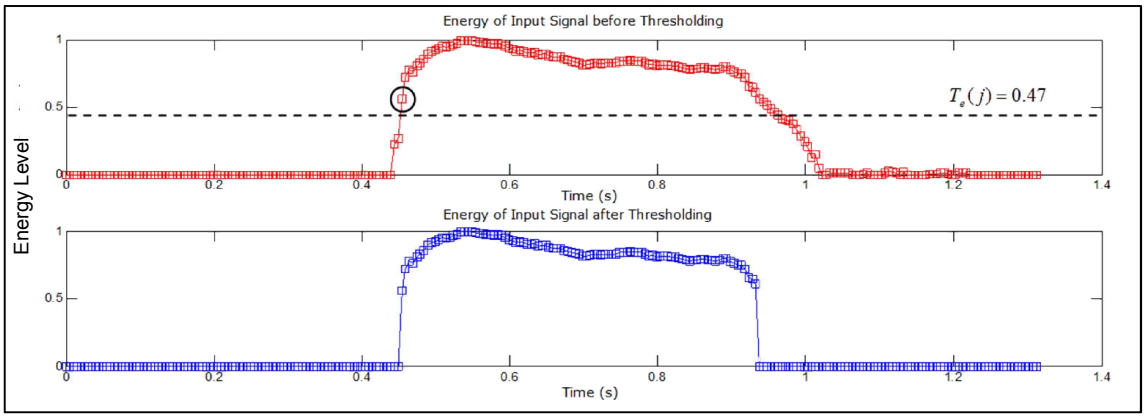



**Figure 3.8** Vowel segmentation of the word "*lûak*" before applying vowel threshold (top) after applying vowel threshold (bottom)

A magnitude difference function (MDF) pitch extraction method is based on the calculation of speech waveform. We therefore estimate fundamental frequency ($F_0$) by using only the vowel parts, called vowel-MDF ($V_{MDF}$). We design this process by simple algorithms and thus its calculation costs are low. $V_{MDF}$ is applied to the selected frame $j$ in the previous section. Each $j$ frame calculates $V_{MDF}$ within a frame length $N$ = 256 and frame shift $M$ = 128. The following Equation is applied to the selected frame:

$$V_{MDF}(j,n) = \sum_{i=1}^{2M} \left| s(x+i) - s(x+i+n) \right| \tag{9}$$

where $s(i)$ is an input vowel signal at the $i$-th sample, $x = (j-1)M$, and $n$ is a lag value $n$ = [1, 2,..., $N$].

In order to eliminate the effect of the $V_{MDF}$ complex trajectory, the following threshold is applied first to $V_{MDF}(j, l)$ where $l$ is the lag that covers possible pitch frequency range of men and women, between 50 and 500 Hz. In this thesis, we assign $L = [l_{20}, l_2, \ldots, l_{160}]$ and $W$ is equal to 141, total number of those lags.

$$\overline{P(j)} = \frac{1}{W}\sum_{l=20}^{160}(V_{AMDF}(j,l)) \tag{10}$$

$$T_p(j) = \overline{P(j)} - \sqrt{\frac{1}{p}\sum_{l=1}^{p}(V_{AMDF}(j,l) - \overline{P(j)})^2}. \tag{11}$$

In Equation (11), The value of $T_p(j)$ is defined as a pitch period threshold. The first term of $T_p(j)$ is the average value of $V_{MDF}(j,l)$ in the selected $j$-th frame. The second term of $T_p(j)$ indicates the variance of $V_{MDF}(j,l)$. Therefore, the threshold is given as the averaged lower bound of $V_{MDF}(j,l)$ variations.

The integer-valued pitch period is selected from candidate local minimum points $l_i$ where $V_{MDF}(j,l) < T_p(j)$, $i = [1, 2, \ldots, n]$. If the estimated candidates $l_i$ show the positions of all local minimum, $l_i$ is decided as the integer-valued pitch period of $V_{MDF}$. The $V_{MDF}$ pitch period ($P_v$) is satisfied as follows:

$$P_v(j) = l_1(j) \qquad ; \qquad l_1 > 0. \tag{12}$$

In order to evaluate our proposed method performance, we compare the proposed method to average magnitude difference function (*AMDF*) and autocorrelation function (*AC*) methods. The *AMDF* is similar to $V_{MDF}$. It divides $V_{MDF}$ by the range of lag value which is equal to $N = 256$. It estimates the pitch by selecting the minimum point of *AMDF (j,n)* as similar way as $V_{MDF}$. The *AC* is convolved with a time-lagged version of itself. The *AMDF* is defined as:

$$AMDF(j,n) = \frac{1}{N}\cdot\sum_{i=1}^{2M}\left|s(x+i) - s(x+i+n)\right| \tag{13}$$

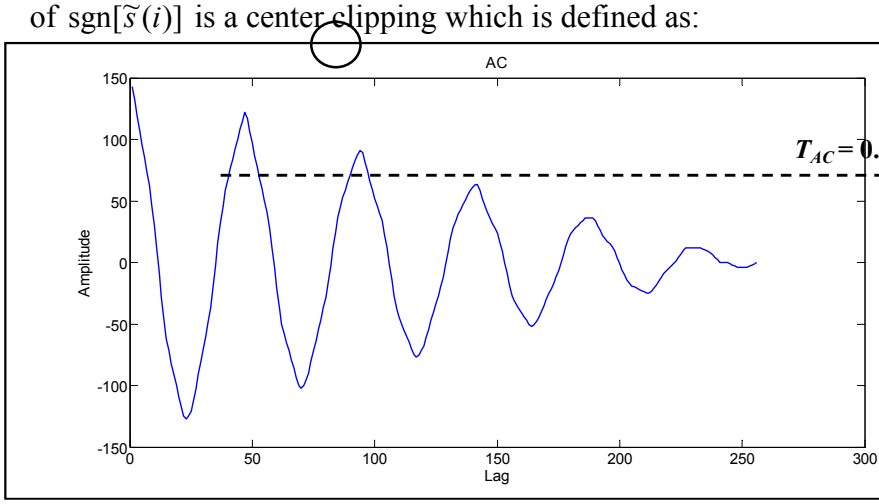
where $s(i)$ is an input vowel signal at the $i$-th sample, $x = (j-1)M$, and $n$ is a lag value $n = [1, 2, \ldots, N]$.

*AC* is given as:

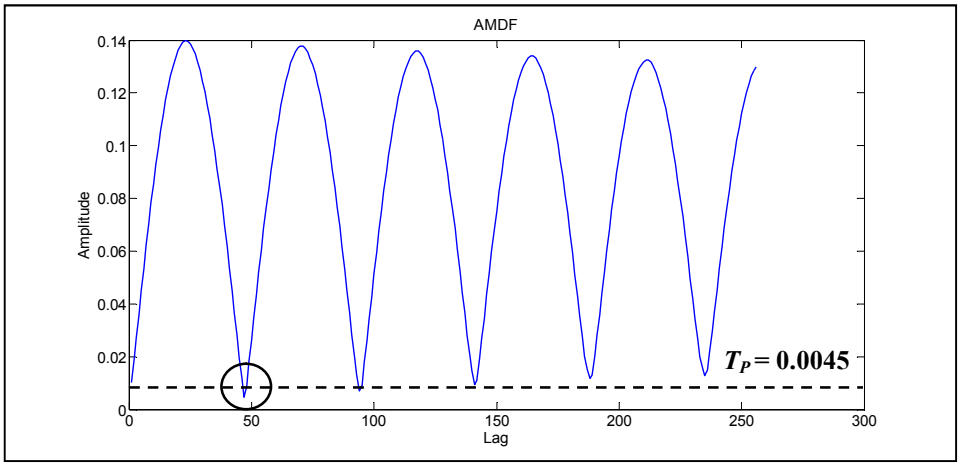
$$AC(j,n) = \sum_{i=1}^{2M}(\mathrm{sgn}[\tilde{s}(i)]\cdot\mathrm{sgn}[\tilde{s}(i+n)]). \tag{14}$$

$$\text{sgn}[\tilde{s}(i)] = \begin{cases} 1, s(i) > T_c \\ -1, s(i) < -T_c \\ 0, otherwise. \end{cases}$$

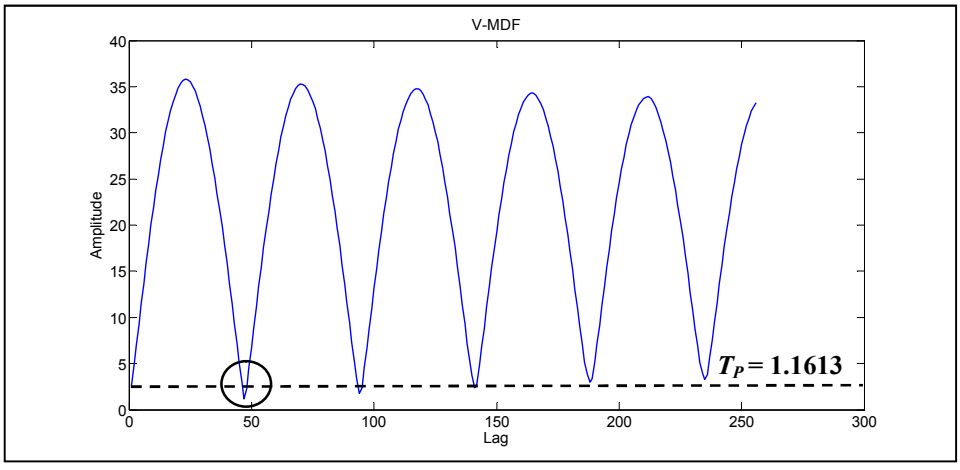
where $n$ is a lag value $n$ = [1, 2,..., $N$-$i$] and the function of sgn$[\tilde{s}(i)]$ is a center clipping which is defined as:




(a)




(b)




(c)

**Figure 3.9** $F_0$ feature waveforms of the word "*l~~a~~ak*" where $j$ = 31 and $P_v$ = 48 (a) *AC* (b) *AMDF* (c) $V_{MDF}$

We assume that the silence and unvoiced regions are equal to 30% of the total of the center clippings. We therefore estimate the pitch candidates from the maximum point of *AC(j,n)* where its value is greater than *AC* threshold ($T_{AC}$). The *AC* threshold ($T_{AC}$) is given as:

$$T_{AC} = 0.3 \cdot \sum_{i=1}^{N} \left| \tilde{s}(i) \right| \tag{15}$$

The potential pitch frequency range is generally found between 50 and 250 Hz for men while the range usually falls between 120 and 500 Hz for women [34]. We thus assign extract $V_{MDF}$, *AMDF*, and *AC* between the lag 20 and 160 with the sampling frequency 11.025 kHz.

Figure 3.9 shows the $F_0$ feature waveforms of the word "*lûak*" in frame 31 (a) autocorrelation function (*AC*) (b) average magnitude difference function (*AMDF*) (c) vowel- magnitude difference function ($V_{MDF}$). The pitch threshold of each function is presented in the dashed lines while the pitch points are in the circles. Although those functions indicate different waveforms and different thresholds, the results of the pitch points are the same.

In Figure 3.9, *AMDF* provides the lower amplitude, compared with $V_{MDF}$. In addition, it has to divide the summation of absolute difference by frame length. This results in a higher calculation cost than $V_{MDF}$. We therefore propose $V_{MDF}$ for our tone recognition.

## 3.3 $F_0$ Estimation

Pitch period is in fact a real number since the original speech is continuous-time signal. In order to find the optimal pitch period, the integer-valued pitch period $l_1$ can be optimized by an optimal pitch period coefficient $l_0$. Because the signal statistics slightly change with time [34], this thesis estimates $l_0$ by the association between the current samples and past samples ($P_1$), and between the current samples and consecutive samples ($P_2$) where $n$ is set from 1 to $N$ = 256 and a frame shift $M$ = 128. The association between the current, past, and consecutive samples can be given by

$$P_1(j) = s(l_1 + n + m) - s(l_1 + n) \tag{16}$$

$$P_2(j) = s(l_1 + n + m) - s(l_1 + n + m + 1) \tag{17}$$

where $m = (j - 1)M$, and $P1$ is the pitch period.

Figure 3.10 shows the input signals-association between $l_1(j)$ and $e(j)$ where a frame $j$ = 6. In Figure 3.10, $l_1$ is the current sample. *frameA*, *frameB*, and *frameC* are the current, past, and consecutive frames, respectively. $P_1(j)$ = *frameA* - *frameB*, and $P_2(j)$ = *frameA* - *frameC*. Each *frame(i)* is operated within a frame length $N$.

We assume that when we associate $P_1$ with $P_2$, $P_1$ is the sum between the product of coefficient $l_0$ with $P_2$ and the error $e(j)$.

$$P_1(j) = l_0 \cdot P_2(j) + e(j) \tag{18}$$

In order to find the optimal pitch period, we assign $N_0(j)$ as the normalized sum of squared error. To minimize $N_0(j)$, the optimal pitch period coefficient $l_0(j)$ can be found by differentiating $e(j)$ with respect to $l_0(j)$ and the result can then be set to zero. The $N_0(j)$ and $l_0(j)$ are defined by

$$N_0(j) = \frac{\sum_{n=1}^{N}(P_1(j) - l_0 \cdot P_2(j))^2}{\sum_{n=1}^{N} P_1^2(j)} \tag{19}$$

$$l_0(j) = \frac{\sum_{n=1}^{N} P_1(j) \cdot P_2(j)}{\sum_{n=1}^{N} P_2^2(j)} \tag{20}$$

where frame length $N = 256$.

Finally, we determine the fundamental frequency $F_0(j)$ by the optimized pitch period. Namely, $l_0$ and $l_1$ are the denominators used to indicate $F_0$. The $F_0(j)$ is defined by

$$F_0(j) = \frac{F_s}{P_v(j) + l_0(j)} \tag{21}$$

where $F_s$ is the sampling rate 11.025 kHz.

The results of the word "*hnâa*" in feature extraction and $F_0$ estimation where $j = 6$ are shown in Figure 3.11. In the top waveform, the vowel signals detected from the feature extraction process in Section 2.1 are shown. The middle waveform shows the trajectory of the proposed $V_{MDF}(j)$. The $T_p(j)$ is shown in the dashed line and the local minimum points, having values lower than $T_p(j)$, are presented in the circles. Although the positions of 44 (circle 1) and 89 (circle 2) are less than $T_p(j)$, $P_v(j)$ is given as the position of $l_1(j) = 44$. Finally, Figure 3.11 (bottom) shows the $F_0(j)$ curve of falling tone. The $F_0(j)$ value is defined as $F_s$ divided by $l_1(j) + l_0(j)$.

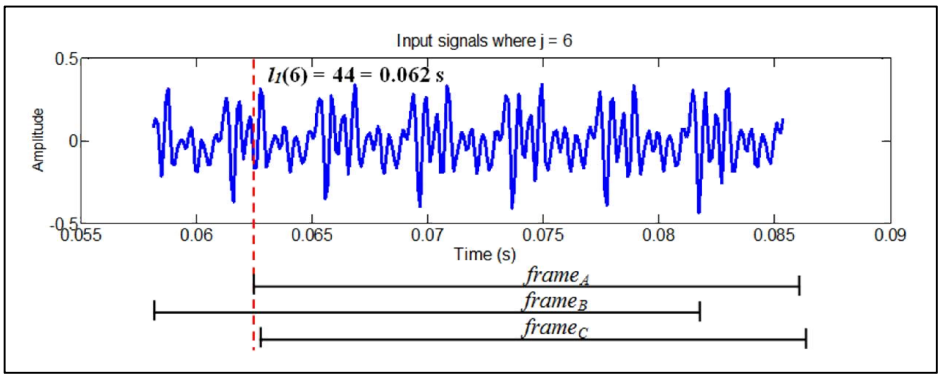



**Figure 3.10** Association between the current, past, and consecutive samples where a frame $j = 6$ and $l_1(6) = 44$

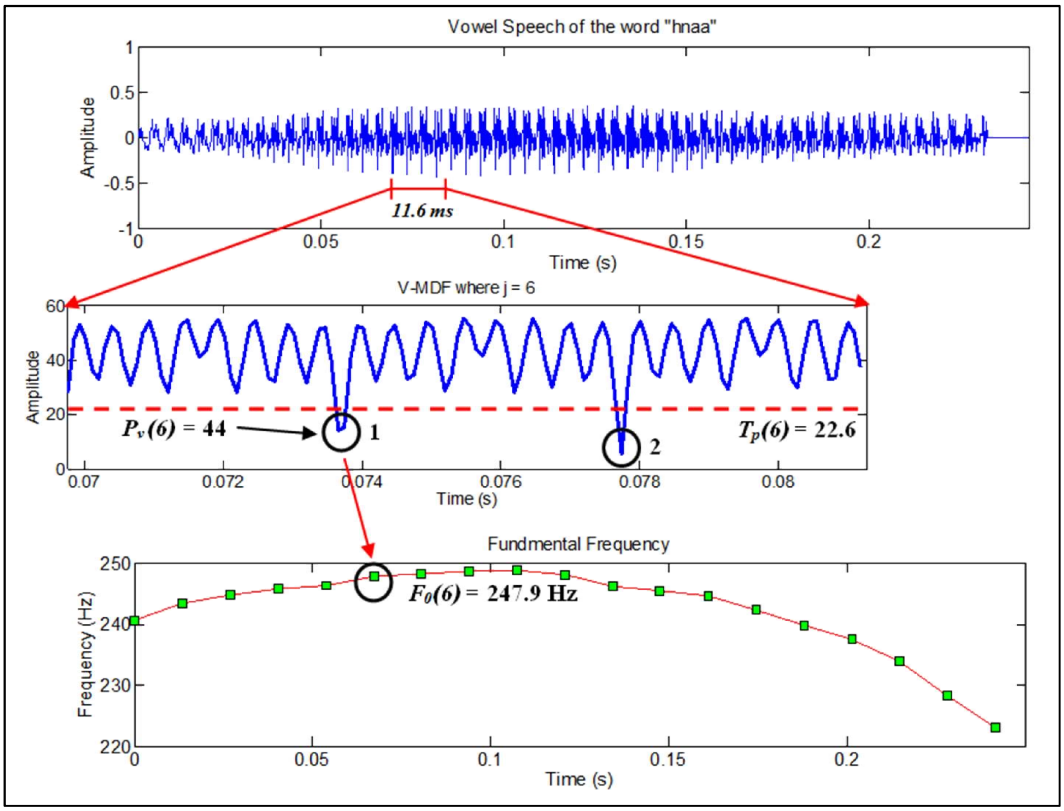



**Figure 3.11** Example of the word "*hnâa*"; vowel signal (top), $V_{MDF}$ based pitch detector (middle), and $F_0$ characteristic of falling tone (bottom)

## 3.4 Tone Decision

The proposed method classifies tones from the trajectory of $F_0(j)$ curve where $j$ = 1, 2, …, $k$ and $k$ is the frame ending. The typical $F_0$ trajectories of five Thai tones are shown in Appendix A.

There are certain features used in the decision tree. Three condition steps are considered to classify five Thai tones.

Firstly, we determine the maximum $F_0$ of each individual word ($M_w$) as shown in Equation (22), then the maximum $F_0$ of the first half syllable ($M_{h1}$) and the second half syllable ($M_{h2}$) are determined in Equation(23) to Equation (24). If $M_w$ is equal to $M_{h1}$, either a low or falling tone is a tone result. If $M_w$ is equal to $M_{h2}$, either a rising or high tone is a tone result. Due to a mid tone characteristic, the $M$w of the mid tone can be equal to $M_{h1}$ or $M_{h2}$.

Secondly, slope ($S$) in Equation (25) is used to observe the fall and rise of the $F_0$. Absolute slope ($|S|$) in Equation (26) observes vertical and horizontal variation. If $|S| \leq$ 3.0, only the mid tone is decided as a tone result.

Finally, $F_0$ is divided into a first half and a second half syllable, then we compare the value of $F_0$ ($j$) to observe the initial and last parts of the both half syllables. The initial part of the first half syllable compares the values of $F_0$ ($j$) at $j$ = 1 and 2, $F_0$ (1) and $F_0$ (2). The last part of the first half syllable compares the values of $F_0$ ($rd(k/2) - 1$) and $F_0$ ($rd(k/2)$) where $rd(x)$ shows the rounded off value of $x$. The second half syllable

compares $F_0$ ($j$) in the same way as the first half syllable. The initial part of the second half syllable compares the values of

$F_0$ ($rd(k/2)$+1) and $F_0$ ($rd(k/2)$+2) . The last part of the first half syllable compares the values of $F_0$ ($k$ − 1) and $F_0$ ($k$). The tone results are decided by a decision tree algorithm shown in Figure 3.12. The following features are used in the tone decision process:

$$M_w = \max[F_0(j) : j = 1,2,...k] \tag{22}$$

$$M_{h1} = \max[F_0(j) : j = 1,2,...rd(k/2)] \tag{23}$$

$$M_{h2} = \max[F_0(j) : j = rd(k/2)+1, rd(W/2)+2,...,k] \tag{24}$$

$$S = \sum_{j=2}^{k} (F_0(j) - F_0(j-1)) \tag{25}$$

$$|S| = \sum_{j=2}^{k} |F_0(j) - F_0(j-1)| \tag{26}$$

where $max[x_i : i = 1, 2, \ldots, k]$ means that the maximum value of $xi$ is selected among $i = 1, 2, \ldots, k$.

Figure 3.12 describes the 5 Thai tones decision tree in the decision process. If the maximum value, $M_w$, is located in the first half trajectory and its slope is going down, a low or falling tone should occur.

In addition, if $F_0(1) > F_0(2)$, a low tone is determined. Otherwise, if the central part of the trajectory is rising from the start point and then going down to the end point, the result is a falling tone.

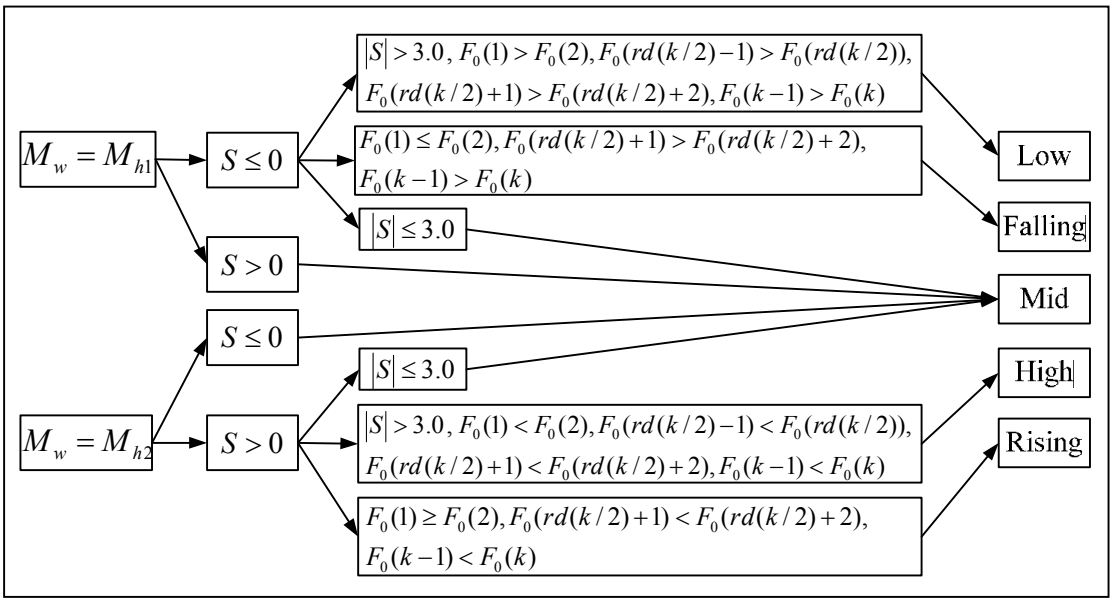



**Figure 3.12** Decision tree of 5 Thai tones [64]

If $|S|$ is small or $S$ is greater than 0, the tone result is a mid tone. If the maximum value, $M_w$, is located in the second half trajectory and its slope is rising, a high or rising tone should occur. If $F_0(1) < F_0(2)$, a high tone is decided. Otherwise, if the central part of the trajectory is going down from the start point and then rising to the end point, a rising tone is the tone result.

In addition, if $|S|$ is small or $S$ is less than or equal to 0, the tone result is a mid tone.

Finally, we compare $F_0$ trajectories between *AC*, *AMDF*, and $V_{MDF}$ in Figure 3.13. These functions result in a mid tone even though their $F_0$ trajectories are slightly different. While the $F_0$ trajectories of *AMDF* and $V_{MDF}$ are equal, *AMDF* requires more operations than in $V_{MDF}$.

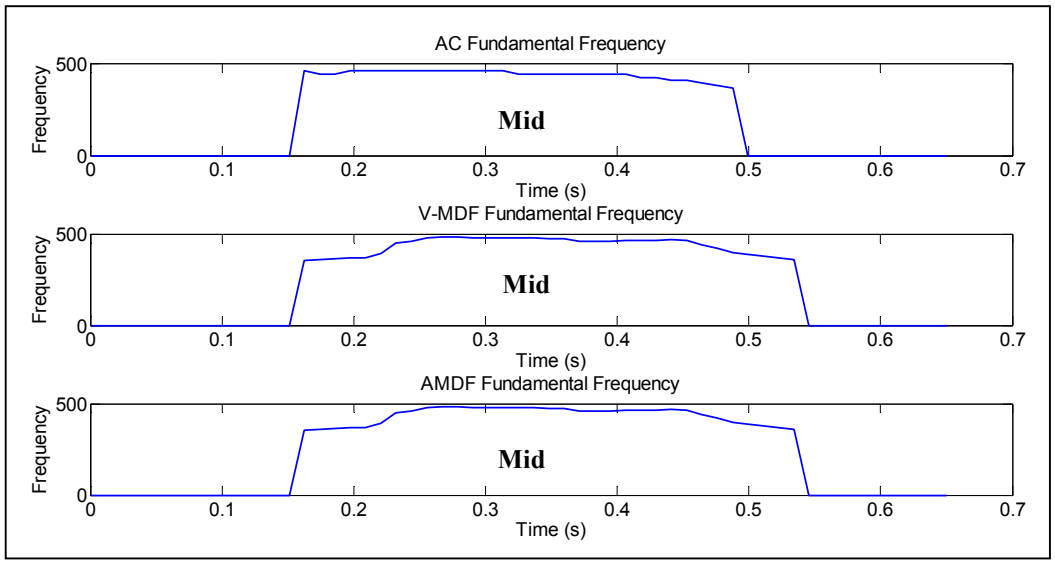



**Figure 3.13** $F_0$ trajectories of the word "*lûak* "; *AC* (top), $V_{MDF}$ (middle), and *AMDF* (bottom)