PITCH DETECTION ALGORITHM: AUTOCORRELATION METHOD AND AMDF

Li Tan and Montri Karnjanadecha

Department of Computer Engineering
Faculty of Engineering
Prince of Songkhla University
Hat Yai, Songkhla
Thailand, 90112
E_mail: litan212@hotmail.com, montri@coe.psu.ac.th

ABSTRACT
This paper describes the pitch tracking techniques using autocorrelation method and AMDF (Average Magnitude Difference Function) method involving the preprocessing and the extraction of pitch pattern. It also presents the implementation and the basic experiments and discussions.

KEYWORDS
Pitch, Pitch Detection Algorithm, Autocorrelation function, Speech Recognition System, Center-clipping, Pitch Contour

1. INTRODUCTION

Pitch detection is very important for many speech processing algorithm. Speech recognition system of tonal language use pitch tracking for tone recognition, which is important in disambiguating the myriad of homophones. Pitch is also crucial for prosodic variations in text-to-speech systems and spoken language systems. The fundamental frequency (*F0*) is the main cue of the pitch. However, it is difficult to build a reliable statistical models involving fundamental frequency *F0* because of pitch estimation errors and the discontinuity of the F0 space. Thus, a reliable pitch detection algorithm (PDA) is a very important component in many speech processing systems.

In the paper, the principles of the two pitch detection algorithms, preprocessing and the extraction of pitch pattern techniques are introduced. The implementation of them is described. Then the experiments and discussions are presented. Finally it's the conclusions of the paper.

2. BACKGROUND

2.1 Autocorrelation Method and AMDF

Basically, pitch detection algorithms use short-term analysis techniques. For every frame x_m we get a score $f(T|\boldsymbol{x_m})$ that is a function of the candidate pitch periods T. Algorithm determine the optimal pitch by maximizing (1).

$$T_m = \underset{T}{\operatorname{argmax}} f(T \mid x_m) \qquad (1)$$

A commonly used method to estimate pitch is based on detecting the highest value of the autocorrelation function in the region of interest. Given a discrete time signal $x(n)$, defined for all n , the auto-correlation function is generally defined in (2):

$$R_x(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m) \qquad (2)$$

The autocorrelation function of a signal is basically a (non-invertible) transformation of the signal that is useful for displaying structure in the waveform. Thus, for pitch detection, if we assume $x(n)$ is exactly periodic with period P, i.e., $x(n) = x(n + P)$ for all n, then it is easily shown that:

$$R_x(m) = R_x(m + P), \qquad (3)$$

i.e., the autocorrelation is also periodic with the same period. Conversely, periodicity in the autocorrelation function indicates periodicity in the signal.

For a nonstationary signal, such as speech, the concept of a long-time autocorrelation measurement as given by (2) is not really meaningful. Thus, it is reasonable to define a short-time autocorrelation function, which operates on short segments of the signal as:

$$R_x(m) = \frac{1}{N} \sum_{n=0}^{N'-1} [x(n+l)w(n)][x(n+l+m)w(n+m)], \quad o \le m \le M_0 \qquad (4)$$

where $w(n)$ is an appropriate window for analysis, N is the section length being analyzed, N' is the number of signal samples used in the computation of $R(m)$, Mo is the number of autocorrelation points to be computed, and l is the index of the starting sample of the frame. For pitch detection applications N' is generally set to the value in (5):

$$N' = N - m \qquad (5).$$

So that only the N samples in the analysis frame (i.e., $x(l)$, $x(l+1), \ldots, x(l + N - 1)$) are used in the autocorrelation computation. Values of 200 and 300 have generally been used for Mo and N , respectively, it is corresponding to a maximum pitch period of 20 ms (200 samples at a 10 kHz sampling rate) and a 30 ms analysis frame size. [1,3]

A variation of autocorrelation analysis for measuring the periodicity of voiced speech uses the AMDF, defined by the relation in (6):

$$D_m = \frac{1}{L} \sum_{n=1}^{L} |x(n) - x(n-m)|, \quad m = 0,1,..m_{\max} \qquad (6)$$

Where $x(n)$ are the samples of input speech and $x(n-m)$, are the samples time shifted m seconds. The vertical bars denote taking the magnitude of the difference $x(n) – x(n-m)$ Thus a difference signal Dm, is formed by delaying the input speech various amounts, subtracting the delayed waveform from the original, and summing the magnitude of the differences between sample values. The difference signal is always zero at delay = 0, and is particularly small at delays corresponding to the pitch period of a voiced sound having a quasiperiodic structure.

The AMDF [4] is a variation of ACF (Autocorrelation Function) analysis [1] where, instead of correlating the input speech at various delays (where multiplications and

summations are formed at each value), a difference signal is formed between the delayed speech and the original, and at each delay value the absolute magnitude is taken. Unlike the autocorrelation or cross-correlation function, however, the AMDF calculations require no multiplications, a desirable property for real-time applications.

For each value of delay, computation is made over an integrating window of *N* samples. To generate the entire range of delays, the window is "cross differenced" with the full analysis interval. An advantage of this method is that the relative sizes of the nulls tend to remain constant as a function of delay. This is because there is always full overlap of data between the two segments being cross differenced.

In extractors of this type, the limiting factor on accuracy is the inability to completely separate the fine structure from the effects of the spectral envelope. For this reason, decision logic and prior knowledge of voicing are used along with the function itself to help make the pitch decision more reliable. [1,4]

2.2 Preprocessing Technique

From above, we know the autocorrelation function and AMDF can be used to detect the pitch. However the speech signal include very rich harmonic components. The minimum F0 is about 80 Hz and the maximum is about 500 Hz. Most of them are in the range of 100-200 Hz. Thus the signal may involve 30-40 harmonic components. And the F0 component is often not the strongest one. Because the first formant usually is between 300-1000 Hz. That is, the 2-8 harmonic components usually stronger than fundamental component. The rich harmonic components let the pitch tracking become very complex. It usually has the harmonic errors and sub-harmonic errors. To improve the reliability some pre-processing of signal is necessary.

Since, the range of F0 is generally in the range of 80-500 Hz, then the frequency components above 500 Hz is useless for pitch detection. Thus a low-pass filter with pass-band frequency above 500 Hz would be useful in improving the performance of pitch detection. Generally, we use the low-pass-filter with 900 Hz.

Also to reduce the effects of the formant structure on the detailed shape of the short-time autocorrelation function, the nonlinear processing is usually used in pitch tracking.

$$Y(n)=C[x(n)] \qquad (7)$$

One of the nonlinear technique is center-clipping of speech which is first introduced by *M. M. Sondhi* [3]. The relation between input *x(n)* and *y(n)* is:

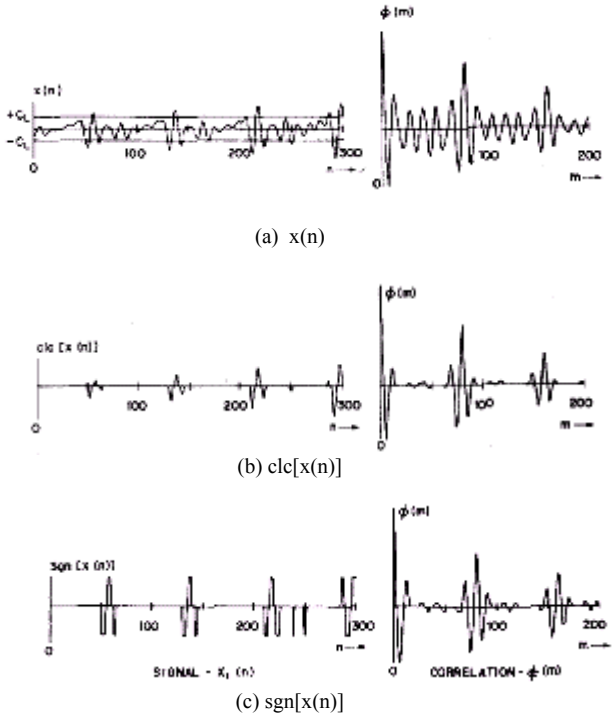
$$y(n) = clc[x(n)] = \begin{cases} (x(n) - C_L), & x(n) \geq C_L \\ 0, & |\, x(n) \,| < C_L \\ (x(n) + C_L), & x(n) \leq -C_L \end{cases} \qquad (8)$$

Another nonlinear clipping we call is infinite-peak-clipping. The function is described in (9):

$$y(n) = \mathrm{sgn}[\, x(n)] = \begin{cases} 1, & x(n) \geq C_L \\ 0, & |\, x(n) \,| < C_L \\ -1, & x(n) \leq -C_L \end{cases} \qquad (9)$$

where C_L is the clipping threshold. Generally C_L is about 30% of the maximum magnitude of signal. In application the C_L should be as high as possible. To get the high C_L, we can catch the peak value of the first 1/3 and the last 1/3 of signal and use the less one to be the maximum magnitude. Then we set the 60-80% of this maximum magnitude to be C_L.

The effect of center-clipping and infinit-peak-clipping is clearly shown in the Fig. 1 (a, b, c). From Fig. 1 (b), after center-clipping, the autocorrelation only leave several pulse that show the reduction of the confused secondory peak. From Fig. 1 (c), the first peak is very clear. Also the secondary peak



(a) x(n)

(b) clc[x(n)]

(c) sgn[x(n)]

Fig. 1 the autocorrelation of x(n), clc[x(n)], sgn[x(n)] (adapted from [5])

value is reduced. All of these show that the center-clipping and infinite-clipping is effective in reduce the effects of the formant structure [2,3,5].

2.3 Post-processing

Generally, the pitch determination described above is still error-prone. The erroneous voiced/unvoiced decisions and inaccurate voiced pitch hypotheses can lead to noisy and undependable feature measurements. Then a smoothing stage is necessary in improving the performance of the system. The most common smoothing techniques includes: median filter, linear smoothing and dynamic programming technique. According to the reliability of pitch tracking algorithm, generally the median-filter is used. In the method of median-filter, it uses a moving window with the length *L*. The value at point n is determined by the data from point *n-L* to point *n+L*. Then the median value in these *2L+1* points is chooses as the value the point. [3]

2.4 Feature Extraction

After getting the smoothed pitch contour, we fit it into the 3rd order polynomial using least-mean-square or project it onto some basis functions using Orthogonal polynomial approximation.

For least-mean-square approximation, it express the approximation function as the sum of weighted observation value.

$$f_{LMS} = \sum_{i=1}^{N} a_i x_i \qquad (10)$$

Where f_{LMS} is the estimated function, a_j is the weighted coefficients, x_I is the observation *item (x1=1,x2=x,x3=x^2, x4=x^3)*. Then it is to minimize the expectation value of approximation error *(e=f_{LMS}-f)* to get the weighted coefficients a_i. In order to minimize the expectation value, we need to get the derivation of the expectation value and set it to zero. Then the coefficients will be calculated through equation (11).

$$\sum_{j=1}^{N} E(x_i x_j) a_j = E(f x_i) \qquad (11)$$

Orthogonal polynomials are defined in terms of their behavior with respect to each other and throughout some predetermined range of the independent variable. In the case of the vectors, if the set was completed it was said to span a vector space and any vector in that space could be expressed as a linear combination of orthogonal basis vectors. The first four discrete Legendre polynomials can be chosen to represent the pitch contour. They are shown in equation (12):

$$\Phi_0\left(\frac{i}{N}\right) = 1$$

$$\Phi_1\left(\frac{i}{N}\right) = \left[\frac{12\times N}{N+2}\right]^{1/2}\left[\left(\frac{i}{N}\right) - \frac{1}{2}\right]$$

$$\Phi_2\left(\frac{i}{N}\right) = \left[\frac{180N^3}{(N-1)(N+2)(N+3)}\right]^{1/2}\left[\left(\frac{i}{N}\right)^2 - \left(\frac{i}{N}\right) + \frac{N-1}{6\times N}\right]$$

$$\Phi_3\left(\frac{i}{N}\right) = \left[\frac{2800N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right]^{1/2} \bullet \left[\left(\frac{i}{N}\right)^3 - \frac{3}{2}\left(\frac{i}{N}\right)^2 + \frac{6N^3-3N+2}{10\times N^2}\left(\frac{i}{N}\right) - \frac{(N-1)(N-2)}{20\times N^2}\right] \qquad (12)$$

These polynomials are normalized in length to [0,1]. Where i is from 0 to *N*, *N+1* is the length of pitch contour and *N* should be bigger than 3. Legendre polynomials is a kind of Orhtogonal polynomials with the simplest weight function which is equal to 1. They are chosen to represent the pitch contour because they resemble to the basic pitch contour patterns. A pitch contour segment *f(i/N),* can then be

$$\hat{f}\left(\frac{i}{N}\right) = \sum_{j=0}^{3} a_j \times \Phi_j\left(\frac{i}{N}\right), \quad 0 \le i \le N \qquad (13)$$

approximated as (13):
Where

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} f\left(\frac{i}{N}\right) \times \Phi_j\left(\frac{i}{N}\right)$$

The reconstructed pitch contour will not lose much information since orthogonol polynomials up to degree of three are used to fit it. [7,8]

3. IMPLEMENTATIONS

3.1 Modified Autocorrelation Method

According to the discussion above, the modified autocorrelation pitch detector based on the center-clipping method and infinite-clipping is used in our implementation. Fig. 2 shows a block diagram of the pitch detection algorithm. The method requires that the speech be low-passed filtered to 900 Hz. The low-pass filtered speech signal is digitized at a 10-kHz sampling rate and sectioned into overlapping 30-ms (300 samples) sections for processing. Since the pitch period computation for all pitch detectors is performed 100 times/s i.e, every 10 ms, adjacent sections overlap by 20 ms or 200 samples. The first stage of processing is the computation of a clipping threshold C_L for the current 30-ms section of speech. The clipping level is set at a value which is 68 percent of the smaller of the peak absolute sample values in the first and last 10-ms portions of the section. Following the determination of the clipping level, the 30-ms section of speech is center clipped, and then infinite peak clipped. Following clipping the autocorrelation function for the30-ms section is computed over a range of lags from 20 samples to 160 samples (i.e., 2-ms-20-ms period). Additionally, the autocorrelation at 0 delay is computed for voiced/unvoiced determination. The autocorrelation function is then searched for its maximum value. If the maximum exceeds 0.55 of the autocorrelation value at 0 delay, the section is classified as voiced and the location of the maximum is the pitch period. Otherwise, the section is classified as unvoiced.

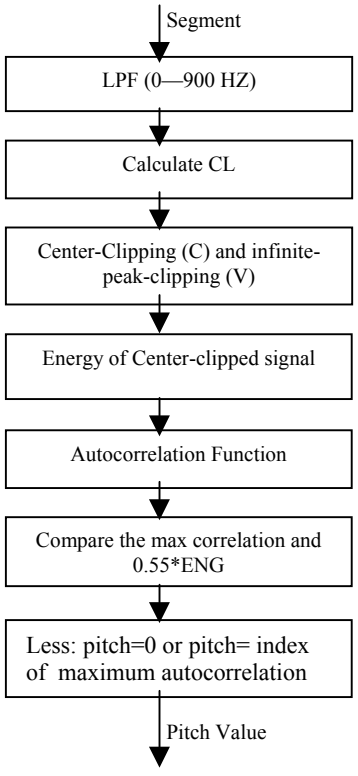


Fig 2. Block Diagram of Pitch Detection Algorithm using Modified Autocorrelation Method

3.2 AMDF

We only implement a coarse quantization. We leave the voice/unvoiced detection and the decision logic as the further work. Fig. 3 shows a block diagram of the AMDF pitch detector. The speech signal, is initially sampled at 10 kHz.

Then the signal pass a low-pass filter (0-900 Hz) and set the first 20 samples to be zero. The clipping threshold is then calculated and the center-clipping is done on the signal. Then average magnitude difference function is computed on the center-clipped speech signal at the lag (20—140 samples) through the signal from 20 to 160 samples. The pitch period is identified as the value of the lag which the minimum AMDF occurs. Thus a fairly coarse quantization is obtained for the pitch period.

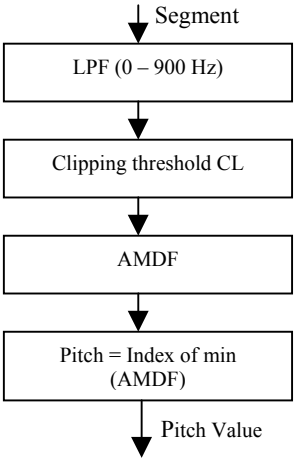


Fig. 3 Block Diagram of the Coarse Pitch Detection using AMDF

Also the 5-point median-filter and feature extraction using LMS and Orthogonal polynomials are implemented according to the introduction above.

4. EXPERIMENTS AND DISCUSSION

4.1 Experiments Setting

The experiments that we have done mainly include two parts. First part is emphasis on the observation of the results of these two pitch detection algorithms. And the pre-processing effects like the processing of low-pass-filter and center-clipping. The voiced/unvoiced determination in autocorrelation method is also tested. The speech that we used in our experiments is from Thai continuous digit database. Here for observing the effects, we have done the above experiments for some speeches from the database. Considering almost all of them shows the similar results. Here we only use one continuous speech with information "07229" and one single Madarine speech "hao(3)" which is considered more difficult in pitch tracking because of its big variation. Second part is worked on a small database which is based on 4-continuous-Thai-digit sentence. The sentences are chosen according to the general distribution of 3 tones in Thai digit. It includes 14 sentences with 23 1st tone, 10 2nd tone and 23 4th tone. To consider this is only for testing, we record the sound in the office environment. Sampling frequency is 16K Hz. And we collect 4 male's sound and 2 rounds per person. Finally we get 112 speeches. All of the speech is hand labeled with the wave-surfer software. In our testing, we use the 1st round speech of each person as training set. And the following are as the testing set. We use our implementation to detect the pitch contour and extract the pitch feature. Then we use a three-layer feedforward NN (neural network) with 4 inputs and 5 outputs as the framework. The hidden layer and training epocs is determined by the test.

4.2 Autocorreltion and AMDF on Continuous Speech

To observe the difference between AMDF and Autocorrelation method, we test both of them through a Thai continuous digit "07229", which is shown in Fig. 5. The pitch is shown in Fig. 7. From the figure, the pitch information mainly lies on the voiced part in the speech signal. In the silence part of the pitch is shown as the big variation. In the voiced part the pitch tracking show continuously and smoothly. Then the voiced/unvoiced decision is proved to be a very important part of pitch detection. Also although the pitch track shown in Fig. 6 can describe the trend of the pitch, it still exists some error points which need the further processing, that we say, smoothing. Also in Fig. 7, it shows both results for Autocorrelation method and AMDF. We can see that both methods can give us accepted result.

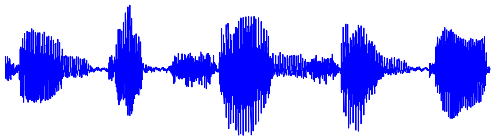


Fig 5. Waveform of Thai Digit("07229")

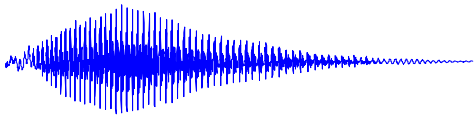


Fig. 6 Waveform of Mandarin Speech "hao" with 3rd tone

4.3 Voice/Unvoiced Decision

In the implementation of autocorrelation method, we use the 0.55 of the frame energy as the threshold to detect the voice/unvoice decision. Fig. 8 shown the experiment's results of it. From Fig. 8, it can detect the voiced part of the speech basically although some decision logics need to be further considered.

4.4 Smoothing

From the above figure, we know, the smoothing of the pitch contour is necessary after a single pitch contour got. Here we choose a single speech word to be the object. Generally the pitch of 3rd tone in Mandarin is more difficult than other tone because of its big variations. Here we choose the Mandarin word "hao" with 3rd tone in the experiment. The waveform is shown in Fig. 9. Also we use the median filter to smooth the pitch using Autocorrelation method. Fig 9 shows the effective of median filter. But the median-filter can't delete the several continuous error points. Also the difference between two continuous frames is examined. If it is greater than a predetermined threshold, the one lie farther away from the mean is treated as error and modified.

4.5 Effects of Pre-processing

Also in order to observe the effects of low-pass-filter and center-clipping, we did the experiments on the speech "hao (3)". The results is shown in Fig. 10 and Fig. 11. From Fig. 10 which is using AMDF algorithm, we don't find the big effect of the pre-processing here. We consider that it might the effects of formant structure in AMDF method is not big. But the pre-processing reduced the data and then increased the processing speed. But the effects of LPF in autocorrelation method are quite clear and shown in Fig. 11 In Fig. 11, the

error points reduced from 10 to 4 after adding the processing of LPF.

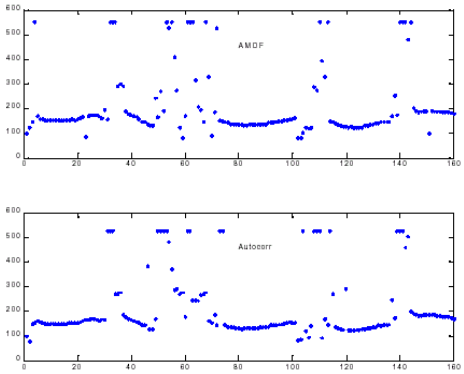


Fig. 7 Pitch Track Using Autocorrelation Method and AMDF (up: AMDF down: Autocorrelation)

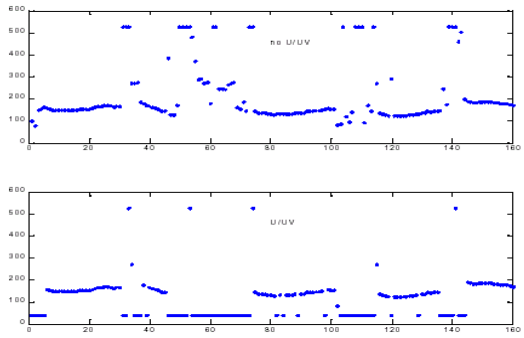


Fig. 8 the Voice/Unvoiced Detection in Autocorrelation method(up: no V/UV down: V/UV)

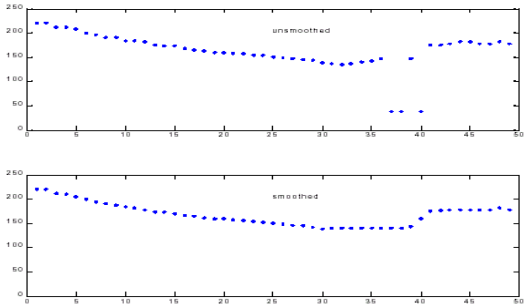


Fig. 9 Smoothing Pitch contour of "hao(3)" using median-filter

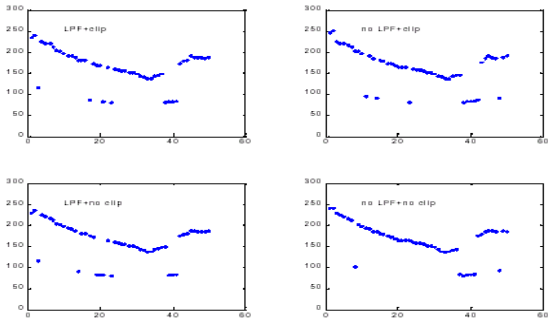


Fig. 10 Effects of Pre-processing technique using AMDF (up-left: LPF+clipped up-right: no LPF+clipped down-left: LPF+no clipped down-right: no LPF+no clipped)

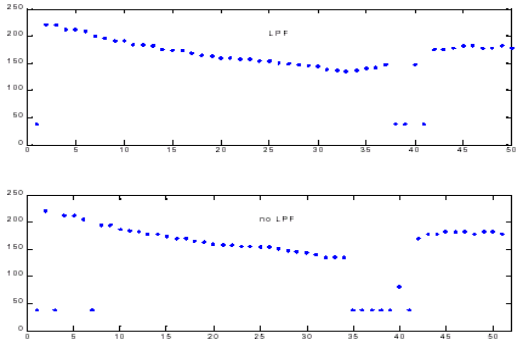


Fig. 11 Effects of LPF in autocorrelation method (up: LPF down: no LPF)

4.6 Feature Extraction

Pitch information mainly lies on the trend of pitch contour. As introduced above, two methods, LMS and Orthogonol polynomial, are used to extract the pattern of pitch contour. The experiment is shown in Fig. 12. According to the figure, both of them are working well. But finally which one can get better performance in recognition system needs the further research and experiments. Also Fig. 13 shows the shape of the four discrete Legendre bases for the space of pitch contour length. From Fig. 13, we can see that the four discrete Legendre bases are quite resemble to the basic pitch contour.

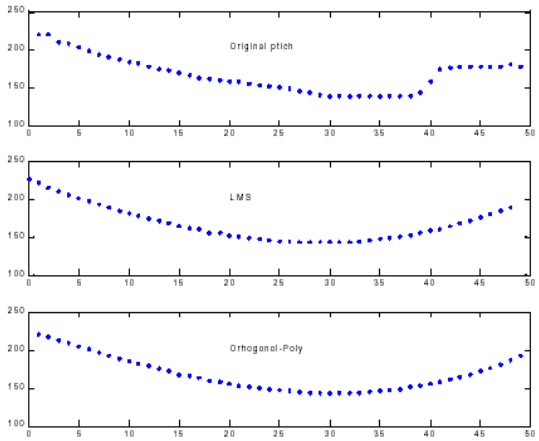


Fig. 12 The pitch pattern extracted 0(up: original-pitch mid: LMS bottom: Legendre Polynomials)

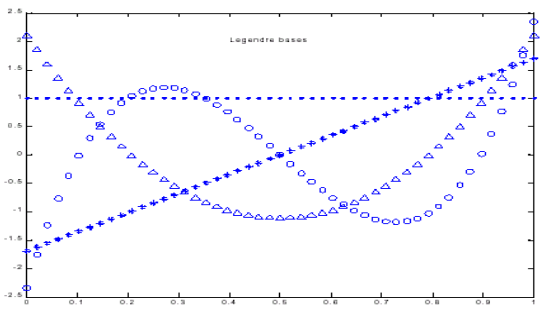


Fig. 13 Four Discrete Legendre bases (Point: base 1 *: base 2 Δ: base 3 o: base 4)

4.7 Classification

This is the second part of our experiments. Feature extraction using the implementation described above and classifier using NN are used. According to our observation, we use the autocorrelation pitch detection and orthogonal polynomial in out testing. All feature vectors are normalized to lie between –

$$normF_i = 2.0 \times \left(\frac{F_i - \min F_i}{\max F_i - \min F_i} \right) - 1.0 \qquad (14)$$

1.0 and 1.0 using the min-max normalization.

The total percentage of the testing is 79.02% (177 from 224). the confusion-matrix is shown in table 1.

Table 1. Confusion-matrix of Tone Classification for Thai Digit

Tone	1	2	4	Percent(%)
1	69	4	19	75
2	5	33	2	82.5
4	15	2	75	81.52
				79.02%

From table 1, we can see the lose of accuracy mainly lies in the confusion between 1st tone and 4th tone. The reason for this result may lie in the 5-Thai-tone contour which is shown in Fig. 15 and the effects of continuous speech.

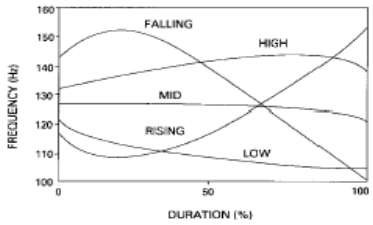


Fig. 14 Average F0 contours of the five Thai tones produced in isolation (adapted from [9])

From here, we can see the initial level of tone 1 and tone 4 is similar. Also because of the continuous effect of speech, the trend of tone can't meet the final level for tone 4 and it let the tone 1 end in a higher level than the isolated case. Also here only 4 feature is used in classification. So it's possible the accuracy will be improved if more feature is added.

5. CONCLUSIONS

The work that we described here is the two pitch detection algorithms and the related techniques including preprocessing post-processing and extraction of pitch pattern. According to our observing of the experiments. We found that both autocorrelation method and AMDF algorithm can provide the accepted results. Through the observing of preprocessing technique in both techniques, we didn't find the big effects of preprocessing on AMDF. But the obvious effects of low-pass-filter is shown in the experiment using autocorrelation method. At the same time, we have tested the smoothing using median-filter and voiced/unvoiced decision in autocorrelation method. Both of them showed the positive results. Finally, we used two methods to extract the pitch pattern through the smoothed pitch contour. According to the experiments figure, both of them works quite well. But in this case we need the quite smoothed pitch segment. For pitch detection the voice/unvoiced determination and the segmenting of pitch contour are another important issue that we didn't discussed here. We will put them in our further work. Moreover, a simply classification testing has been done on our implementation. The results show the basic working of our implementation. The 79.02% accuracy is reached. And big confusion lies between tone 1 and tone 4. Anyway the work described here is only based on the observation and the basic testing. From the work that we described here, we still can't say it will work very well. The final evaluation needs us to consider more and use it in the real tone-classification system. And according to the results of classification performance the methods will be evaluated. And it will be our further work.

5 REFERENCES

[1]. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal. "A comparative performance study of several pitch detection algorithms". *IEEE Transactions on Audio, Signal, and Speech Processing 24*, 399-417 1976.

[2]. M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans.Audio Electroacoust.,* vol. AU-16, pp. 262-266, June 1968.

[3]. Yi Kechu, Tian Fu, Fu Qiang, "YU YIN XIN HAO CHU LI", China Machine Press, BeiJing, 2000

[4]. M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J.Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-22, pp. 353-362, Oct. 1974.

[5]. Lawrence R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection" *IEEE* Trans. Acoust, Speech, Signal Processing, VOL. ASSP-25, NO. 1, 1977

[6]. X. Huang and A. Acero, H. Hon," Spoken Language Processing: A Guide to Theory, Algorithm, and System Development", Prentice Hall, 2001

[7]. S.-H. Chen and Y.-R.Wang. "Vector quantization of pitch information in Mandarin speech". *IEEE Transactions on Communications 38*(9), 1317-1320 1990.

[8]. C. Wang, "Prosodic Modeling for Improved Speech Recognition and Understanding", Ph.D. dissertation, MIT, June 2001

[9]. S. Potisuk, M. P. Harper., and J. Gandour. "Classification of Thai tone sequences in syllable-segmented speech using the analysisby-synthesis method", *IEEE Transactions on Speech and Audio Processing*, 7(1): 95-02,1999.