2011 17th Asia-Pacific Conference on Communications (APCC)
2nd – 5th October 2011 | Sutera Harbour Resort, Kota Kinabalu, Sabah, Malaysia

Melody To Musical Notation Translating System

Chooi Ling Si Toh, Chee Kyun Ng and Nor Kamariah Noordin
Department of Computer and Communication Systems Engineering, Faculty of Engineering,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
cell_st@hotmail.com, {mpnck, nknordin}@eng.upm.edu.my

Abstract— **This paper presents a system that translates the captured or recorded melody into musical notations automatically and instantly into a developed stave whereby a musician can compose music directly without any extra process or procedure, in real-time environments. In this translating system, the frequency of a captured or recorded melody is first analyzed through a microphone or musical instrument for its fundamental frequency. The analyzed fundamental frequency is then compared with the predefined frequency of musical notes. The matched musical note frequency will be distinguished at the developed musical stave interface instantly. This developed system can facilitate a composer in automatically translate his melody to musical notes without having to manually writing it down based on the melody he plays.**

Keywords- Musical Translating System; Musical Composing System; Musical Notation; Melody; Fundamental Frequency.

I. INTRODUCTION

Music composition is always ascribed to the art work of people who owned musical background. In order to compose a piece of music chord, ones must have the knowledge of musical notation elements which are pitch and rhythm. Pitch governs the melody and harmonic whereas rhythm elaborates the timing of musical sounds and silences [1]. With this knowledge, composers are able to elucidate their composition in a written form – the Music Score.

A great piece of composition is the conjunction of inspiration and creativity of a composer. Inspiration is mostly sourced by composer's mood, character, memories and their observation to the surroundings. The inspired melody line is then enriched by the creativity of the composer, where harmony and dynamic will be exerted to thicken the performance of the inspired melody line. Music is always the best medium to express the feeling and virtual idea across the boundary of geography as it is perceived as a global language that everyone does speak. It is the poet writing on our soul where it touches the soul without any literary reading [2].

In recent year, music industry has occupied the knowledge of computer science and engineering to develop the music score where the composer can elaborate the musical notation through mouse click while composing music [3], [4]. Also, the pitch and rhythm are displayed neatly at the monitor compared to traditional hand-written form. Furthermore, hardware devices also involve in music production in order to produce a high quality of music product [5]. The electronic musical instruments have been invented tremendously as the application of electronic industry is expanding. These musical instruments are mainly used in contemporary musical composition. The electronic piano are currently developed to integrate with personal computer to provide a platform for composer to compose their melody art work. A special designed software tool is used to enable the pitch of melody to be displayed according to the triggered keystroke on the piano keyboard [6].

However, the composing music activities should not restrict to merely individual who have the musical background. Everyone should be a music composer as long as they have the inspiration with a melody in their thought. Melody is a subjective and it differs among individuals. A great melody might be inspired by ones who do not have the music background. Thus, the great melody line will be abandoned when it does not being recorded or written down. This is ascribed to the waste of art as the precious idea which could not be expressed due to the lack of literary in musical notation. Although the advance of integration of computer science and engineering has developed various software and hardware equipments to enhance the musical composition workflow, but there is no any system tool that suitable for such individuals to express their idea in musical written form.

In this paper, a developed melody to musical notation translating system is presented to provide a solution for the composer to generate musical score efficiently. This system provides a platform that synthesizes the musical notation of the captured or recorded melody through singing or playing any musical instruments. The system is divided into three main parts. There are WAV File panel for the on-demand system, Microphone panel for the real-time system, and Result panel for the musical notation displaying system. The WAV File panel allows user to conduct pitch recognition or musical translation on the recorded monophony acoustic signal in .WAV file format. The Microphone panel allows user to conduct pitch recognition or musical translation in real-time environment by singing through microphone or playing musical instrument through a direct audio cable. The Result panel displays the corresponding musical notation for the analyzed pitch of the recorded or captured melody on the developed musical staves.

The rest of this paper is organized as follows: Section II presents the characteristic of musical acoustic signal. Section III discusses how to detect the pitch of recorded acoustic signal. The development of the melody to musical notation translating system is described in Section IV. The performance evaluations of the developed system are discussed in Section V. Finally, this paper is concluded in Section VI.

978-1-4577-0390-4/11/$26.00 ©2011 IEEE

II. CHARACTERISTIC OF MUSICAL ACOUSTIC SIGNAL

The musical acoustic signal is a periodic signal with frequencies that are multiple integers ("harmonics") of a fundamental frequency. The proportion is according to the associated harmonics in this sum to determine the sound timbre, while the sound pitch is determined by the fundamental frequency. Henceforth, the frequency of a musical acoustic is called the fundamental frequency [6]. A harmonic of a wave is a component frequency of the signal that is an integer multiple of the fundamental frequency. If the fundamental frequency is f, the harmonics have frequencies 2f, 3f, 4f. etc. The harmonics have the property that they are all periodic at the fundamental frequency; therefore the sum of harmonics is also periodic at that frequency [6].

Harmonic frequencies are equally spaced by the width of the fundamental frequency. These frequencies can be found by repeatedly adding of that frequency. For example, if the fundamental frequency is 25 Hz, the frequencies of harmonics are 50 Hz, 75 Hz, 100 Hz, etc. A plucked guitar string or a struck drum head or struck bell, which naturally oscillated with not only one, but several frequencies which are known as partials. Many of these partials are integer multiples of the fundamental frequency; these are called harmonics [6]. Harmonic wave motion is depicted as in Figure 1.

When a note is played on an musical instrument, such as middle "C" is played on a piano, it does not comprise of just one frequency but a complex and made up of many frequencies that are combined to yield the audible sound signal. Table 1 shows an example of Middle C frequency value and its initial harmonics. When these periodic waveform sums up, the frequency waveforms for the middle "C" (C4) note is formed as illustrated in Figure 2. Note that the musical signal is not a pure sine wave although human tends to hear a musical acoustic that may sounded at a particular frequency [6].

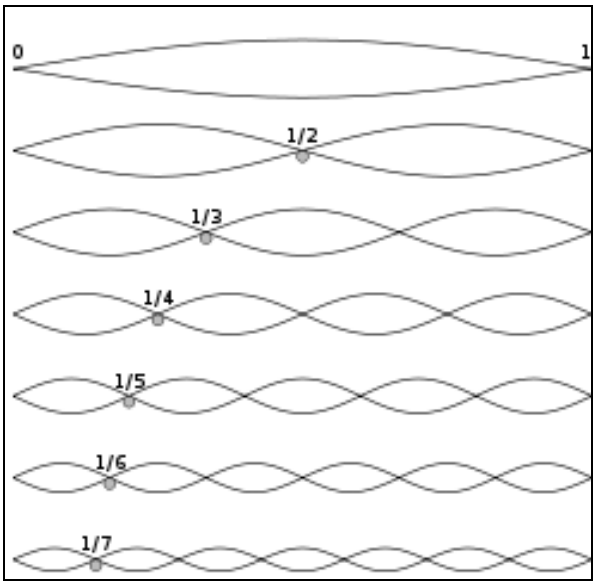

Figure 1. The illustration of harmonics for a vibrating string.

TABLE I. FUNDAMENTAL FREQUENCY AND HARMONICS OF MIDDLE "C".

C4 (Middle C)	262 Hz	Fundamental
C5	523 Hz	First Harmonic
G5	785 Hz	Second Harmonic
C6	1046 Hz	Third Harmonic
E6	1318 Hz	Forth Harmonic

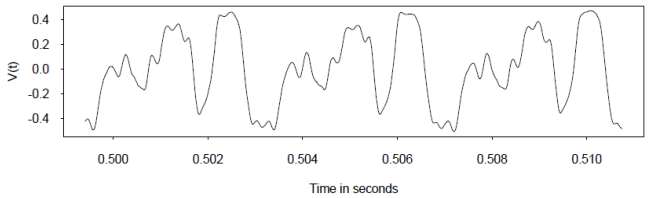

Figure 2. The waveform of middle "C" (C4).

The first or predominant frequency that human heard is fundamental frequency whereas the additional frequencies that make up the note are referred to harmonics which forms the timbre of musical sound signal. Nevertheless, the musical sound signal is not restricted to only pitch musical instrument, but also human vocal [6].

III. MELODY PITCH DETECTION ALGORITHM

The acoustics signal such as human voice and musical signal are ascribed to harmonic signal. To identify the pitch of harmonic signal, cepstral analysis is applied instead of spectral analysis. The cepstrum analysis method is known as a harmonic signal analysis tool commonly used to measure the fundamental frequency of speech. It tends to separate a strong pitched component from the rest of spectrum [7].

Cepstrum applies the principle where a frequency spectrum of a time harmonic signal that consists of a series of impulses at the fundamental and its harmonics, occurs at integer multiples of the fundamental. The cepstrum of a discrete-time signal x(n), CEPx(k) is defined as an inverse N-point Discrete-Time Fourier Transform (DTFT) multiplies with the logarithm of N-point Discrete Fourier Transform (DFT) for x(n) [7], and it is given by

$$CEPx(k) = \frac{1}{N}\sum_{k=0}^{N-1} \log\left[\left|\sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}kn}\right|\right] e^{j\frac{2\pi}{N}kn} \quad (1)$$

The cepstrum of a voiced speech segment is having a peak at the fundamental period of the segment. The discrete function CEPx(k) is used to search for the largest peak in the pitch period region of interest. If this peak value is above the detection threshold, the pitch period is defined as the location of this peak. If this peak is below this threshold, a pitch will not be detected [7]. For discrete-time signals, the cepstrum of a signal is defined as the inverse discrete time Fourier transform (IDTFT) of the logarithm of the magnitude of the DTFT of the signal. That is, the cepstrum of a signal x[n], c[n] is defined as

$$c[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X\left(e^{iw}\right)\right| e^{iwn}\, dw \tag{2}$$

where the DTFT of the signal is defined as

$$X(e^{iw}) = \sum_{n=-\infty}^{\infty} x[n] e^{-iwn} \tag{3}$$

Note that c[n], being an IDTFT, is nominally a function of a discrete index n. If the input sequence is obtained by sampling an analog signal, i.e., x[n] = xa(n/fs), then it would be natural to associate time with the index n in the cepstrum. Bogert et al. in [3] introduced the term quefrency for the name of the independent variable of the cepstrum in (2). This new term is useful to describe the fundamental properties of the cepstrum. For example, low quefrencies correspond to slowly varying components in the log magnitude spectrum, while high quefrencies correspond to rapidly vary components of the log magnitude.

To apply the derived mathematical function of cepstrum in speech processing, the discrete Fourier transform (DFT) which is computed with a Fast Fourier Transform (FFT) algorithm, is used with a sampled (in frequency) version of the DTFT of a finite-length sequence [1], [3]. The windowed signal is then transformed into frequency domain to obtain the spectrum that contained in the windowed signal by using Discrete Fourier Transform (DFT). The function of signal, X[k] after the DFT is defined as

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-i(2\pi k/N)n} \tag{4}$$

The log magnitude of X[k] is computed for power spectrum. The function of logarithm X[k], $\hat{X}[k]$ is fixed to

$$\hat{X}[k] = \log\left|X[k]\right| + i\arg\{X[k]\} \tag{5}$$

The function of cepstrum for x[n] discrete signal, $\overline{\hat{X}}[n]$ where the DTFT in (2) has been replaced by the finite DFT computation which can be derived as [7]

$$\overline{\hat{X}}[n] = \frac{1}{N}\sum_{n=0}^{N-1} \hat{X}[k] e^{i(2\pi k/N)n} \tag{6}$$

After the substitution of (4) into (5) and (5) into (6), the predefined function of cepstrum as (1) can be obtained. Figure 3 depicts the operations of cepstrum [3]. Cepstral peaks are the strong peak corresponding to the pitch period of the voiced speech segment that being analyzed [7].

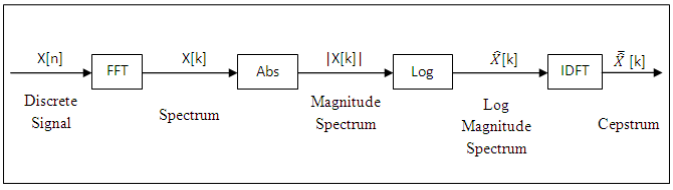

Figure 3. The computed Cepstrum using DFT.

A 512-point FFT was found sufficient to produce an accurate computation of the cepstrum. The cepstral peaks corresponding to the voiced segments are clearly resolved and quite sharp. Hence, the peak picking scheme is exploited to determine the cepstral peak in the interval of 2.5 - 15 ms, which corresponding to pitch frequencies between 60 - 400 Hz [4]. The result of cepstrum computation is a time sequence, just as the input signal itself. If the input signal has a strong fundamental pitch period in the region interested, the pitch period will show up as a peak in the cepstrum. The fundamental period of this pitch can be determined by calculating the time distance from time 0 to the time of peak. Figure 4 depicts the pitch determination from cepstrum plot [7]. From the cepstrum plot for a recorded note, the cepstral peak that marked by an asterisk is found at 2.52 ms. This indicated the note is cycled at a period of 2.52 ms and thus the fundamental frequency of this recorded note is 396 Hz.

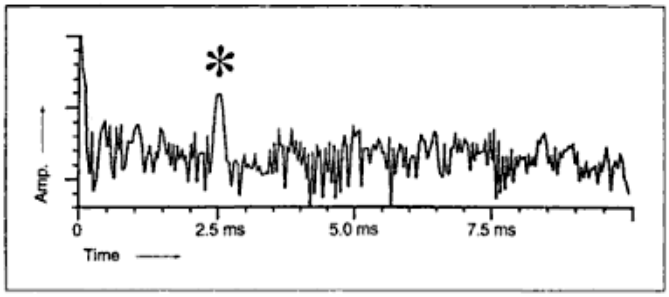

Figure 4. An example of pitch extraction.

The input signal for cepstrum analysis has to be segmented by sliding a window function across the signal as the input signal is an infinite impulse signal. The dataset after windowing is the input signal for (4) [3]. Figure 5 illustrate the windowing operation. The strength and existence of a cepstral peak for the voiced speech is depending on a variety of factors, including the length of the analysis window that applied to the signal and the formant structure of the input signal. The window length and the relative positions of the window and the speech signal will have considerable effect on the height of the cepstral peaks. If the window length is less than two pitch period long, a strong indication of periodicity cannot be expected. The longer the window, the greater the variation of the speech signals from the beginning to the end.

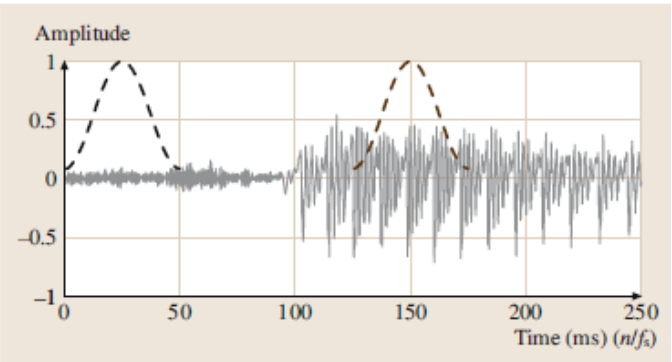

Figure 5. Segment of sampled speech waveform.

Therefore, considering the tapering effect of the analysis window as shown in Figure 5, the window length is set to 40 ms to capture at least two clearly defined periods in the windowed speech segment [4]. The Hamming window is applied to truncate the lengthy sound data. The window function of Hamming window is then defined as

$$W(n) = \begin{cases} 0.54 - 0.4\cos(\frac{2\pi n}{N}) & for \quad 0 \le n \le N \\ 0 \end{cases} \qquad (7)$$

where N is the sample point of the window. The magnitude is the response of Hamming window in time and frequency domains as illustrated in Figure 6 respectively.

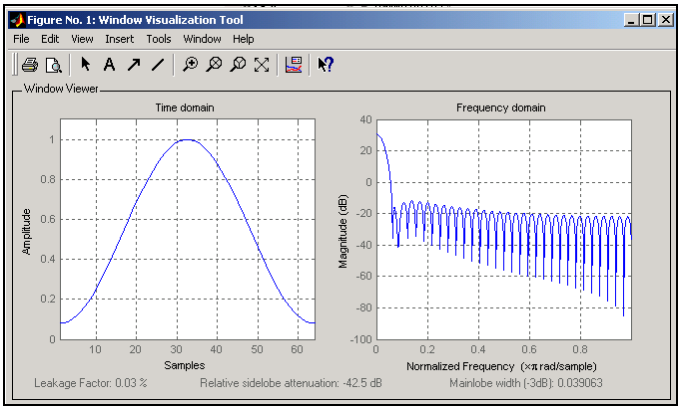

Figure 6. The magnitude response of Hamming Window in both time and frequency domains.

IV. System Development

The developed melody to musical notation translating system is initiated with input of the acoustic wave of melody in analog form for the analog to digital signal conversion. Once the analog signal is sampled and converted to digital signal, the sampled data of digital signal is stored as information that carries the characteristic of the fundamental frequency of the captured acoustic signal. These sampled data is then retrieved for the pitch determination by using the cepstrum analysis. The cepstrum analysis extracts the pitch of a concerned acoustic signal region by determining the peak of the fundamental frequency in the cepstrum. After the pitch value is determined, the corresponding musical notation will be synthesized based on its standard reference and displayed on a music score as the interface template. In order to make its system performance more precise, the evaluation of sound level and voice activity detection including the background noise filtering will be taken into consideration to assure only the effective voice region are processed for the cepstrum analysis.

Hence, the developed musical translating system consists of five main stages such as sound data collection, voiced region determination, fundamental frequency analysis, pitch matching and normalization, and autoshape coordinate allocation as shown in Figure 7.

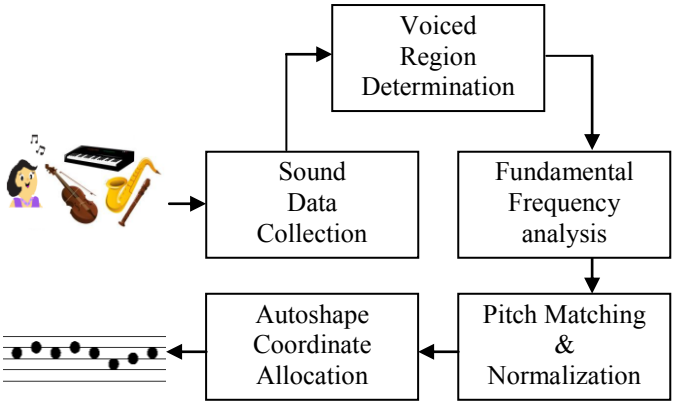

Figure 7. System overview of musical translating system.

A) Sound Data Collection

The preferable acoustic signal is being sampled at 44100 Hz or above. Data quantization will be in 16-bit for better digitization and reducing lost of data. It is also preferable to record the sound in an environment with less noise signal. This is to ensure the characteristic of the sampled sound data is not being attenuated such that it retains as much as possible of the original sound data characteristic.

The sources of collected musical acoustic signal could be from various inputs either human vocal singing or the family of musical instruments such as keyboard, string, brass and woodwind. The collected lengthy digitized sound data is then segmented into smaller frame with the length of power of two of FFT points before preceded to the pitch analysis.

B) Fundamental Frequency analysis

Acoustic signal is a harmonic signal which contains more than one spectrum peak. Cepstrum analysis is implemented to determine the fundamental frequency of the segmented acoustic sound data. The time-domain sound data is transformed into frequency-domain using FFT algorithm with the purpose of analyzing the sound spectrums. Since the sound data is a harmonic signal, the resulted spectrums are further analyzed to restrict its power values for determining the cepstral peak. By determining the restricted frequency with the highest power, the fundamental frequency of the windowed region can be found. The whole process of cepstrum analysis is shown in Figure 8.

C) Pitch Matching & Normalization

The estimated fundamental frequency value of the segmented sound data rarely hits the exact frequency for the precise pitching of musical note due to the digitization lost and noises. Thus, a frequency range is defined to allocate the estimated fundamental frequency to proper music note. The upper bound and lower bound of the frequency ranges for all standard musical notes are assigned by

Upper bound = standard pitch frequency + diff (8)

Lower bound = standard pitch frequency – diff (9)

where

diff = Difference between two consecutives notes/2 (10)

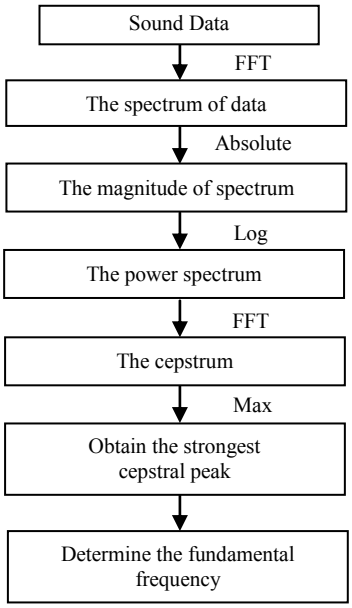

Figure 8. The flow of cepstrum analysis.

The identical computed music notes tend to repeat at every short distance of time as the sound data is segmented into a small frame for cepstrum analysis. Thus, these computed musical notes will be normalized before being recognized and displayed on music score template to avoid the redundant notation displays for same sound data.

D) Autoshape Coordinate Allocation

The computed musical note from the previous stage is projected to be displayed on music score template as its standard notation. The autoshapes such as circle, horizontal line are used to construct the musical notation elements. These autoshapes are drawn on axes accordingly to display the computed or recognized musical notes.

E) Voiced Region Determination

In order to yield a precise system performance where only voiced signal will be involved in fundamental frequency analysis, the captured acoustic signal will be analyzed for voiced region. Voice activity detection (VAD) and sound intensity level evaluation will be implemented to extract only the effective voiced region are processed for the cepstrum analysis. The VAD applies the concept of zero crossing rates to detect the active voice region. The rate of the framed sound data for crossing the zero level is used to determine the framed sound data is a voiced region or an unvoiced region. In order to determine the zero crossing rate of a framed sound data, the amplitude of both consecutive sound data within the frame is evaluated for its sign changes as

$$z(n) = \frac{1}{2}\sum_{m=1}^{N} |\,\mathrm{sgn}[x(m+1)] - \mathrm{sgn}[x(m)]\,| \qquad (11)$$

where

$$\mathrm{sgn}[x(m)] = \begin{cases} +1 & x(m) \geq 0 \\ -1 & x(m) < 0 \end{cases} \qquad (12)$$

A positive sound data will gives a +1 sign meanwhile a negative sound data gives a –1 sign. For an unvoiced region, the sound data tends to cross the zero level at higher rate as the amplitude of unvoiced signal is changes frequently from negative to positive and vice versa. Thus, it tends to accumulate the summation of sign in (11) and yield to a high zero crossing rate. While for a voiced region, the sound data are either remain at positive amplitude or negative amplitude. This will yield a low zero crossing rates. Thus, by determining the zero crossing rate, the voiced and unvoiced region of an acoustic signal can be determined.

The sound intensity level for the framed sound data is set with a threshold level of 50 dBA. To determine the sound level of a captured acoustic signal, the frequency spectrum of the segmented acoustic signal is obtained by using FFT algorithm so that the sound data is processed in frequency domain. In order to replicate human hearing response, the frequency spectrum sound data is modified by the A-weighting filter. The A-weighting filter will filter the very low and high frequencies sound signal. The total energy of captured acoustic signal is then measured in the frequency domain by applying Parseval's relation as

$$\text{Total energy} = \frac{1}{N}\sum_{n=0}^{N-1} |x[k]|^2 \qquad (13)$$

where x(k) is the amplitude of sound data and N is the length of sound data. By taking the mean energy across the segmented sound data, the mean energy value is further analyzed to determine the sound level by

$$\text{Sound pressure level} = 10\log_{10}(\text{Mean Energy}) \qquad (14)$$

V. Performance Evualution of The Developed System

To evaluate the accuracy performance of the developed system, a musical note C4 is played and captured by the system. Figure 9 shows plots of the amplitude, spectrum, and cepstrum for the captured C4 note. From the cepstrum plot, it shows that the peak is located at 0.003902s. Thus the fundamental frequency for the captured C4 signal is 1/0.003902 = 256 Hz, which is in the predefined domain of C4 signal.

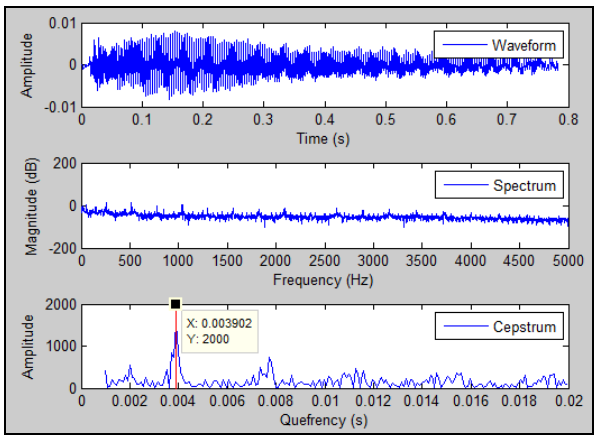

Figure 9. The amplitude, spectrum, and cepstrum plots for the captured C4 note.

When the musical notes C4 D4 and E4 are played and captured by the system. The sound data are analysed and its results in term of amplitude and frequency over timeline are shown in Figure 10. The computed fundamental frequency value are matched with the predefined frequency domain and its corresponding musical notation is displayed as shown in Figure 11. The evaluated accuracy performance of the system is environment dependent. The silent surrounding tends to yield a higher accuracy rate because the noise effect on acoustic signal very low. Figure 12 shows the accuracy performance for each note at ten trials. At higher sampling rates, the acoustic signal is being sampled at higher rate and thus it allows higher degree of preservation for the acoustic signal details. A more precise pitch value would be yield for segmented sound data at higher sampling rates. Figure 13 shows the percentage of note hit over different sampling rates. When the FFT length is longer, the system will load more sound data and only start to perform FFT algorithm after all points are valid. Although FFT algorithm computes DFT in an easier way and avoids redundant calculation, yet it still consumes some time and it becomes more significant for larger number of FFT points. Thus, lower FFT points length will yield shorter computational time. Figure 14 shows the computational time over different length of FFT points.

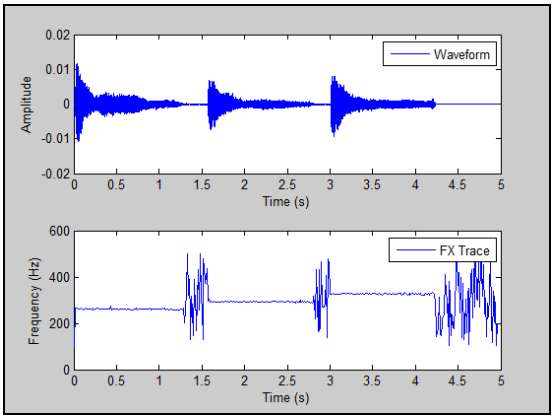

Figure 10. The amplitude and frequency over timeline of the captured C4 D4 E4 musical notes.

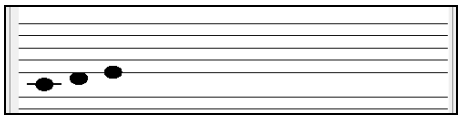

Figure 11. The displayed musical notation of captured C4 D4 E4.

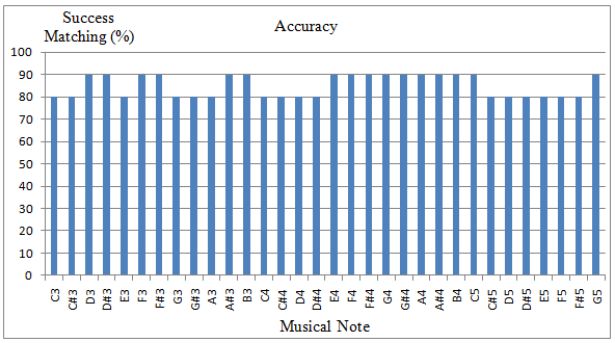

Figure 12. The accuracy performance for each musical note.

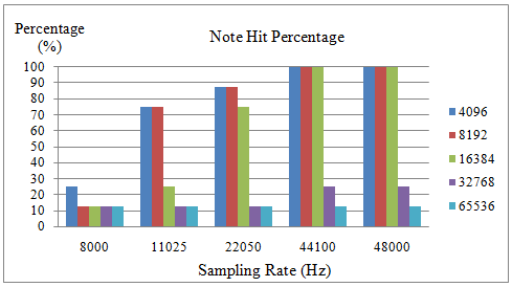

Figure 13. The hit rate across different sampling rate.

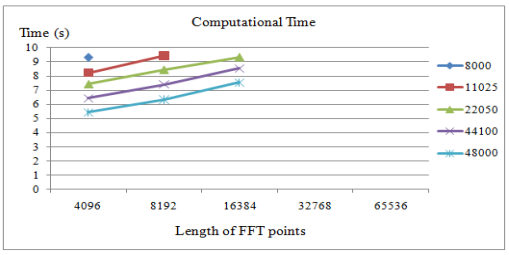

Figure 14. The computational time over different length of FFT points.

VI. CONCLUSION

The melody to musical notation translating system has been developed to display the musical notation for a captured acoustic signal by determining the fundamental frequency of the signal. The performance of the developed system has been evaluated where the optimized performance can be achieved at 48000 Hz sampling rate and 4096 FFT points. This musical translating system is best being implemented in noiseless environment as the acoustic signal is less being attenuated by noise signal and thus yield a better result.

REFERENCES

[1] S. C. Davies and D. M. Etter, "An Adaptive Technique for Automated Recognition of Musical Tones," The Thirtieth Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1138 - 1141, November 1996.

[2] B. F. Miessner, "Electronic Music and Instruments," Proceedings of the Institute of Radio Engineers, vol. 24 , no. 11, pp. 1427 - 1463, 1936.

[3] B. P. Bogert, M. J. R. Healy and J. W. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking," in Proceedings of the Symposium on Time Series Analysis. New York, ch. 15, pp. 209 - 243, 1963.

[4] S. Ahmadi and A. S. Spanias, "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm," IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pp. 333 - 338, May 1999.

[5] M. Ference, H. B. Lemon and R. J.Stephenson, Analytical Experimental Physics, The University of Chigago Press, 3rd Ed., 1956.

[6] N. E. Mastorakis, K. D. Gioldasis, D. Koutsouvelis, and N. J. Theodorou, " Study and Design of An Electronic Musical Instrument Which Accurately Produces The Spaces of The Byzantine Music," IEEE Transactions on Consumer Electronics, vol. 41, no. 1, pp. 118 - 124, February 1995.

[7] M. P. Norton and D. G. Karczub, Fundamentals of Noise and Vibration Analysis for Engineers, 2nd Ed., Cambridge University Press, September 2003.