Developing Speech Recognition System for Quranic Verse Recitation Learning Software

B.Putra[1], B.T. Atmaja[2], D.Prananto[3]
Department of Engineering Physics
Sepuluh Nopember Institutes of Technology
60111 Surabaya, Indonesia
{[1]budiman.fgi, [2]btatmaja,[3]dprananto}@gmail.com

***Abstract*— Quran as holy book for Muslim consists of many rules which are needed to be considered in reading Quran verse properly. If the recitation does not meet all of those rules, the meaning of Quran verse recited will be different with its origins. Intensive learning is needed to be able to do correct recitation. However, the limitation of teachers and time to study Quran verse recitation together in a class could be an obstacle in Quran recitation learning. In order to minimize the obstacle and to ease the learning process we implement speech recognition techniques based on Mel Frequency Cepstral Coefficient (MFCC) features and Gaussian Mixture Model (GMM) modeling, we have successfully designed and developed Quran verse recitation learning software in prototype stage. This software is interactive multimedia software which has many features for learning flexibility and effectiveness. This paper explains the developing of speech recognition system for Quran learning software which is built with the ability to perform evaluation and correction in Qur'an recitation. In this paper, the authors present clearly the built and tested prototype of the system based on experiment data.**

Keywords-Qur'an verse recitation, speech recognition, Mel Frequency Cepstral Coefficient (MFCC), Gaussian Mixture Mode (GMM)l.

I. INTRODUCTION

Al Qur'an is the holy book of Moslem consisted of life guidance which has to be understood clearly by them. Because of that, it is important for the Moslem to be able recite Qur'an verse correctly. Unfortunately, that is not easy because the Qur'an recitation has many rules which have to be strictly considered on the whole words. Moreover, if the recitation does not meet the rules, the recitation will give different meaning from the origins. So, the Moslems have to learn intensively in order to be capable in reciting Qur'an verse properly.

In contrast, the conventional learning method for Qur'an verse recitation learning is still uses *face to face* method which the student has to be taught directly by the teacher. Limited time to study together in a class and also the limited number of teachers result in lack learning process. Then the other learning lack process is the absence of lecturer because in conventional method, the lecture holds the important role in there. The development of learning software for Qur'an verse recitation is aimed to ease people in studying how to recite Quranic verses by them with correct and proper manner without abandoning the conventional way, the *face to face* method.

There are some methods which can be used to design and develop learning software for Quranic verse recitation. In the previous research which has done by Razak [4], the implementation of speech recognition for Quranic utterance style recognition using Hidden Markov Model (HMM) was developed. In this research, we use speech recognition techniques as part of template referencing method to develop the software. The techniques include Mel Frequency Cepstral Coefficient (MFCC) and Gaussian Mixture model (GMM) which use the threshold of log-likelihood value as the evaluation of utterances. This software will carry out correction and rectification of reading on each learning session if there is any error in reading. It is expected that people could learn to read the Quran easily without feeling doubt about the accuracy of pronunciation and recitation of Quranic verse recitation law (*tajweed*), because they would feel to have a virtual mentor which always makes corrections in learning process.

This paper is a revised version of conference paper presented at International Conference of Informatics for Development held Yogyakarta, Indonesia. Some of parts are modified and expanded from the origins. The speech recognition theory explained is revised to be more comprehensive with the research and more detail from the origins. Then, the research methodology part is modified with major changes to be clearer than before. In that part, we add the user interfaces of software as subsection to give the illustration for reader about our prototype software looks like and features built in this research. In addition, some minor revision also has done in other parts such as introduction part and the Quranic laws and pronunciations parts.

The rest of this paper is organized as follows. Section 2 presents the theory of Quranic verse recitation law and pronunciation. The next section will explain speech recognition techniques and its sub-theory which is necessary in building this system. The implementation and configuration on the building stage of software are described in the next section in Research Methodology. Then, the experiment conducted and its result is explained in Experiment Result section. Finally, the Conclusion and Discussion will be presented in Section 5.

This paper is a revised and expanded version of a paper entitled *"Prototyping of Quranic Verse Recitation Learning Software Using Speech Recognition Techniques Based on Cepstral Feature"* presented at *International Conference*

II. QURANIC VERSE RECITATION LAW AND PRONOUNCIATION

A. Makhraj

Makhraj means place of discharge. Makhraj in hija'iyah letters means the place where hija'iyah letters come out from mouth. Considering that Arabic letters differ from latin letters, hence Arabic letters pronounced in different manner to malay words pronunciation. Pronunciation of Arabic words/letter is determined by makhraj of the letter.

B. Mad

Mad means elongate tone. In tajweed course there are two kind of mad, i.e. mad ashli/tabi'i and mad far'i. Ashli means principal and far'i means subsidiary.

Mad ashli is read in two harakat. *Mad ashli* occur when there is *alif* after letters with *fathah*, *ya* consonant after letters with kasrah, and *waw* consonant after letters with dhomah. Furthermore, long-sign readings can be utilized if *alif*, *waw* consonant or ya consonant are not used.

Mad far'i has many different types with different vowel length. The application of the *Mad* is also different for every qiraat.

C. Law of "Nun" Consonant

The law of *nun* consonant and tanween can be classified in some types. *Izhhaar* means clearly read. Izhaar is readings where *nun* consonant or tanween meet *alif* (□), *hamzah* (□), *'ain* (□), *ghain* (□), *ha* (□), *kha* (□), *and ha'* (□) and read with clear sound.

Idgham means to get into/to change the tone of *nun* when nun consonant or tanween meet idgham letters. Each idgham readings read in two harakat. Idgham Letters are *ya* (□) ra (□), *mim* (□), *lam* (□), waw (□), and *nun* (□).

Idgham bilaghunnah is an inverse of *idgham bighunnah*, where the tone is not inserted into the nose. The letters of *idgham bilaghunnah* are "*lam*" and "*ra*".

Iqlab occur when *nun* consonant or tanween meet *ba* (□). The tone of "*ba*" in Iqlab readings changed to "*mim*" accompanied with drone. Iqlab readings are read in two harakat.

Ikhfa' means to hide/vague. *Ikhfa'* readings is read by vague voice of nun when nun consonant or tanween meet *ikhfa'* letters. All the readings with ikhfa' read in two harakats. *Ikhfa'* letters are the letters except in *izhhaar, idgham and iqlab.*

D. Law Mimi Consonant

Idgam mutamatsilayn occurs when mim consonant meet "*mim*", where the tone of mim in mim consonant is inserted to the next letter tone with buzz tone. This readings is read with two harakats.

If mim consonant meet "ba", then the tone of mim in mim consonant read with vague with a bit of buzz. It is read with two harakats.

If *mim* consonant meet letters other than "*mim*" and "*ba*", then the tone of mim in mim consonant read obviously. *Izhaar syafawi* is read with one harakat.

Ghunnah means to buzz. *Ghunnah* occur in two cases, that is when "*mim*" and "*nun*" use tasydid sign. *Ghunnah* is read with two harakats.

III. SPEECH RECOGNITION TECHNIQUES

Signal processing was utilized to obtain the characteristics of pronunciation which latter be used as identifier of faults and correction. Extraction of MFCC coefficient, signal energy, delta MFCC and delta-delta MFCC is conducted to subtract the feature of voices.

A. Voice Signal and its Occurrence Process

Voice is signals which greatly influenced by frequency and a form of discrete signal which is influenced by time. The main component in voice production system is vocal tract. Vocal tract is a resonance tube-shaped object in voice production system which has three main parts called pharynx, nasal cavity and oral cavity. The vocal tract varies in shapes according to soft plate (*velum*), tongues, lips and jaw which overall called as articulators.

See Figure 1 below, the vocal organ consists of (1) Nasal cavity, (2) Hard palate, (3) Alveolar ridge, (4) Soft palate (Velum), (5) Tip of the tongue (Apex), (6) Dorsum, (7) Uvula, (8) Radix, (9) Pharynx, (10) Epiglottis, (11) False vocal cords, (12) Vocal cords, (13) Larynx, (14) Esophagus, and (15) Trachea.

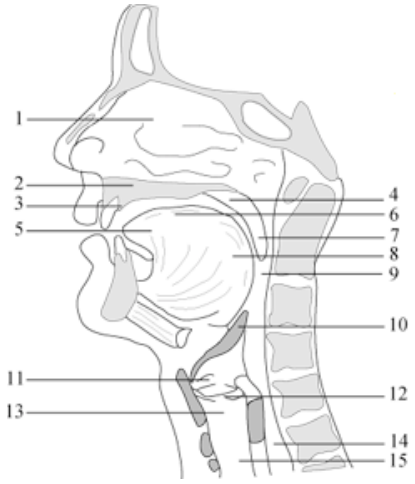


Figure 1. Human vocal organs [14]

The process of human voice production can be explained as follows; the air flows from lungs to the trachea, a tube composed of cartilage rings, and goes through larynx to the vocal tract. Larynx reacts as a gate between lungs and mouth. Larynx composed of epiglottis, vocal cords and false vocal cords. These three components are closed when human shallow food, so that the food does not enter the lungs, and open again when the human inhale. Phoneme in English can

be classified in terminology of "manner of articulation" and "place of articulation".

Manner of articulation is concentrated on air flow, it means it concerns about track and the level of vocal that goes trough. Manner of articulation and voicing is divided into three big classes of phoneme. Phoneme that produce employ voicing and solely stimulate the vocal tract on glottis called *sonorant* (vowels, diphthongs, glides, liquids and nasals). They have continuous, intents and periodic phonemes characteristics.

Voiced sounds is produced by air pressure which flows trough vocal cords while vocal cords squeezed to open and close quickly to produce a series of puffs periodic which have fundamental frequency (first harmonics) same as vocal-cord's vibration frequency. The frequency of vocal cords depends on the level of solidity, tension, length of vocal cords and air flow effect which produced in the glottis, a chamber between vocal cords. The component of this frequency is composed of a number of harmonics from fundamental frequency. A sound that produced without vibration in vocal cords is called unvoiced [2].

The following picture illustrates a segment on vowel /ix/. A quasi-periodicity signal on voiced speech can be seen here.

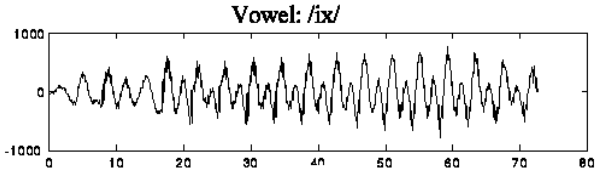


Figure 2. Illustration of vowel segment /ix/ [2]

Fricative sound is generated from the confinement of vocal tract and air flow pressure with high enough velocity to make turbulence. The turbulence is for instance /hh/ or /s/.

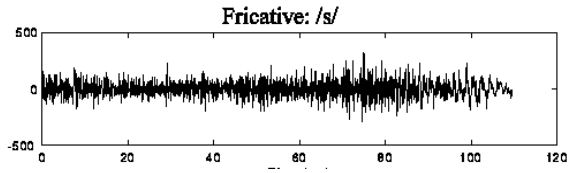


Figure 3. Fricative /s/ [2]

Plosive or stop sounds is generated from vocal tract blocking process by closing the lips and nasal cavity, enabling lateral air pressure, and followed by a beat. This mechanism will generate /p/ and /g/ voices.

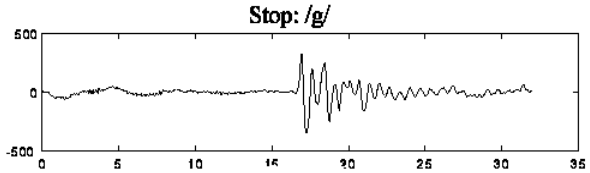


Figure 4. Plosive or stop sounds /g/ [2]

Affricate is a combination of stop and fricative sounds. Stops, fricative and affricates collectively called as obstruent *phoneme* which is weak enough and periodic, and basically is a form that generated by blocking stimulus on main vocal tract. Vowel is basically described in terminology of tongue position and lips

B. Speech Signal Processing

Speech signal processing is intended to obtain cepstral feature of human voice. The process of voice signal processing consists of Sampling, Frame Blocking, Windowing, Discrete Fourier Transform (DFT), Filter Bank, Discrete Cosine Transform (DCT) and calculating of dynamical coefficient and spectrum energy.

1) Sampling

Human voices will generate continuous analog signals. Therefore, the analog signal is chopped in certain interval of time. Discrete series sample x [n] is obtained from continuous signal x(t),

$$x[n] = x(nT) \tag{1}$$

Where T is sampling period and i/T = Fs is sampling frequency in unit of sample/second. The value of n is the number of samples. According to the Nyquist sampling theory, minimal sampling frequency required is twice of original maximal signal.

$$FSampling \geq 2 \times FSignal \tag{2}$$

2) Frame Blocking

Frame Blocking is segregation of voices into several frames where one frame consisted of several samples. This process is needed to transform a non-stationer signal to a quasi-stationer signal so it can be transformed from time domain to frequency domain with Fourier transform. Human voice signal indicate quasi-stationer characteristic in range of time from 20 to 40 milliseconds. Hence, in that range of time the Fourier transform can be performed.

3) Windowing

Voice signal which is chopped into frames will lead to discontinuity in initial and final signal. The discontinuity leads to data error in the process of Fourier transform. Windowing is needed to reduce the effect of discontinuity in chopped signal. If window is defined as w(n), where $0 \leq n \leq N-1$ and N is number of samples in each frame, then the result of windowing process is;

$$w(n) = x(n)W(n), \quad 0 \leq n \leq N-1 \tag{3}$$

Windowing which is used in this research is Hamming windowing.

$$W_{hann} = \begin{cases} (0.52 - 0.46\cos(2\pi n/(N-1)) & 0 \le n \le N-1 \\ 0 & others \end{cases} \quad (4)$$

4) Discrette Fourier Transform (DFT)

Fourier transform is performed to transform from time domain to frequency domain. DFT is a specific form of integral Fourier equation;

$$Y(\omega) = \int w(t) e^{-j\omega} dt \quad (5)$$

The DFT can be obtained by changing the variables time (t) and frequency (w) into discrete form:

$$Y(k\omega_0) = \sum_{n=0}^{N-1} w(nT) e^{-jk\omega_0 nt} \quad (6)$$

5) Mel Frequency Cepstral Coefficient (MFCC)

The most important information of human voice signal is located at high frequency. This important information determines the characteristic of human voice and Mel Scale is utilized to accommodate this characteristics. The relation between Mel and actual frequency is according to various researches about perception of voice reception by human ear .

$$Mel = 1000 \times \log_2(1+\omega) \quad (7)$$

In the implementation, this scaling is interpreted with Mel Filter Bank where each value of frequency magnitude is filtered by triangle filter series with Mel frequency as middle frequency. The triangular filter represents the process of Mel scaling in the signal.

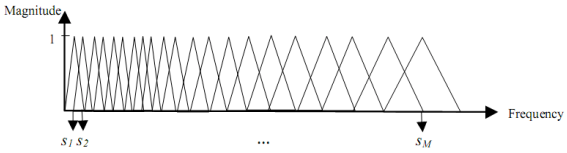


Figure 5. Mel filter bank construction

After the magnitude of signal spectrum X[k] filtered by Mel Filter Bank, computation of logarithmic value of energy is conducted to each of output band from each filter. Logarithmic signal energy process is utilized to adapt the system just like human ear.

$$s[m] = \ln\left[\sum_{k=0}^{N-1} |Y[k]|^2 H_m[k]\right] \qquad 0 \le m \le M \quad (8)$$

To obtain the MFCC, the result of energy logarithmic is processed with Discrete Cosine Transform (DCT).

$$c[n] = \sum_{m=0}^{M-1} s[m] \cos\left(\frac{\pi n(m-0.5)}{M}\right) \quad (9)$$

6) Calculation of Dynamical Cepstral Coefficient and Spectrum Energy.

In order to add the pattern of recitation speech, the system is equipped with dynamical cepstral coefficient and spectrum energy. The dynamical cepstral coefficient is inherited from MFCC each frame separated to two forms coefficient, delta MFCC and delta-delta MFCC

Delta MFCC naturally is the first regression of MFCC in each frame. The equation below is the computation of delta MFCC.

$$d_i = \frac{\sum_{n=1}^{N} n(c_{n+1} - c_{n-1})}{2\sum_{n=1}^{N} n^2} \quad (10)$$

Where d_i is the delta coefficient at frame i computed in terms of the corresponding basic coefficients c_{n+1} to c_{n-1}. The same equation is used to compute the acceleration coefficients by replacing the basic coefficients with the delta coefficients.

Then, the delta-delta MFCC or acceleration of MFCC can be computed as follows,

$$a_i = \frac{\sum_{n=1}^{N} n(d_{n+1} - d_{n-1})}{2\sum_{n=1}^{N} n^2} \quad (11)$$

where a_i is the delta coefficient at frame i computed in terms of the corresponding basic coefficients d_{n+1} to d_{n-1}

C. Gaussian Mixture Model (GMM)

Gaussian probability density function (pdf) is bell-shaped one dimensional function which is defined by two parameters, that is mean μ and variant σ or covariant Σ. In D dimension it can be formulated as;

$$N(\mathbf{x};\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \quad (12)$$

where μ is vector mean and Σ is covariant matrix.

Gaussian mixture model (GMM) is a mixing of several Gaussian distribution or representation of the existence of subclasses in a class. The probability density function of GMM is described as sum of multiplication of weight with Gaussian probability.

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{c=1}^{C} \alpha_c N(\mathbf{x};\boldsymbol{\mu}_c, \Sigma_c) \tag{13}$$

α_c is weight of mixed component c, where $0 < \alpha_c < 1$ for each component and $\sum_{c=1}^{C} \alpha_c$. Whereas, the parameter distribution,

$$\boldsymbol{\theta} = \{\boldsymbol{\alpha_1}, \boldsymbol{\mu_1}, \Sigma_1, ..., \boldsymbol{\alpha_C}, \boldsymbol{\mu_C}, \Sigma_C\} \tag{14}$$

It is the definition of Gaussian mixture probability density function parameter. The figure blow is the illustration of GMM with three mixtures

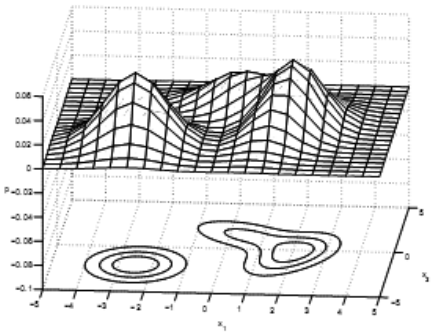


Figure 6. Example of GMM dimension-2 with 3 mixtures

IV. RESEARCH METHODOLOGY

This part will explain about the prototype system and the testing procedure to obtain the system performance. Also, in the last of this part the result of recognition performance test is reported.

A. *Description of software and features*

This software is a prototype stage software used to learn how to read Al Qur'an correctly. Commonly, Al-Quran recitation learning book require more competent supervisor. However, this software is an interactive multimedia software accompanied with pronunciation and verse recitation law correction in each courses.

This software consists of three modules or levels depend of the difficulty and learning material, i.e. basic, intermediate and advance. In the basic level, the course includes only correction in *makhraj*. The module included in this level consists of letters of *Hijaiyah*. Then, in intermediate level, the courses consist of law of recitation and also correction in *makhraj/*pronunciation. In this level there include only one kind of recitation law such as *Idhar, Ikhfa'* or *Iqlab* only. The highest level, advance, include courses with combination of more than one law of recitation and still with pronunciation correction. Moreover in this last level, both pronunciation and verse recitation law are evaluated. So, the higher level we take the higher difficulty of the course. If user makes an error during the learning, the error message will appear on the software which will guide the user to correct the reading and pronunciations.

In order to support the learning process, each module in the software equipped with the sample recitation. This is needed to make the user understand how to recite correctly. Also by this recitation sample, the user can copy the recitation to ease his learning process.

As the software supporting features, software is also equipped with a tutorial how to read *Al-Qur'an* recitation correctly as guidance and pre-learning. The tutorial is recommended to read every time we make error so we know why it is wrong.

B. *Design of Correction System*

The figure below shows the general step of designing and application of software,

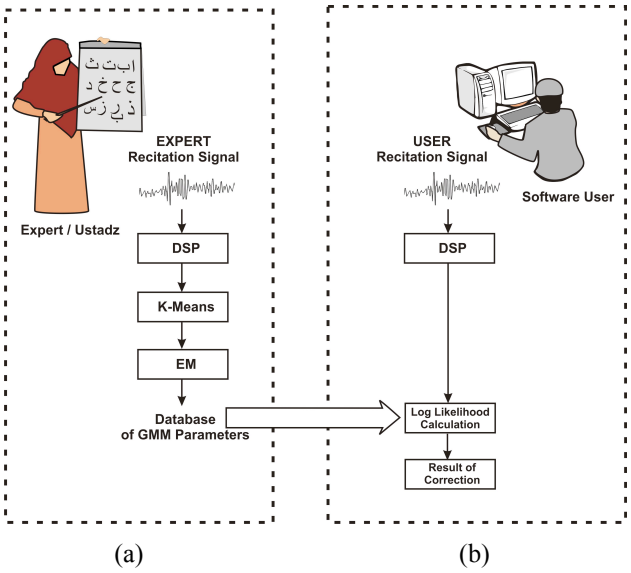


Figure 7. (a) Designing and template making step of software using expert/ustadz recitation (b) Process of software application used by user

The basic idea of the correction and evaluation system is template matching using speech recognition, which the cepstral feature of the voice of each reading is taken for main pattern of recognition. The voice used as main pattern is recorded from the expert of Qur'an recitation. Also this voice can be played in software as module recitation sample voice.

The Gaussian Mixture Model (GMM) is used to model the speech features of recitation. In this research, the number of mixture in each GMM consisted of 6 mixtures. This number is selected based on consideration of minimum software computation time.

Formulation and design is done prior to software construction and design. General formulation of the parameters of GMM modeling which is used to recognize the reading is obtained from this stage.

The stage of design and formulation is composed of digital signal processing (DSP), GMM modeling and logging the GMM parameters into software database. Digital signal processing is consisted of several sub process as follows, see Figure 8. The incoming voice signal is processed in order to extract the cepstral feature parameters. The incoming voice signal is sampled with sampling frequency of 8000 Hz, accordance with Nyquist rule, and then divided into time slots (framing) with frame time length 40 ms and overlapping time 20 ms or about 50%. The number of frame is separated depend on the recitation word number or *harakat*. Then, each frame is passed through Hamming window to reduce signals discontinuity after framing. The signal is then transformed to frequency domain by using DFT with N = 1024. The signal is passed through the Filter bank of 24 triangles mel filter. The DCT with MFCC coefficient of 14 is done after filtering process. Other feature calculation like signal energy, delta-MFCC and delta-delta MFCC were done after DCT process.

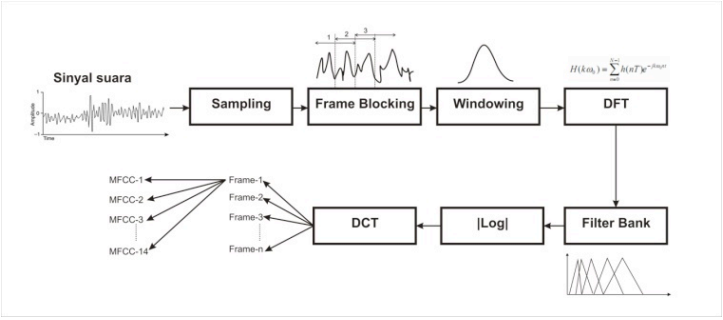


Figure 8. Digital Signal Processing for Extracting Cepstral Feature

The four signal feature vectors in each group word are used as input of the observation data in the GMM. The number of state and model which is used to design the software is not specified to adapt with verses section or the session of each reading in the course. The process of GMM modeling can be seen in figure below. The first step is initialization of clustering as many as number of GMM mixtures. After initialization by K-Means clustering, the next process is GMM training to obtain the maximum likelihood using Expectation-Maximation (EM) algorithm. In this training use maximum iteration until 1000 and convergence value is 0.001. After several iterations, if convergence or maximum iteration is achieved the iteration can be stopped and all the parameter is stored into the software database. Each recitation law correlated with word numbers (*harakat*) in the course has one GMM model for correction as the reference template of recitation model in the application stage.

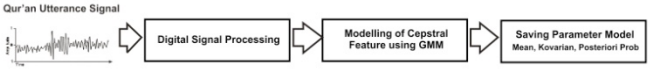


Figure 9. Process of Software Construction

After database is filled by GMM parameters, the software can be used in application or software using stage. At this stage, correction process is done with similar process at the construction and formulation stage. However, as figure 7b on the top, at this stage likelihood estimation with GMM parameter for each recitation law/pronunciation is done after signal processing process. Furthermore, if the likelihood value is less than the predetermined threshold value then certainly an error occurred in reciting the readings, see figure 10. The error message a is appeared in order to warn the user to reread the readings with proper and correct law of recitation.

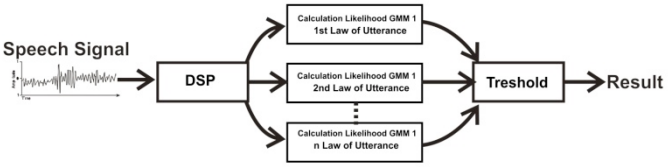
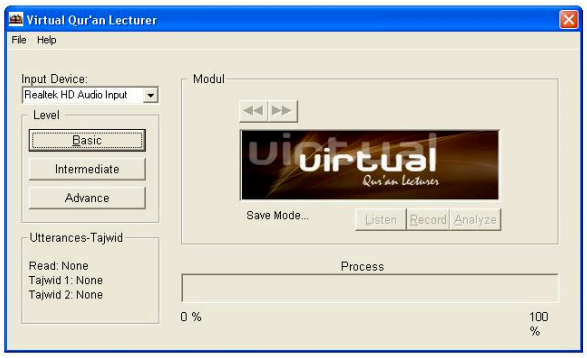


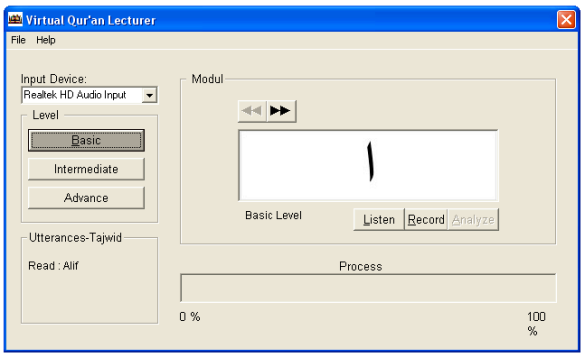
Figure 10. The Correction Process of Software

C. *User Interfaces (UI)*

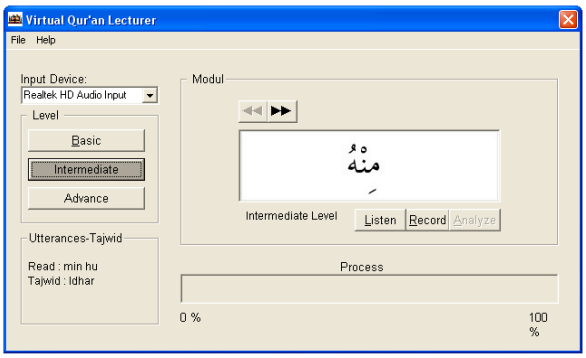
As user software, the interface is important part attracting the user to use the software. Below is the user interface of software we built in three level modules,



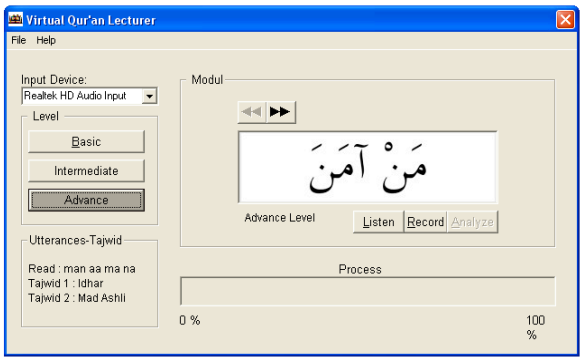
(a)



(b)



(c)



(d)

Figure 11. User infeace software (a) in standby mode (b) level Basic Module (c) level Intermediate Module (d) level Advance Module

We can see figure on the top the user interface of this prototype software is designed simply. In further research, we develop this interface in object to ease the user in learning process.

V. Result And Discussion

In order to test the reliability and accuracy of correction, experiment for each recitation law in the software is done with ten speakers to read the readings in wrong and right manner. This test used clean speech signal as the template.

1) *Analysis of pronunciation correction (Makhraj)*

The accuracy of correction for *hija'iyah* letter pronunciation is obtained from the average result for all *hija'iyah* letters. For example, how much *alif* recited true by the system compared with real true value of pronunciation of *alif* and then other *hija'iyah* letters. The average result from all *hija'iyah* letters is calculated as the correction accuracy which the value is 90%. This value is higher than we expected. However, this value is premature data as the experiment data is not enough for evaluate the speech recognition based system.

2) *Accuracy of Quranic Recitation Law*

Accuracy of Quranic recitation law can be obtained by testing the system of Quranic recitation. For a word containing some letters, the law might be *idghom, ihkfa'* and *idhar*. How many law from some Arabic words is detected true compared with real true value is obtained as the accuracy of Quranic recitation law. For recitation law correction test, the result suggests that the correction accuracy is fairly poor (70%), hence, the system needs a reconfiguration in order to improve the correction of recitation law.

3) *Combination of several Makhraj and recitation law*

The last accuracy test for the system is the combination of *makhraj* and Quranic recitation law. It can be obtained by testing both *makhraj* and Quranic recitation law from some Arabic words and compare the result with real true value. For combination of several *makhraj* and recitation law test, the result suggests that the correction accuracy is also poor (60%), hence, the system need a reconfiguration back in order to improve combination of *makhraj* and recitation law.

VI. Conclusion

The prototype has been designed as an interactive multimedia Quran recitation learning software using cepstral feature and GMM modeling as the basis of speech recognition technology. The results of the research suggest that the correction accuracy of the software is 70% for pronunciation, 90% for recitation law and 60% for combination of pronunciation and recitation law. The performance of correction accuracy can be improved by changing the configuration and template of speech recognition used. .

We are working on adding the speech enhancement and data training in object to improve the performance of learning software. Also, recently we are developing the features of software in order to improve the feasibility of this software to be used.

References

[1] B.Putra, B.T. Atmaja, D.Parananto. 2011. Prototyping of Quranic Verse Recitation Learning Software Using Speech Recognition Techniques Based on Cepstral Feature. Proceeding of International Conference of Informatics Development 26 Novermber 2011, Yogyakarta, Indonesia.

[2] Bardici, Nick. Speech Recognition using Gaussian Mixture Model. PhD Thesis. Blekinge Institute of Technology. 2006

[3] Moreno, Pedro J. Speech Recognition in Noisy Environments. PhD Thesis. Carnegie Mellon University. 1996

[4] Razak, Zaidi. Quranic Verse Recitation Feature Extraction Using Mel-Frequency Cepstral Coefficient (MFCC). Journal University of Malaya. 2007

[5] Berouti, M., Schwartz, R. And Makhoul, J. . Enhancement Of Speech Corrupted By Acoustic Noise. Proc. Of IEEE ICASSP, Pp. 208-211, Washington DC.1979

[6] Bhatnagar, B.E., Mukul. A Modified Spectral Subtraction Method Combined With Perceptual Weighting For Speech Enhancement. MSc Thesis. The University Of Texas At Dallas. 2002

[7] Guard, Cedric. Speech Recognition Based On Template Matching and Phone Posteriori Probability. MSc Thesis. IDIAP Research Institute. 2007.

[8] Patel, Ibrahim. Speech Recognition Using HMM With MFCC-An Analysis Using Frequency Specral Decomposion Technique. Signal & Image Processing : An International Journal(SIPIJ) Vol.1, No.2, December 2010.

[9] Lima, Carlos. Spectral Normalisation MFCC Derived Features for Robust Speech Recognition. SPECOM'2004: 9th Conference Speech and Computer St. Petersburg, Russia September 20-22, 2004.

[10] Bala, Anjali. Voice Command Recognition System Based On MFCC And DTW. International Journal of Engineering Science and Technology Vol. 2 (12), 2010, 7335-7342.

[11] Shaneh, Mahdi. Voice Command Recognition System Based on MFCC and VQ Algorithms. World Academy of Science, Engineering and Technology 57, 2009.

[12] Rabiner, Lawrence R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE Vol 77 No 2, February 1989.

[13] B.H. Juang, Lawrence R Rabiner. Hidden Markov Model for Speech Recognition. Technometrics, Vol. 33, No. 3. (Aug., 1991), pp. 251-272.

[14] Review of Speech Synthesis Technology. http://www.acoustics.hut.fi/publications/ files/theses/lemmetty_mst/index.html. [Online]. Available: http://www.acoustics.hut.fi/ publications/files/theses/ lemmetty_mst/chap3.html

AUTHORS PROFILE

Budiman Putra received his Bachelor of Engineering in Engineering Physics majoring Instrumentation and Control System Engineering from Department of Engineering Physics, Sepuluh Nopember Institute of Technology Surabaya, Indonesia. Currently, he is a Defense and Control System Engineer in PT Len Industri (Persero), Bandung, Indonesia. His research interests include Digital Signal Processing, Image Processing, Artificial Intelligence and Electronic Control System.

Bagus Tris Atmaja received his Bachelor and Master of Engineering degree from Department of Engineering Physics, Sepuluh Nopember Institutes of Technology in 2009 and 2012 respectively. Now, he is a research sudent in Usagawa-Chisaki Lab, Kumamoto Univ., Japan. His research interests include Artificial Intelligence, Digital Signal Processing, Acoustics and Vibration.

Dwi Prananto received his Bachelor degree in Engineering Physic from Sepuluh Nopember Institute of Technology (ITS), Surabaya, Indonesia. Currently, he is Master Student at Department of Physics, Tohoku University, Japan. His research interests include Smart Material, Thermoelectric Materials and Instrumentation Scientific.