Prototyping of Quranic Verse Recitation Learning Software Using Speech Recognition Techniques Based on Cepstral Feature

B. Putra
Department of Engineering Physics
Institut Teknologi Sepuluh
Nopember (ITS)
Surabaya, Indonesia

B.T. Atmaja
Department of Engineering Physics
Institut Teknologi Sepuluh
Nopember (ITS)
Surabaya, Indonesia

D. Prananto
Department of Engineering Physics
Institut Teknologi Sepuluh
Nopember (ITS)
Surabaya, Indonesia

***Abstract*—Al-Quran, the holy book of moslem, is written and recited in its native language, the Arabic. For Indonesian it is troublesome, because it is not their native language. Proper pronunciation of the verses is important, because different pronunciation will gives different meaning. Intensive learning is needed to be able to recite Quranic verses properly. However, the limitation of teachers and time to study Quran verse recitation together in a class could be an obstacle in Qur'an recitation learning. Hence, a learning tool is needed to be developed in order to overcome the obstacle. By implementing speech recognition techniques based on cepstral feature coefficient and Gaussian Mixture Model (GMM) modeling, we have designed and developed Quranic verse recitation learning software. This software is interactive multimedia software which is developed to help and to ease people in learn how to recite Qur'an properly. This system is equipped with the ability to perform correction in Qur'an recitation and guide user to recite Qur'an in correct and proper manner with the help of tutorial provided. In this research paper we have built and tested the prototype of this system.**

Keywords- qur'anic verse recitation, learning software, speech recognition, cepstral feature

I. INTRODUCTION

The conventional learning method for Quran verse recitation learning is face to face method. Limited time to study together in a class and also the limited number of teachers result in lack learning process. The development of learning software for Quran verse recitation is aimed to ease people in studying how to recite Quranic verses with correct and proper manner without abandoning the conventional way, the 'face to face' method. It means that the software is just an additional tool to improve their Quranic verse recitation skills by themselves, not as main tool for learning Quran.

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 [3], the implementation of speech recognition for Quranic utterance using Hidden Markov Model (HMM) was developed. In order to improve the reliability of the system, in this research, we use speech recognition techniques as part of template referencing method to develop the software. The techniques include cepstral feature 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 is always make corrections in learning process.

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 *harakah*. *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 *harakah*. *Idgham* Letters are "*ya*", "*ra*", "*mim*", "*lam*", *waw*" and "*nun*".

Idgham bilaghunnah is an inverse of *idgham bughunnah,* 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 *harakah.*

Ikhfa' means to hide/vague. *Ikhfa'* readings read by vague voice of *nun* when *nun* consonant or *tanween* meet *ikhfa'* letters. All the readings with *ikhfa'* read in two *harakah*. *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 *harakah*

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 harakah.

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 *harakah.*

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

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.

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 divideed into three big classes of phoneme. Phoneme that produce employ voicing and solely stimulate the vocal tract on glottis called *sonorants* (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 [1].

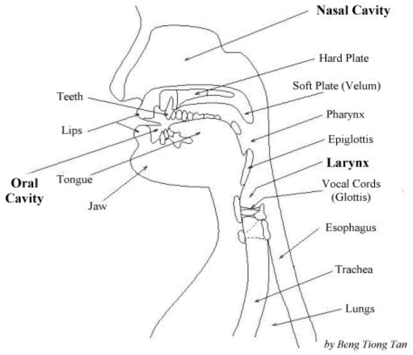

Figure 1. Human's mouth cavity [2]

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

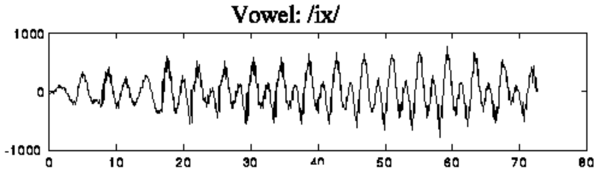

Figure 2. Illustration of vowel segment /ix/ [2]

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/.

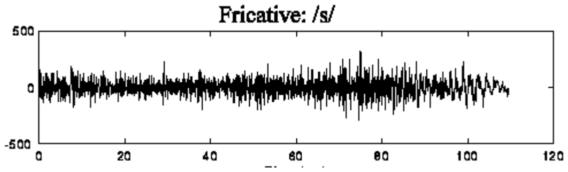

Figure 3. Fricative /s/ [2]

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.

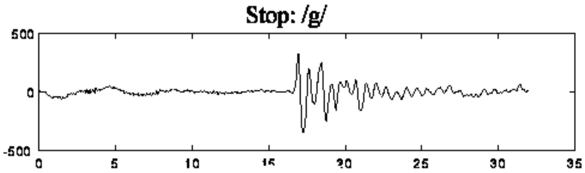

Figure 4. Plosive or stop sounds /g/ [2]

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 and Discrete Cosine Transform.

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),

$$x[n] = x(nT) \qquad (1)$$

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.

$$FSampling \geq 2 \times FSignal \qquad (2)$$

2) Frame Blocking

Frame Blocking is division 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 ≤ n ≤ N-1 and N is number of samples in each frame, then the result of windowing process is;

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

Windowing which is used in this research is Hamming windowing.

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

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;

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

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

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

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 .

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

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.

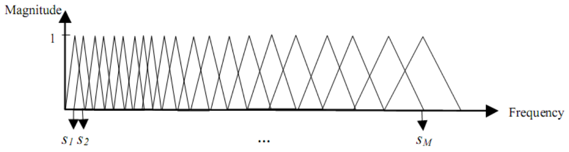

Figure 5. Filter bank construction

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.

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

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

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

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;

$$N(\mathbf{x};\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{10}$$

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.

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{c=1}^{C} \alpha_c N(\mathbf{x};\boldsymbol{\mu}_c,\Sigma_c) \tag{11}$$

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

$$\boldsymbol{\theta} = \{\boldsymbol{\alpha_1},\boldsymbol{\mu_1},\Sigma_1,\ldots\ldots\ldots\ldots,\boldsymbol{\alpha_C},\boldsymbol{\mu_C},\Sigma_C\} \tag{12}$$

It is the definition of Gaussian mixture probability density function parameter.

.

IV. RESEARCH METHOD

This part of research paper 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. General purpose of systems

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

There are several modules or levels in this software depend of the difficulty, i.e. basic, intermediate and advance. In the basic level, the course includes only correction in *makhraj*. 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. The advanced level, include courses with combination of more than one law of recitation and still with pronunciation correction. So, the higher level we take the higher difficulty of the course. If we make an error during the learning, the error message will appear on the software which will guide the user to correct the reading and pronunciations.

The software is 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 System

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 Gaussian Mixture Model (GMM) is used to model the signals. 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 (27) (DSP), GMM modeling and logging the GMM parameters into database. 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 40 ms and overlapping time 20 ms or about 50%. The number of frame is separated depend on the recitation word number. Then, each frame is passed through Hamming window to reduce signals discontinuity after chopping. The signal is then transformed to frequency domain by using DFT with N = 1024. The signal is passed through the Filter bank of 24 triangle 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.

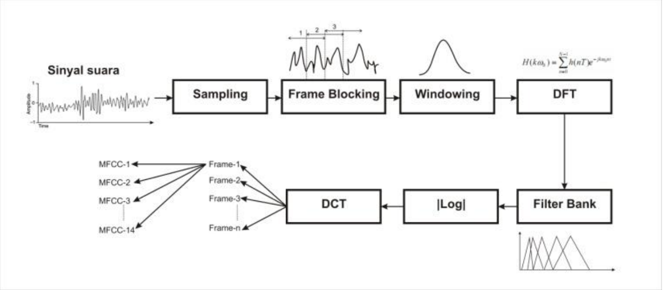

Figure 6. Digital Signal Processing for Extracting Cepstral Feature

The four signal features 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. After initialization, the next process is GMM training to obtain the maximum likelihood. After several iterations, if convergence is achieved the iteration can be stopped and all the parameter is stored into the database. Each recitation law in the course has one GMM model for correction as the reference template of recitation model in the application stage.

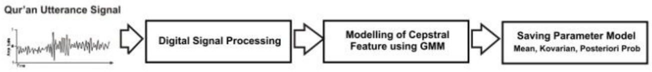
Qur'an Utterance Signal
Digital Signal Processing
Modelling of Cepstral Feature using GMM
Saving Parameter Model
Mean, Kovarian, Posteriori Prob

Figure 7. Process of Software Construction

At the stage of construction and application, correction process is done with similar process at the construction and formulation stage. However, 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. The error message is appeared in order to warn the user to reread the readings with proper and correct law of recitation.

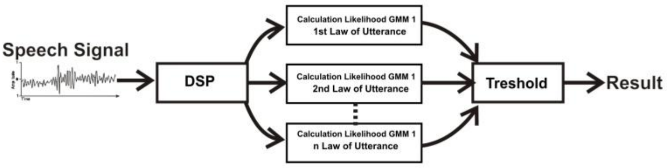

Figure 8. The Correction Process of Software

C. *Result of System Prototyping*

To test the reliability and accuracy of correction, calibration for each recitation law in the software is done with ten speakers to read the readings in wrong and right manner.

1) Analysis of pronunciation correction (Makhraj): Fundamental

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%.

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 poor (70%), hence, the system need 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.

V. 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.

REFERENCES

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

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

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

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

[5] 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

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

[7] 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

[8] 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

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

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

[11] 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

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