Malaysian Journal of Computer Science, Vol. 12 No. 1, June 1999, pp. 28-37

THEORY, METHODOLOGY AND IMPLEMENTATION OF THE MALAY TEXT-TO-SPEECH SYSTEM

Aini Hussain, Salina Abdul Samad and Kuek Teik Soon
Multimedia Signal Processing Research Group
Dept. of Electrical, Electronic & Systems Engineering
Faculty of Engineering
Universiti Kebangsaan Malaysia
43600 Bangi, MALAYSIA
email: aini@eng.ukm.my

ABSTRACT

This paper deals with the improvements and modifications done on the first edition of the Malay Language Text-to-Speech System, SUM (acronym for Sintesis Ucapan Melayu). A simple review on human speech production system and synthetic speech production system is discussed. Modifications on the voicing source include an additional KLGLOTT88 model added into the software. Theory and methodology on the voicing source are discussed in this paper. Characteristics of natural speech such as breathiness and flutter are studied. In addition, more constructive rules are added to improve the capability of the previous Malay TTS SUM system. The synthesiser model is based on the Klatt's formant synthesiser, KLSYN88.

Keywords: ***speech synthesis, text-to-speech, formant synthesiser & KLSYN88.***

1.0 INTRODUCTION

In most synthesis-by-rule systems, the synthesis model is based on a source-filter theory of speech production that was made explicit by Gunnar Fant (1960) [1]. This type of speech synthesis model is achieved by controlling the parameters of the sources and of the filters that shape the sources [2]. The model that is used in our SUM 1 and SUM 2 is the formant synthesiser. The formant synthesiser is well known as the model that is capable of producing the most natural-sounding and intelligible speech at present time. In our case, the synthesiser is used in text processing for text-to-speech (TTS) synthesis.

Automatic synthesis of speech from a general text requires two basic types of procedures: transforming the input text into a set of intermediate parameters with appropriate linguistic label, and the actual generation of a speech waveform based on the linguistic parameters [3]. The text processing phase of a TTS system must accept a standard text as input and produce as output a sequence of codes corresponding to phonemes plus sufficient prosodic information to specify the intensity, duration and fundamental frequency, i.e., the acoustic aspects of intonation. The standard text is composed of words and other symbols such as digits and abbreviations, along with punctuation marks.

In the effort of producing more natural synthesised speech, basic understanding in the mechanism of human speech production system is vital. Thus, we will discuss this field to a certain extent, such as the structure of the larynx, its mechanism in controlling the opening and closure of the glottis and the resulting effect in the speech, especially the breathy effect as in a whisper. In line with this, one model of voicing source is studied and added to our software, which is the KLGLOTT88 [4, 5]. The pulse shape of the glottal volume velocity, $U_g(t)$ produced by the KLGLOTT88 was proposed by Rosenberg (1971).

2.0 HUMAN SPEECH PRODUCTION SYSTEM

Before we carry on our discussion on the synthesis model that we use in the synthetic speech production, it is important for us to understand the mechanism in human speech production and some characteristics of speech signal. This is due to the fact that most of the speech synthesisers are based on human's speech production structure and system. Fig. 1 shows a cross section of human vocal tract. Most of the speech sounds are produced by pulmonic air stream that is expelled from the lungs through muscular action. The air stream then passes through the larynx. Fig. 2 shows the diagram of a larynx. The main components of the larynx are the thyroid cartilage and crycoid cartilage where the cartilage is a kind of firm elastic tissue. Vocal folds are located within these two cartilages.

This is shown clearly by the cross section diagram of the larynx in Fig. 3. The simplified diagram of Fig. 3 is shown in Fig. 4. During normal breathing, the vocal folds are held apart to allow air to flow through the glottis and the air stream is blocked when the vocal folds are brought together. Voicing fundamental period is controlled by the vibration of the vocal folds [4, 6]. A fundamental period can be divided into two portions, an open-glottis cycle and a closed-glottis cycle as mentioned above and it will be further illustrated in the following section.

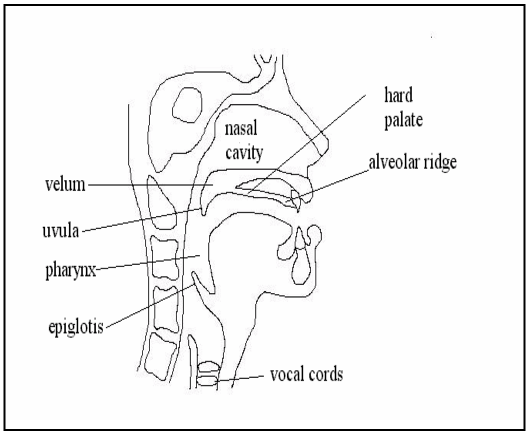

Fig. 1: Human vocal tract [1]

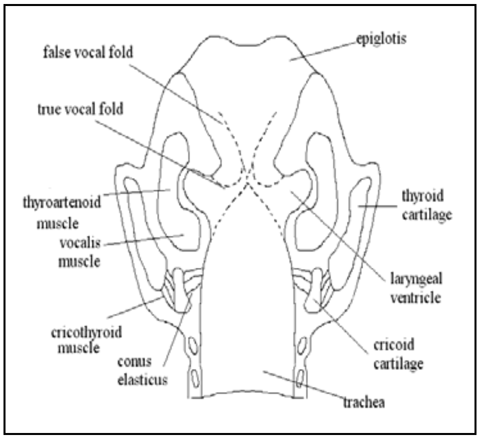

Fig. 2: Larynx [6]

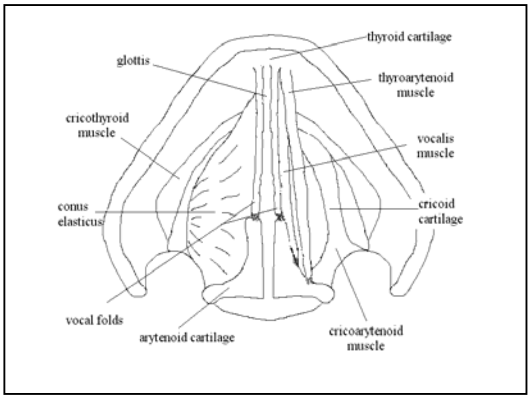

Fig. 3: Superior view of larynx [6]

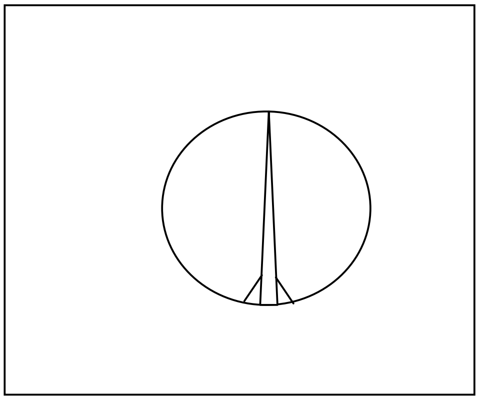

Fig. 4: Simplified diagram of larynx

2.1 Mechanism of Larynx in Speech Production

As shown in Fig. 4, the arytenoid cartilage controls the glottis size. By varying the distance between the two cartilages, we can produce various types of voicing such as creaky sound, breathy sound and so on. The three glottis configurations that are controlled by the arytenoid cartilages in the larynx are discussed in this section. As discussed earlier, the fundamental period is divided into an open-glottis cycle and a closed-glottis cycle. The duration of each cycle contributes to the characteristics of voicing, for example, for a breathy sound, the open-glottis cycle is prolonged so as to let more air stream expelled from the lungs at a longer period and results in a breathy effect.

The three glottis configurations with the glottal volume velocity waveform, Ug(t), are shown in Fig. 5 [4]. In Fig. 5, row (a) shows the 3 glottis configurations, row (b) shows the volume velocity waveform, Ug(t), and row (c) shows the spectrum of voicing source. The first column configuration shows a laryngealised voicing source. Note that the arytenoids are brought together and the glottis is closed.

This configuration produces creaky voice and the glottal volume velocity waveform is a relatively narrow glottal pulse. This indicates that the duration of the open portion of a fundamental period is substantially lowered during laryngealisation [4]. The characteristics of a modal voice are illustrated in the second column configuration. The vocal folds are nearly approximated with a slightly opened glottis. The volume velocity waveform has a longer duration of the open portion, and this allows more air flowing through the glottis to add some breathiness to the voice
.

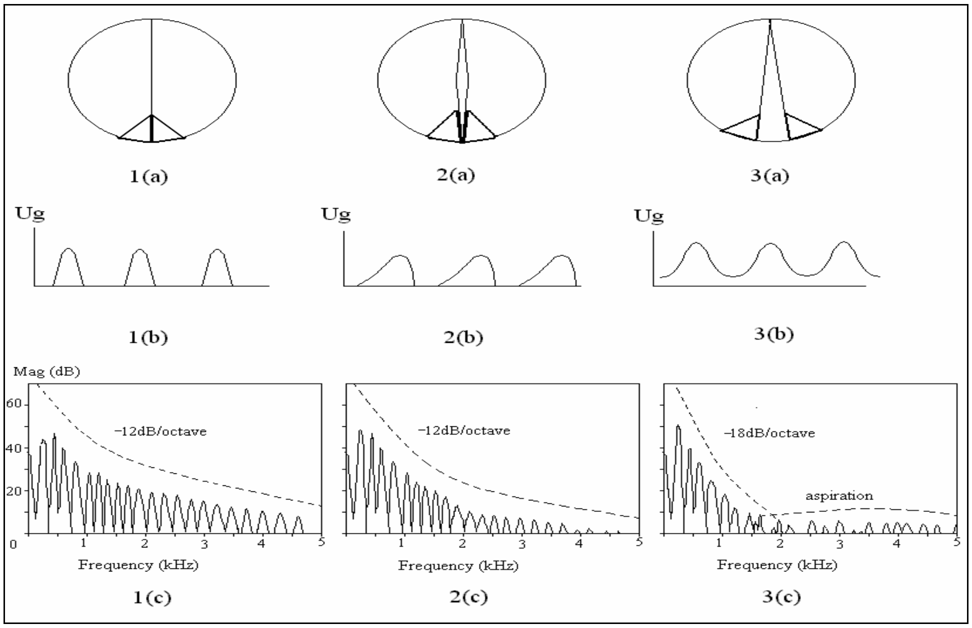

Fig. 5: The Three Glotties Configurations

The spectrum of the normal voicing source is shown in 2(c) with an average falloff of about -12dB/oct. The third column configuration shows a breathy glottal configuration. The arytenoid cartilages are separated at the back and causes large amount of airflow resulting in turbulent aspiration noise. Hence, the spectrum of the voicing source is a combination of harmonics and random noise. As we can see from left to right in row 3(c), the spectrum shows a weakening high frequency harmonic being replaced by aspiration noise due to the breathiness.

2.2 Speech Production System

2.2.1 Source Filter Model

Most of the speech synthesisers nowadays are based on a speech production system known as the source-filter model. The source-filter model view speech signal as being produced by sound sources exciting a linear system (the vocal tract) [6, 7, 8]. The four possible combinations of a sound source are:

(a) no source (silence)
(b) voiced source only
(c) mixed voice source and noise source
(d) noise source only

Fig. 6 shows the block diagram in a speech production process based on the source-filter model. The source consists of pulsed airflow through the glottis at the fundamental period. This can be expressed as a harmonic spectrum |S(f)| by the Fourier transformation of the time domain glottal waveform. The spectrum of vocal tract transfer function is represented by |T(f)| and the radiation characteristics of sound can be represented by |R(f)|.

The final speech signal is the speech pressure waveform |P(f)| which can be expressed by the following equation:

$$|P(f)| = |S(f)|\,|T(f)|\,|R(f)| \qquad (1)$$

The peaks in |T(f)| are referred to as formants [8]. Note that the formant structure in the final speech signal is a combination of the resonances of the vocal tract transfer function and the harmonic structure of the source.

2.2.2 KLGLOTT Voicing Source Model

In the effort to increase the naturalness of the synthesised speech, breathiness is introduced. KLGLOTT88 is one of the voicing source model that takes this into consideration. By introducing an open quotient in the fundamental period, breathiness is added in the voicing source (as discussed earlier)[4],[5]. Figure (7) shows the waveform of the KLGLOTT88. KLGLOTT88 is based on the pulse shape proposed by Rosenberg (1971), and it can be expressed by the equations below:

$$\begin{aligned} U_g(t) &= aT^2 - bT^3\ , \text{ for } 0 < T < O_qT_0; \\ U_g(t) &= 0 \qquad\quad\ , \text{ for } O_qT_0 < T < T_0; \end{aligned} \qquad (2)$$

in which a and b are constant depending on the voicing amplitude, AV and the length of open quotient O_qT_0 :

$$a = \frac{27AV}{4T_0{O_q}^2}\ , \qquad b = \frac{27AV}{4T_0^2O_q^3} \qquad (3)$$

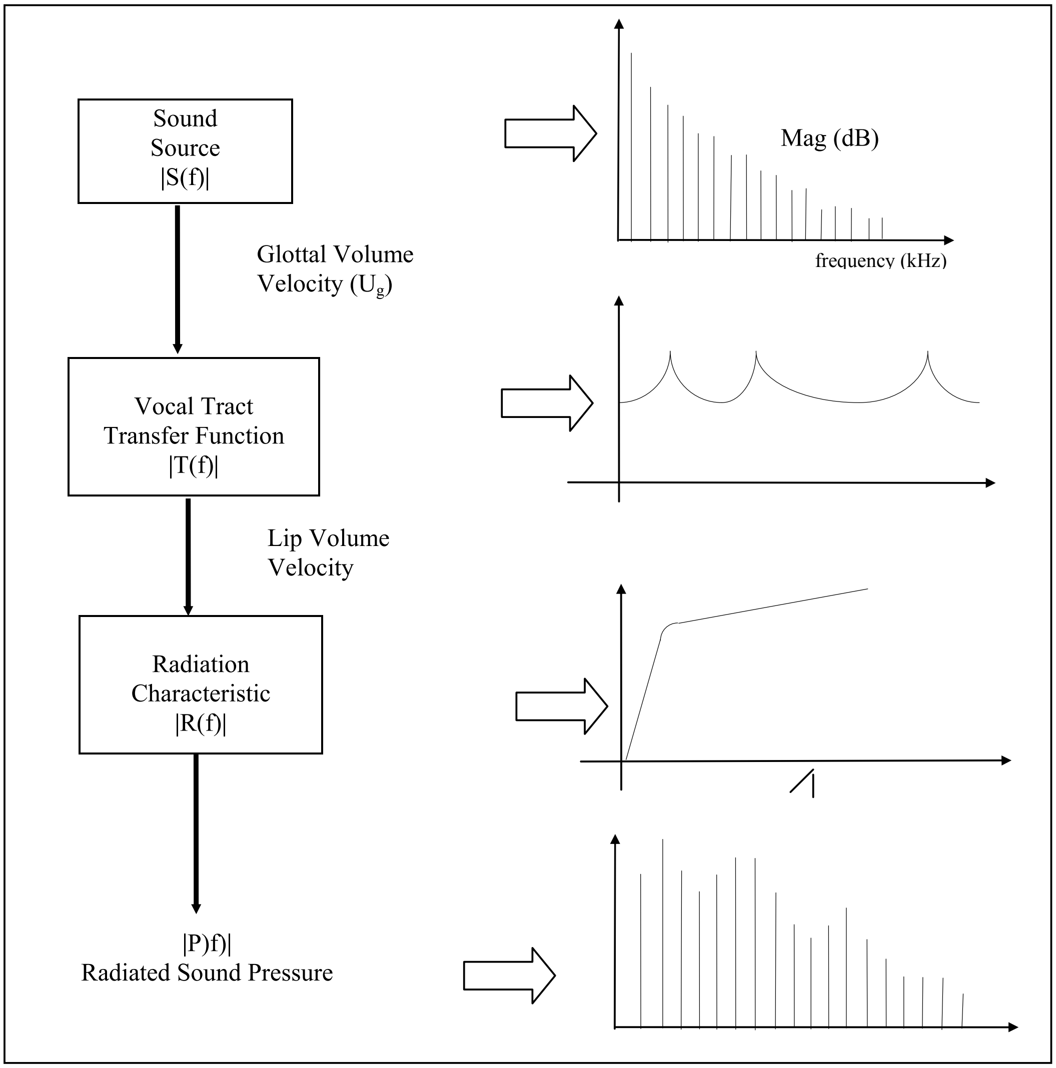

Fig. 6: Block diagram representing the source-filter theory of speech production [8]

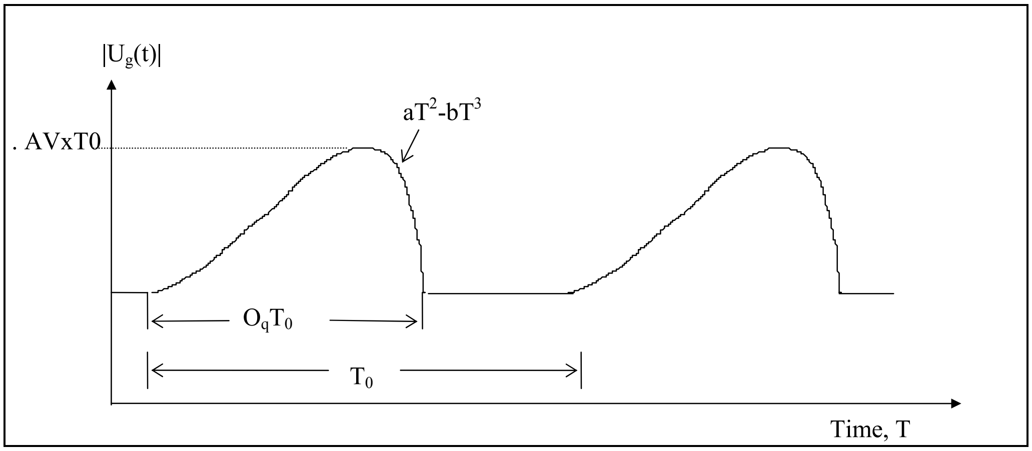

Fig. 7: KLGLOTT88 Voicing Source Mode

The spectrum of the KLGLOTT88 model is given by:

$$U_k(\upsilon)=\frac{27jAV}{2O_q(2\pi\upsilon)^2}\left[\frac{je^{-j2\pi\upsilon O_qT_o}}{2}+\frac{1+2e^{-j2\pi\upsilon O_qT_0}}{2\pi\upsilon O_qT_0}+3j\frac{1-e^{-j2\pi\upsilon O_qT_0}}{(2\pi\upsilon O_qT_0)^2}\right] \tag{4}$$

In our software, the sampling rate used is 11025Hz while the parameters updating rate is 5ms, in other words, parameters are being updated every 56 samples. Hence, we define 56 samples as 1 frame and the number of frames in one segment of phoneme depends on the duration of the phoneme. Fig. 8 shows the diagrammatic representation of the updating parameter methodology used in our software by using two types of voicing sources: impulse voicing source and KLGLOTT88 voicing source. As can be seen from Fig. 8, in every voicing cycle,T_0, voicing pulse is being generated and at the same time, parameters are being updated.

2.2.3 Fluttering Characteristics

It is well known that a constant fundamental frequency, F_0, is to be avoided in speech synthesis because the result is a peculiarly mechanical sound quality. Hence, another parameter, FL is suggested [4] and is added into our synthesiser. In SUM 1, the contour of the fundamental frequency is based on a straight base line [8, 9]. In the effort to achieve a more natural speech, we include a mechanism for introducing a slow quasi-random drift to the F0 contour through the FL control parameter, in fact, the quasi-period component is the sum of three slowly varying sine waves [4]:

ΔF_0 = (FL/50)(F_0/100)[sin(2π12.7t) + sin(2π7.1)t + sin(2π4.7t)] Hz. {5}

In our case, the smallest time range unit for the changes of F_0 is 1 frame (56 samples), in other word, ΔF_0 is added to the F_0 contour at the same instant when parameters are being updated. This is illustrated in Fig. 9.

2.2.4 Tracheal Coupling

According to D. Klatt, effects of tracheal coupling can be modelled in a formant synthesiser by adding one or more paired pole-zero resonators to the cascade model of the vocal tract transfer function. This addition of resonators and anti-resonators can improve the synthesis of breathy vowel [4]. D. Klatt suggested that tracheal resonances are often seen in breathy vowels at frequencies of about 550, 1300 and 2100Hz. Hence, we use three resonators (RTP1, RTP2, RTP3) and three anti-resonators (RTZ1, RTZ2, RTZ3) to resonate tracheal coupling frequencies. The variables BTP (bandwidth of the tracheal pole) and BTZ (bandwidth of the tracheal zero) have default values of 180 Hz [4].

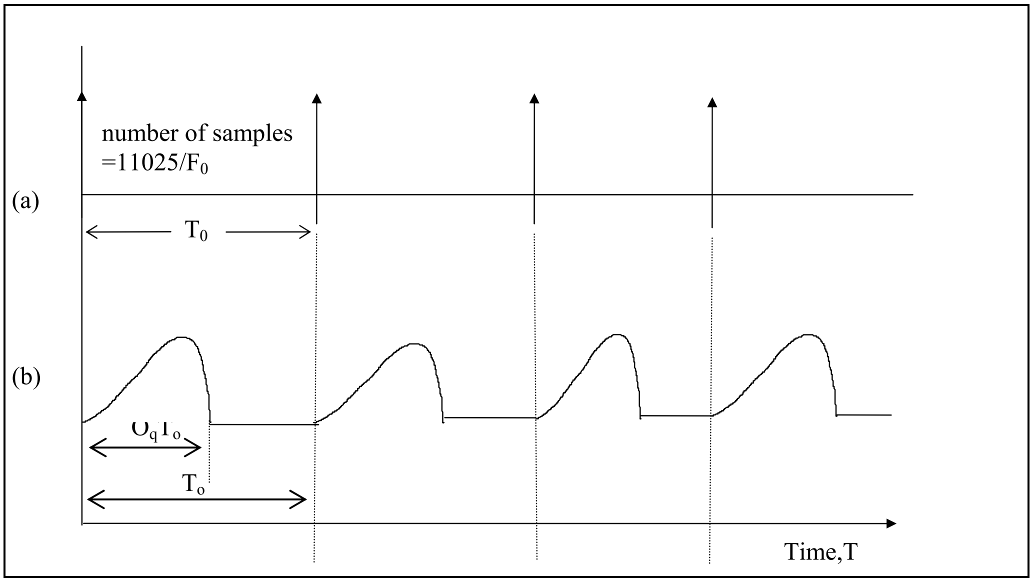

Fig. 8: Updating Parameters using (a) Impulse Voicing Source and (b) KLGLOTT88 Voicing Source

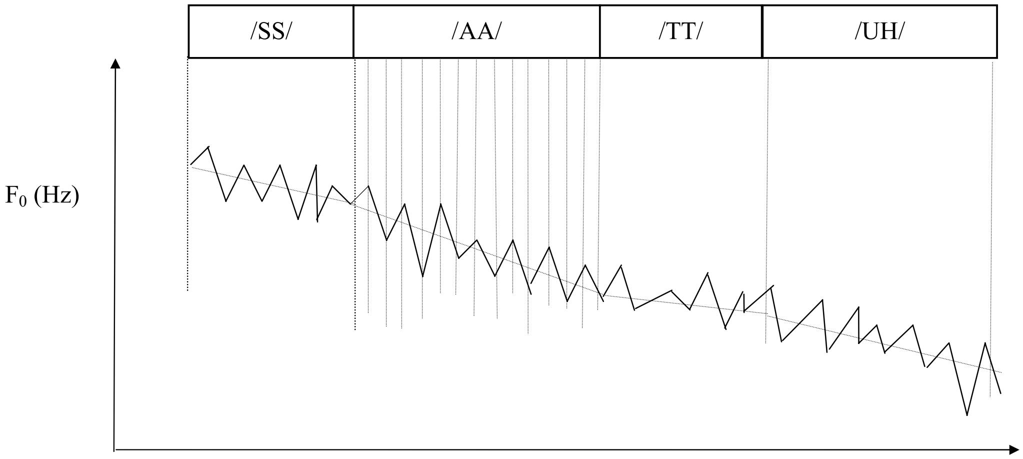

Fig. 9: Adding flutter into the fundamental frequency contour

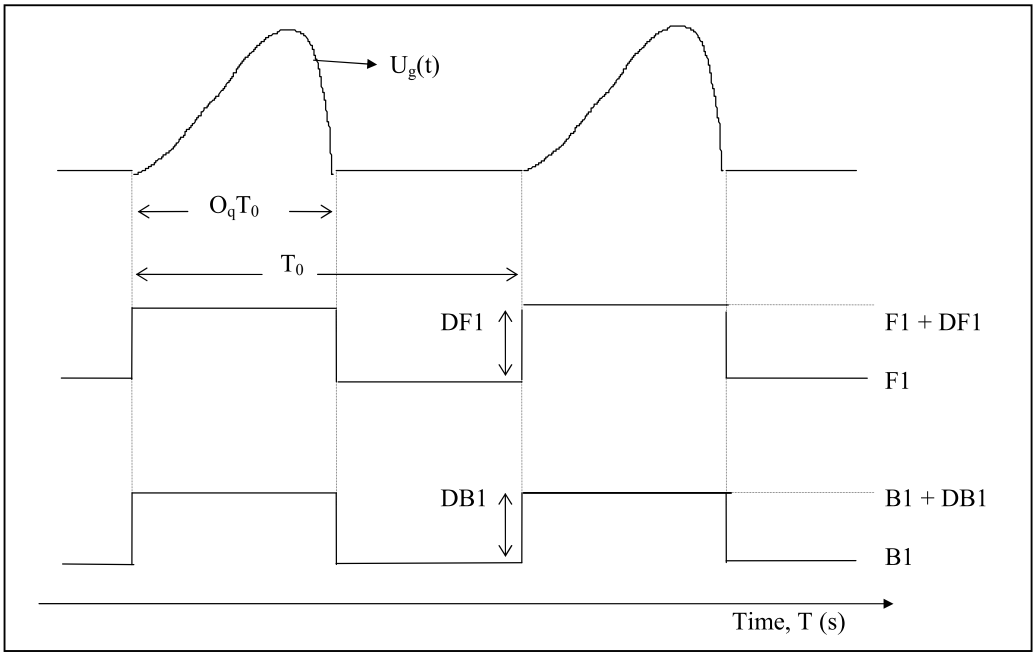

Fig. 10: Changes of F1 and B1 in "Square-Wave" Fashion during Open Quotient, O_qT_0

2.2.5 Variation in F1 and B1 During Open Quotient of Voicing

To synthesise voiced and breathy sounds, the first formant frequency F1 is increased by 10% during the open quotient of each voicing cycle, while the first formant bandwidth B1 is increased by 400 Hz [4]. F1 and B1 change abruptly at the instant of glottal opening and glottal closure, in other word, it occurs in a "square-wave" fashion. This is illustrated in the Fig. 10. This changes in the first formant frequency and first formant bandwidth also help in synthesising breathy sound.

3.0 ADDENDUM

A few rules have been added into our software SUM to improve its intelligibility and to add more features to our software. Currently, our software uses two types of voicing sources: the impulse source and the KLGLOTT88 voicing source. As KLGLOTT88 does consider the open quotient and close quotient of voicing cycle, thus it can be used to produce breathy sounds. Also included is improvement in the quality of breathy vowels such as the vowel /AA/ in the word "Hari". Due to the aspiration and breathy effects of the phoneme /HH/, the initial portion of the vowel /AA/ should also be added some breathy effect. Thus we use KLGLOTT88 voicing source in synthesising the initial portion of the breathy vowel while the remaining part of the vowel uses the impulse-voicing source.

SUM 2 is able to produce two types of voice features: the normal voice and the whispering effect. Besides, we also tried to improve the naturalness and the intelligibility of the synthesised speech by adding the fluttering effect (as in real speech). Other attempts to improve the capability and intelligibility of our software have also been done. Now, SUM 2 is able to read numbers based on the values, for example, 123 will be read as “Seratus dua puluh tiga” (one hundred twenty three) and not just “satu”, “dua”, “tiga” anymore. Rules have been set so that SUM 2 can read up to the number 999,999,999,999,999. However, more rules should be added in the future so as to improve the current capability of reading numbers, such as in reading fraction, date, currency and so on.

We also attempted to use the concept of tri-phone in our synthesiser. Currently, we have done it on vowels only; in which we divide vowels into three categories: the initial vowel, the middle vowel and the ending vowel according to the positions of the vowels in a word. We observed that the duration of each vowel varies according to their positions in the word; the middle vowel has the shortest duration because it is limited by phoneme prior to and after the middle vowel; while the ending vowel has the longest duration. Thus, rules have been set to achieve this concept.

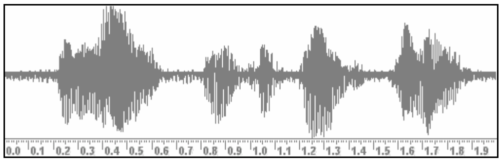

Fig.11 : Original sound (normal)

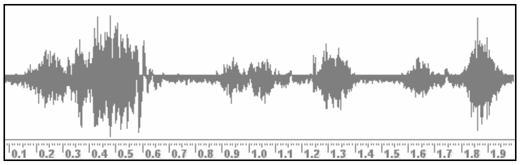

Fig. 12 : Original sound (whispering)

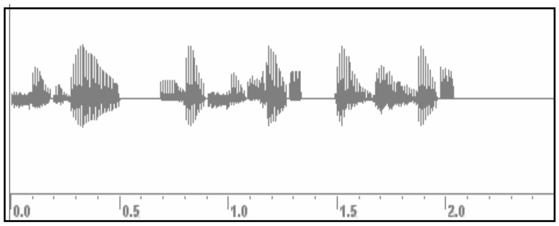

Fig. 13(a) : No breathiness added

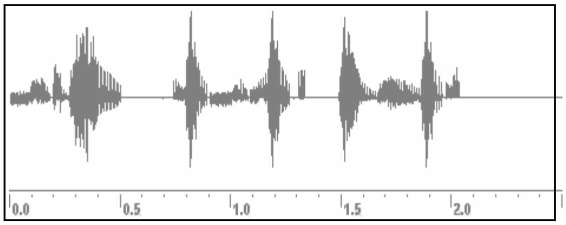

Fig. 13(d) : 50% breathiness added

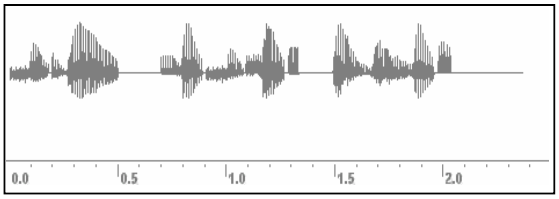

Fig. 13(b) : 10% breathiness added

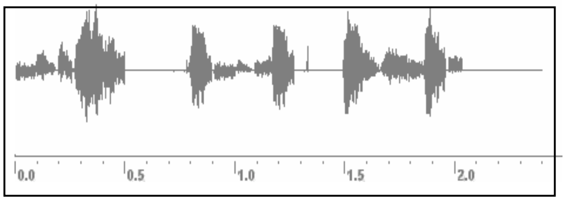

Fig. 13(e) : 90% breathiness added

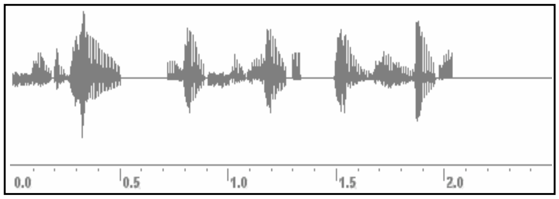

Fig. 13(c) : 30% breathiness added

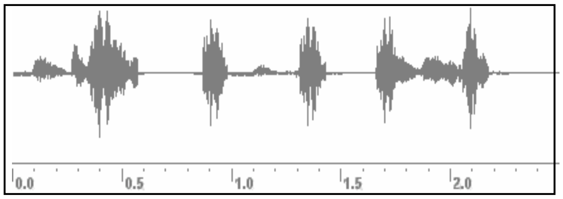

Fig. 13(f) : 100% breathiness added

4.0 RESULT

Here, we will show some of the synthesised speech waveform. First, we will see the effect of adding breathiness at different levels of the synthesised speech. Fig. 11 and Fig. 12 shown below are the original speech waveforms in which the former is a normal voicing effect, while the latter is a whispering effect, and the underlying sentence is “Sila masukkan ayat”.

Fig. 13(a) shows the synthesised waveform of the same sentence with no breathiness added. Fig. 13(b) shows its PSD spectrum, while Fig. 14 (a, b, c, d, e & f) show the synthesised waveforms with different levels of breathiness being added, and Fig. 15(a),(b),(c),(d),(e) &(f) show their PSD spectrum respectively. Note that when we add the breathiness in order to get the whispering effect, we are actually adding some noises into the speech, and these noises are in the range of higher frequencies. In Fig. 14(a)-14(f); as more breathiness is being added to the speech, the PSD spectrum tends to be flattened at higher frequencies. Fig. 15, Fig. 16 and Fig. 17 show the spectrograms of the synthesised speech with different levels of breathiness.

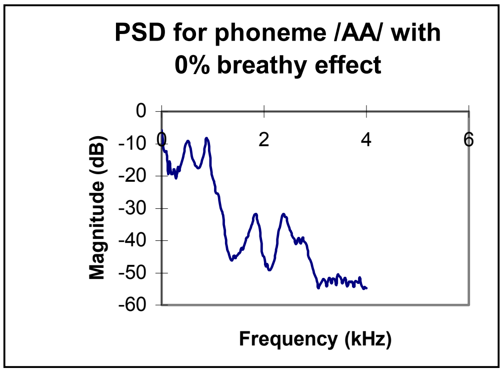

Fig. 14(a)

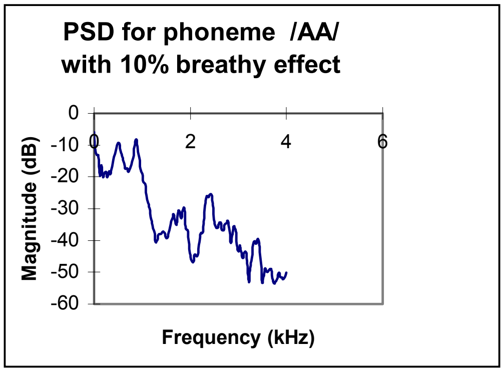

Fig. 14(b)

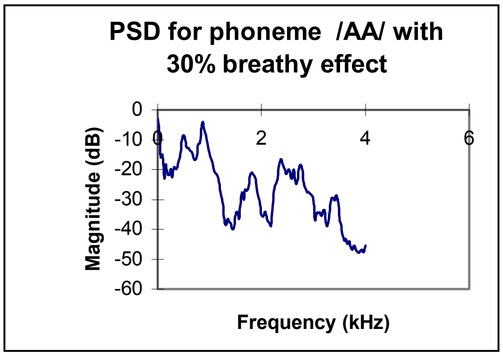

Fig. 14(c)

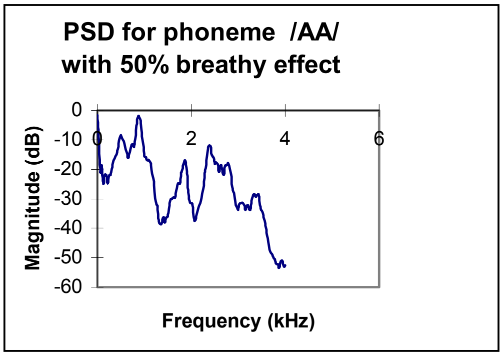

Fig. 14(d)

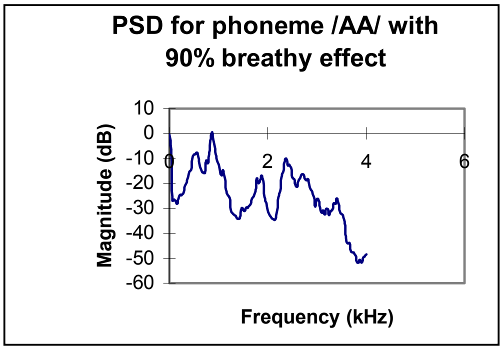

Fig. 14(e)

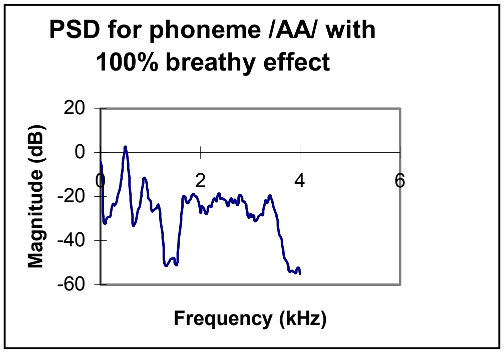

Fig. 14(f)

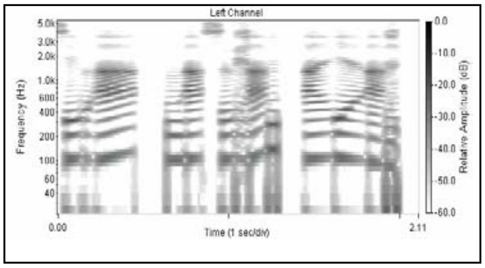

Fig. 15 : No breathiness added

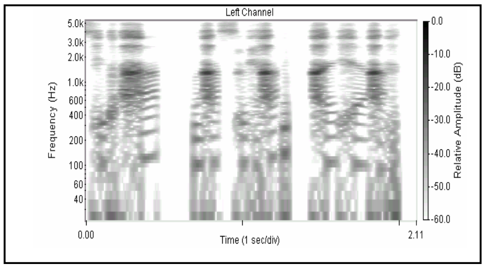

Fig. 16 : 50% breathiness added

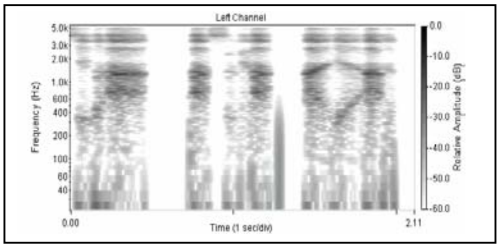

Fig. 17 : 90% Breathiness Added

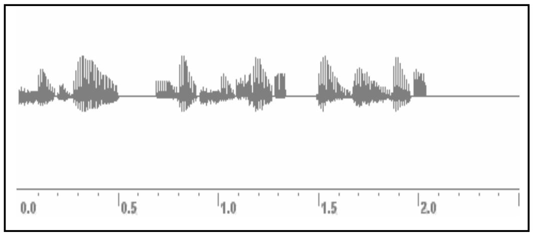

Fig. 18 (a) : No Flutter Added as in S.U.M.1

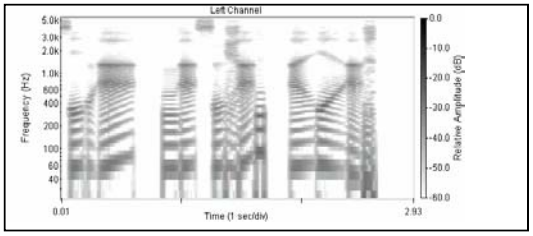

Fig. 19 (a) : No Flutter Added as in S.U.M.1

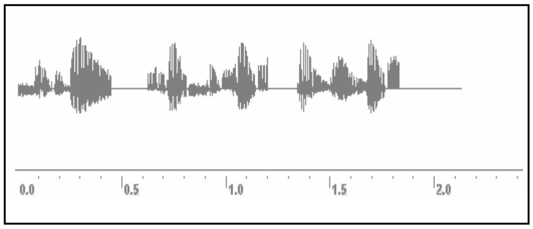

Fig. 18 (b) : 50% Flutter Added

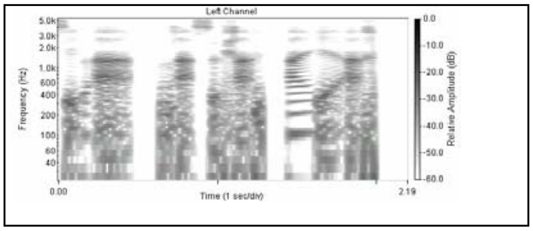

Fig. 19 (b) : 50% Flutter Added

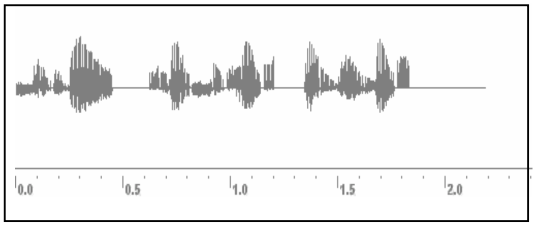

Fig. 18(c) : 90% Flutter Added

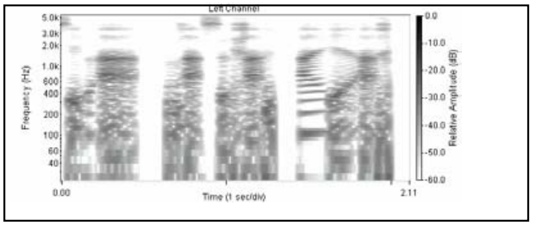

Fig. 19 (c) : 90% Flutter Added

As we can see from the spectrograms, the energy intensity for the higher range of frequencies increases as more breathiness is being added, and this agrees with what we had concluded that the noises that are being added are in the higher range of frequencies. Next, we will see the effect of adding different level of flutter in the speech, as we can see in Fig. 18, Fig. 19 and Fig. 20. In S.U.M. 2, the levels of flutter range from 0 to 9. As mentioned earlier, the main purpose of adding flutter is in an attempt to imitate the real characteristics of natural voice in human beings. The synthesised speech waveforms with flutter effect are more similar to the natural speech waveform as shown in Fig. 11 and look more natural when compared with the S.U.M.1's synthesised speech waveform.

5.0 CONCLUSION

There is still some shortcoming of this software. The synthesised speech lacks in naturalness and this is the main concern. However, with more constructive rules being added, the software is more capable now when compared to SUM 1. The study that has been conducted on the voicing source, flutter effect and breathiness is a very important step in achieving higher quality TTS system. The flexibility of the structure of the software allows future improvements and modifications since there are still untapped potential to be discovered.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, "Digital Processing of Speech Signals" Prentice Hall (1978).

[2] I. H. Witten, "Principles of Computer Speech" Academic Press,Inc. (1982).

[3] D. O'Shaughnessy, "Spectral Transitions in Rule-Based and Diphone Synthesis". Talking Machines: Theories, Models and Designs (1992).

[4] D. H. Klatt and L. C. Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers" J. Acoustic. Soc. Am. 87(2), February 1990.

[5] B. Doval and C. d'Alessandro, "Spectral Correlates of Glottal Waveform Models: An Analytic Study" 0-8186-7919/97 IEEE (1997).

[6] A. Breen, "Speech Synthesis Models: A review", Electronics and Communication Engineering Journal February 1992.

[7] H. Javkin, K.Hata, L.Mendes, "A Multi-Lingual Text-to-Speech System" CH2673-2/89/0000-0242 IEEE (1989).

[8] J. Allen, M. S. Hunnicutt and D. H. Klatt, "From Text-to-Speech~The MITalk System". Cambridge University Press (1987).

[9] A. Hussain, S. A. Samad and H. J. Haur "The Malay Text-To-Speech System". TELEKOM Journal. (To be published in June 1999).

[10] K. N. Stevens, "Speech Synthesis Methods: Komage to Dennis Klatt". Talking Machines: Theories, Models and Designs (1992).

BIOGRAPHY

Aini Hussain received her B. Sc., M. Sc. and Ph.D. degrees in Electrical Engineering from Louisiana State University, USA, University of Manchester Institute of Science and Technology, UK and University Kebangsaan Malaysia in 1985, 1989 and 1997, respectively. Her research interests include signal processing and neural networks.

Salina Abdul Samad obtained her Bachelor of Science in Electrical Engineering from the University of Tennessee, USA and a Ph.D. from the University of Nottingham, England. Her research interest is in the field of digital signal processing, from algorithm design to software and hardware implementation.

Kuek Teik Soon received his Bachelor of Science degree in Electrical Engineering from the Universiti Kebangsaan Malaysia in 1997. He currently works in Intel Corporation as a design engineer.