How to cite this paper:

Nurmaisara Za'ba & Nursuriati Jamil. (2017). Speech to singing synthesis: incorporating patah lagu in the fundamental frequency control model for malay asli song in Zulikha, J. & N. H. Zakaria (Eds.), Proceedings of the 6th International Conference of Computing & Informatics (pp 268-275). Sintok: School of Computing.

SPEECH TO SINGING SYNTHESIS: INCORPORATING PATAH LAGU IN THE FUNDAMENTAL FREQUENCY CONTROL MODEL FOR MALAY ASLI SONG

Nurmaisara Za'ba, Nursuriati Jamil

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40430 Shah Alam, Selangor, Malaysia, nurmaisara_zaba@yahoo.com, *lizajamil@commputer.org*

ABSTRACT. Singing traditional Malay asli music requires a type of ornamentation called patah lagu although it is often not indicated in the musical score. Therefore, new singers often learn through listening to previous performance and depend on rote learning and memory. In this paper, we introduced a new patah lagu contour to be incorporated in a speech to singing synthesis model of Malay traditional asli music. It works by modifying duration and fundamental frequency (F0) contour of input Malay spoken lyrics and convert it into singing voice based on musical notes, length and values from the score. A new contour called patah lagu based on original singing sample is incorporated in the F0 control model to accommodate Malay asli singing synthesis voice. Based on multivariate repeated measures, patah lagu makes the singing voice in the song sounds more natural as compared to without.

Keywords: singing synthesis, fundamental frequency, ornamentation, patah lagu

INTRODUCTION

Singing a traditional song is challenging without former knowledge and understanding of the song. Notated ornaments indication in a song usually varies, thus singers are expected to know how to embellish and alter the rhythms in appropriate places. Even though the indication and ornamentation practice may have been performed differently, its presence is very important in Malay *asli* music. *Asli* music is a famous syncretic Malay traditional music in Malaysia. It is normally played in slow tempo at 60 BPM, and the singing is often accompanied by traditional musical instruments. *Asli* music singing involves traditional melodic ornamentation called *patah lagu. Patah lagu* is a form of ornamentation that is very important in *asli* music. Without *patah lagu*, an *asli* song will sound dreary and dull (Nasuruddin, 2007). However, proper documentation on the ornamentation part in performing *asli* music ensembles is non-existence. This is because *asli* music was created based on play and memorize by native musician according to feel and taste. The music was then rewritten by later musicians but focusing mostly on the melody of the songs. Thus today, variation in *patah lagu* rendition by different performers on a same song can be observed. With these limitations, new singers who wish to learn a*sli* music now and in the next generation have to look for practical instruction and guidance from Malay *asli* connoisseurs. This is especially crucial as the number of Malay *asli* music connoisseurs has dropped to almost none.

The main goal of this paper is to document the *patah lagu* contour in Malay *asli* music and to incorporate it into a speech to singing synthesis system of a Malay traditional *asli* music called *Seri Mersing*. Section 2 presents related work on singing voice characteristics in speech to singing synthesis. Singing voice and *patah lagu* data collection is described in Section 3, followed by our proposed Malay speech to singing synthesis in Section 4. Section 5 describes the survey results on user perception towards the naturalness of singing voices, with and without *patah lagu*. Conclusions are given in the last section.

RELATED WORKS

Singing voice is known to have more dynamic characteristics than speaking voice and the vowels are articulated over a wider range of pitch frequencies. A singer normally expresses song lyrics by sustaining vowels and changing musical notes according to a melody in the score, which produces pitch contour. However, initial study on singing voice and its characteristics is essential to construct natural synthesize singing voices (Saitou et al., 2005a; 2005b). Akagi et al. (2000) introduced the importance of fluctuations in pitch contours by analyzing fundamental frequency (F0) fluctuations in singing voices and sustained vowels. The important types of fluctuations related to singing voice characteristics as described by Saitou et al. (2005b) are vibrato (Ferrante, 2011), overshoot, and fine fluctuations. Overshoot refers to F0 transition that exceeds the target note just after a note change (Saitou et al., 2005a). Vibrato is the quasi-periodic frequency modulation between 5 to 8 Hz of F0 observed in trained singing voices (Scherer et al., 2008). In *asli* music, vibrato cycle has the average rate of around 4 to 6 Hz (Za'ba et al., 2011). Meanwhile, preparation is F0 deflection in the opposite direction of a note change observed just before the note changes. Finally, fine fluctuation is the irregular fine fluctuations related to modulation frequency and modulation amplitude. In this paper, we conclude that all these dynamic characteristics are important aspects to be included in the synthesis system.

Patah lagu is a unique type of ornamentation that is turns, mellismetic type of rendering, embellishing the longer notes with vibrato, with new ideas and accent at the end of a phrase or cadence (Nasuruddin, 2007). The presence of *patah lagu* in *asli* music is essential as it carries identity to the music. Discussion on vibrato and *patah lagu* for traditional Malay *asli* singing synthesis is discussed in detail by Za'ba et al. (2011). The presence of ornamentation is not only important to western music but to all Indian musics. Arora et al. (2006) investigated vocal performance in Indian Classical Raga using PRAAT speech processing tool, by introducing Indian ornamentation such as *gamak* and *meend*. Thus, in this paper, we intend to introduce patah lagu and include its pitch contour into the speech to singing synthesis.

Shih and Kochanski (2001) described a style of speech and singing as a set of localized prosodic features and a set of rules. The styles were described as Stem ML-tags (Soft TEMplate markup language), to control accent shapes, pitch contour and amplitude profiles. Analysing *patah lagu* is required to obtain and describe rules based on prosodic features such as its pitch contour, duration and amplitude.

Based on the above discussion, this section has described the dynamic characteristics aspects of a singing synthesis system, and some introduction to *patah lagu* for a Malay *asli* singing synthesis system. This study is relevant to proof the importance of the *patah lagu* and report the performance of the synthesized *patah lagu* in the Malay speech to singing synthesis system.

THE ORNAMENTATION (*PATAH LAGU*) OF MALAY *ASLI* SONG

All *asli* music should be sung with *patah lagu*, even though the description of the *patah lagu* is not properly written in the score. In this paper, a popular Malay traditional *asli* song called *Seri Mersing* is chosen to be synthesized. Early observation on *Seri Mersing* music score shows no indication of *patah lagu* although it exists in singing performance. This shows that the singers play an important role in improvising *Seri Mersing*. In Indian classical singing, ornamentation is so important to the aesthetics of the genre. An Objective assessment method to distinguish ornamentation rendered by well-trained from an amateur singer based on pitch contour shape was designed to assist judgment process (Gupta et al., 2011). Based on these, an experiment to reveal pitch changes in *patah lagu* is conducted through F0 extraction to confirm its existence in *Seri Mersing.*

Singing Voice Data Collection

Four *asli* music singers comprising two females and two males were gathered to sing *Seri Mersing* in a studio. Two of them are popular *asli* music veteran singers who also teach vocal in *asli* music. The remaining two singers are experienced young singers who had received proper vocal training in *asli* music. The singing voices were recorded in a soundproof studio room at a sampling rate of 48 kHz with 24-bit resolution. During the recording, the singers listened to a minus one of *Seri Mersing* through a headphone, and sing to a microphone and a pop filter. In this experiment, only singing voice data samples are recorded. From the recorded continuous singing voice data samples, all long sustain syllables observed in long musical notes are manually segmented, labeled and processed to extract and reveal *patah lagu* in pitch contour.

Patah Lagu Contour

Based on F0 extraction experimentation using TEMPO in STRAIGHT (Kawahara et al., 1999), it is confirmed that all of the segmented sustain vowels contain *patah lagu.* This is confirmed by a music expert and educator from University Technology of MARA (UiTM). The labeled parts in the singing voice with *patah lagu* were manually segmented using *Audacity* (version 2.0.6). There are 59 labeled *patah lagu* for each singing voice signal with a total of 236 *patah lagu* for all 4 singing voice signals. These *patah lagu* are observed to appear on almost all sustain crochet, minim and semibreve notes in each of the singing voice samples. It shows an average of 79% use of *patah lagu* in singing *Seri Mersing* on every central or long notes. Some of the *patah lagu* are accompanied with natural vibrato around 4 Hz to 5 Hz, and the power changes are synchronized with pitch changes. The contour of the *patah lagu* is also dissimilar in pitch and time/duration by different singer. The experiment showed that female singers appear to have higher pitch range, 300 to 400 Hz, while the males range around 200 to 300 Hz. The pitch contour reveals large, ragged and irregular differences in the length of the successive sound waves and duration, mostly due to age factor. This experimentation is discussed and reported in detail in (Za'ba et al., 2014).

MALAY SPEECH TO SINGING SYNTHESIS

The *patah lagu* contour acquired earlier is further used in the F0 control model of the speech to singing synthesis framework. Speech to singing synthesis is a process that enables users to listen to their singing voice simply by reading from a song lyric. The singing synthesis framework is consists of two models controlling acoustic features unique to singing voices based on F0 and syllable duration as shown in Figure 1.

Speech and Singing Input Data

Speech to singing synthesis requires input of end user's spoken words from the lyrics of *Seri Mersing* and information such as music notes and beat value from the musical score to convert the speech into singing voice. In this paper, the input data consist of recorded voice data of two types: spoken and singing voices. The spoken voice are recorded from four speakers reading the given song lyrics in a neutral reading style. The speaking voices were digitized at 16 bit/48 kHz. They are then manually annotated into five kinds of labels; vowel, consonant, boundary (transition region from vowel to consonant or vice versa), syllable, and silence. The syllables (e.g. 'se', 'ri') are manually segmented and then lengthened or shortened in the duration control model according to its musical value. Finally, they are stored in a data bank as an input data for the speech to singing synthesis system.

The same four singers were also asked to sing *Seri Mersing* using the same lyrics with 48-kz sampling and 16-bit accuracy. The singers are requested to perform the song with their own individual *patah lagu* improvising style. They listened to the melody of *Seri Mersing* from a headphone and sang the song using the same lyrics. Musical information of the song is synchronized with the syllables in the lyrics of *Seri Mersing*. Music tempo for *Seri Mersing* is 60 beat per minute (BPM). All musical note values are converted into time in seconds.

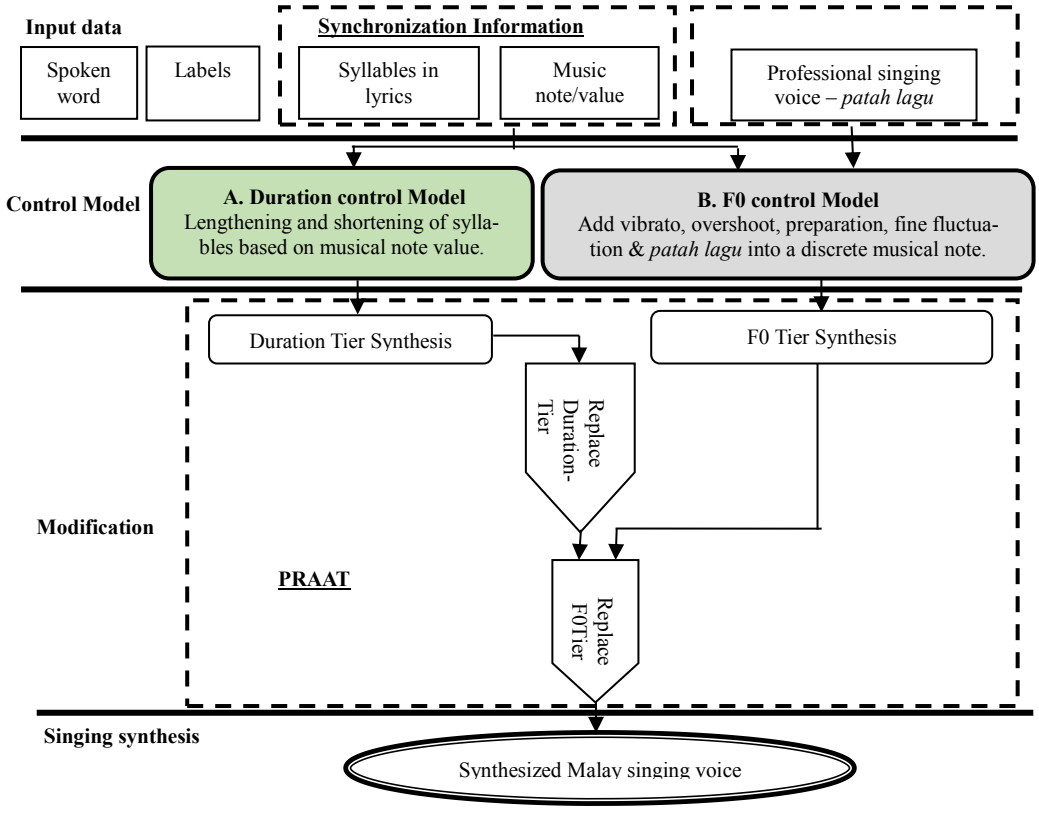

Figure 1. Malay Speech to Singing Synthesis Block Diagram.

Duration Control Model

The duration of syllables in the speaking voice differs from the singing voice. The duration of each syllable is determined by the kind of musical note (e.g., crotchet or quaver) and the given tempo. In this paper, the duration of one beat in musical tempo is defined as 60 beat per

minute. A MATLAB code reads the musical composition and extracts the information as by how much time-fraction in the unit of beats each syllable needs to be extended, and write it to a text file. This text file is then read by PRAAT (Boersma & Weenik, 2010) script to generate a Duration Tier to elongate the duration of syllables. For extending the length of syllable, the vowel part is elongated, keeping the other consonant and boundary, unchanged in duration. The silences are removed.

F0 Control Model

Singing voice contour is generated in this model. Speaking voice is converted into a singing voice by removing the F0 contour of the speaking voice and replacing the target F0 contour of the singing voice generated by F0 control model (Lai & Liang, 2010). The singing voice F0 contour is controlled in terms of global and local F0 variations. Global F0 variations correspond to the sequence of musical notes and local F0 variations include F0 fluctuations. An F0 control model (see Figure 2) is used to generate continuous F0 contour of singing voice from discrete musical information by adding vibrato, overshoot, preparation, fine fluctuation and a new F0, *patah lagu*. In this experiment, a new F0 contour is synthesized which is discrete in frequency having step transitions from one frequency to another. The tonic frequency is defined by the user. To this F0 contour, a vibrato as seen in Eq. (1) is added by using a simple oscillator.

$$pc_{vib}[n] = pc_{discrete}[n] + A\sin(2\pi f_{vib} n) \quad (1)$$

where, *fvib* and *A* are user defined.

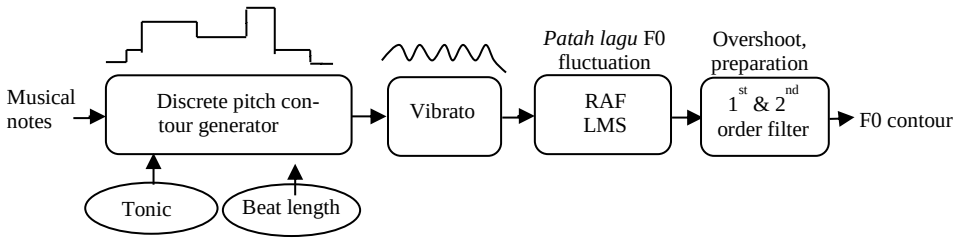

Figure 2. Fundamental Frequency Control Model.

Patah lagu contour that is initially extracted from a few samples of professional singing voices are filtered and simulated using Recursive Adaptive filter (RAF) and Least Mean Square (LMS) algorithm. The filter input is the step transition of pitch or F0 contour (discrete form) from one note to another and the output is smooth transition as in Malay *asli* Music. The desired output is taken as F0 contour of *patah lagu* from recorded singing voices. The filter model is as an all pole system as in Eq. (2). Meanwhile, the transfer function is calculated as shown in Eq. (3).

$$yk = a_0 x_k + b_{1\,yk-1} + b_{2\,yk-2} \quad (2)$$

$$H(z) = \frac{Y(z)}{X(z)} = \frac{a_0}{1 - b_1 z^{-1} - b_2 z^{-2}} \quad (3)$$

The output is then shifted into the appropriate frequency range based on the predefined discrete value by axis shifting. This process is done to avoid “off key” or “singing out of tune” in the synthesized voice. The output or *patah lagu* fluctuation is then added into the F0 contour by replacing discrete pitch contour form.

The simulated F0 contour is then added into F0 control model and then is passed through a fixed parameter filter. The purpose of this filtering is to smoothen and generate overshoot and preparation in the F0 contour. Description of the fixed parameter using first and second order filter is as follows. A continuous time LTI filter was designed and discretized using Bilinear (Tustin) transform. This transformation uses the approximation as in Eq. (4):

$$z = e^{sT_s} \approx \frac{1 + sT_s/2}{1 - sT_s/2} \tag{4}$$

The discrete time transfer function *Hd*(*z*), in terms of continuous time transfer function *H*(*s*) is shown in Eq. (5).

$$H_d(z) = H(s'), \text{ where } s' = \frac{2}{T_s}\frac{z-1}{z+1} \tag{5}$$

Eq. (6) is used for first order filter,

$$H(s) = \frac{1}{\tau s+1} \tag{6}$$

where, τ is the time constant and system gain is 1. Meanwhile, for second order filter, Eq. 7 is utilized:

$$H(s) = \frac{\omega_n^2}{s^2 + 2\delta\omega_n s + \omega_n^2} \tag{7}$$

where, ω_n is the undamped natural frequency, δ is the damping ratio and system gain is 1. The duration and F0 of the input speech is altered and replaced in accordance with the duration and F0 contour from duration control model and F0 control model appropriately. The final Malay singing voice is generated using Pitch-Synchronous Overlap Add (PSOLA) method.

USER PERCEPTION ON THE NATURALNESS OF THE SINGING VOICES.

A survey to evaluate user's perception towards the naturalness of the synthesized singing voices, with and without *patah lagu* was conducted. Twenty-five graduate students with normal hearing ability, listen to paired stimuli through a binaural headphone at a comfortable sound pressure level and rated the naturalness of the synthesized singing voices on a seven-step scale from "-3" to "+3". Paired stimuli having either female or male voices are presented to each subject. The survey focused on the use of vibrato and *patah lagu* is analyzed using both descriptive and inferential statistics. Nine phrases labelled B1 until B9 comprising 2 to 3 words per phrase of *Seri Mersing* song are used in evaluating the naturalness.

Figure 3 illustrates the naturalness of the nine phrases when *patah lagu* is included. Based on Figure 4, each phrase is getting more natural with the incorporation of *patah lagu* especially for phrase B1, B2 and B8. For other phrases, although the inclusion of *patah lagu* made the phrases slightly more natural as compared to without the *patah lagu*, the difference is only minimum. In order to further confirm the difference, multivariate repeated measures is used for evaluation. Significance level of 10% is used since the sample size is small (n = 25).

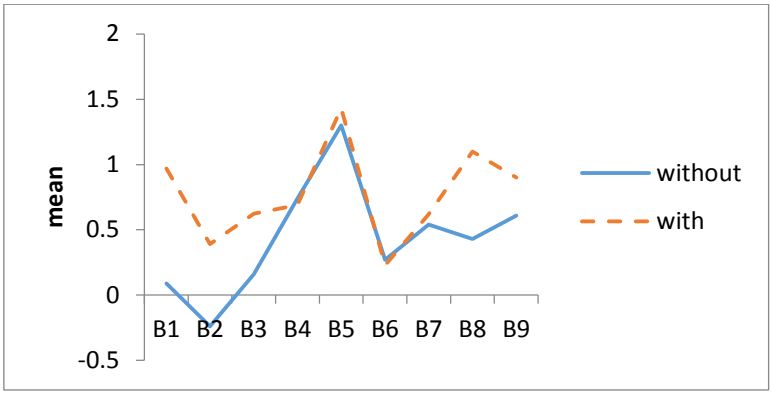

Figure 3. Profile Plot for *Seri Mersing* Phrases With and Without *Patah Lagu*.

Based on multivariate repeated measures, the result indicates that adding *patah lagu* into the F0 control model did make the singing voice in the song sounds more natural as compared to without, T^2 (9, 16) = 31.88, p = .064. However, not all sustain phoneme needs the addition of *patah lagu* to make it more natural. Result from Bonferroni confidence interval as shown in Table 1, points out that only three phrases (B1, B2 and B8) had significant difference in naturalness when *patah lagu* is added since 0 is not included in the interval. The other six phrases did not have any significant differences.

Table 1. Bonferroni Confidence Interval.

Sample-size	Variables	T2	F	df1	df2	p-value
25	9	31.8758	2.3612	9	16	0.0642

Mean vectors results are significant. The Bonferroni confidence interval B1 is: (0.19, 1.57). The Bonferroni confidence interval B2 is: (0.01, 1.25). The Bonferroni confidence interval B3 is: (-0.09, 1.01). The Bonferroni confidence interval B4 is: (-0.83, 0.73). The Bonferroni confidence interval B5 is: (-0.34, 0.60). The Bonferroni confidence interval B6 is: (-0.80, 0.72). The Bonferroni confidence interval B7 is: (-0.58, 0.74). The Bonferroni confidence interval B8 is: (0.27, 1.07). The Bonferroni confidence interval B9 is: (-0.39, 0.97).

CONCLUSION

This paper described *patah lagu* and its importance in Malay *asli* music. It also introduced a new fluctuation called *patah lagu* into a Malay speech to singing synthesis system. The proposed Malay speech to singing synthesis can convert spoken voice into singing voices by adding acoustic features unique to singing voices into F0 contour and by lengthening or shortening the duration of each syllables. Based on the study, we found that *patah lagu* is an important type of ornamentation or without it an *asli* music will sound dreary and dull. The evaluation results revealed that the proposed work is capable of converting Malay spoken voices into singing voices and it is proven that adding *patah lagu* into the F0 control model makes the singing voice in the song sounds more natural as compared to without. However, not all sustain syllable needs the addition of *patah lagu* to make it more natural. This study brings benefit to *asli* music industry and also to musician to record and promote *asli* music singing for younger generation using technology. The contribution is on Malay speech to singing synthesis system which demonstrates the potential of constructing more application in singing synthesis particulalrly on the perception and production of ornamentation or *patah lagu* in the synthesized singing voices. This work, however, focuses only on two types of

control models; F0 and duration. The synthesized singing voice should sound more natural if spectral charateristics and control is added into it.

REFERENCES

Akagi, M., & Kitakaze, H. (2000). Perception of synthesized singing voices with fine fluctuations in their fundamental frequency contours. *Proceedings of INTERSPEECH*, 458-461.

Arora, V., Behera, L., & Sircar, P. (2009). Singing voice synthesis for Indian Classical Raga system. *Proceedings of Signals and Systems Conference (ISSC 2009)*, IET Irish, 1-6.

Boersma, P., & Weenink, D. (2010). Praat: doing phonetics by computer [Computer program], Version 5.1. 44.

Ferrante, I.: Vibrato rate and extent in soprano voice. (2011). A survey on one century of singing. *The Journal of the Acoustical Society of America*, 130(3), 1683-1688.

Gupta, C., & Rao, P. (2011). Objective Assessment of Ornamentation in Indian Classical Singing. *8th International Symposium, CMMR.*

Kawahara, H., Masuda-Katsuse, I., & De Cheveigné. (1999). A. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, 27(3), 187-207.

Lai, W. H., & Liang, S. F. (2010). (2010). An F0 Control Model for Singing Synthesis based on Proportional-Integral-Derivative controller. *Proceedings of the 10th IEEE International Symposium on Signal Processing and Information Technology*, 182-185.

Nasuruddin, M.G.: Traditional Malaysian Music. (2007). Dewan Bahasa dan Pustaka, Kuala Lumpur.

Saitou, T., Unoki, M., & Akagi, M. (2005). Development of an F0 control model based on F0 dynamic characteristics for singing-voice synthesis. *Speech communication*, 46(3), 405-417.

Saitou, T., Goto, M., Unoki, M., & Akagi, M.: (2005). Vocal conversion from speaking voice to singing voice using STRAIGHT. *Proceedings of INTERSPEECH*, 4005-4006.

Scherer, R.C., Radhakrishnan, N., Boominathan, P. & Tan, T. (2008). Rate of change of F0 in performance singing. *J. Acoust. Soc. Am*., 123/5/Pt.2, 3379.

Shih, C., & Kochanski, G.P. (2001). Synthesis of Prosodic Styles. *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*.

Za'ba, N., Jamil, N., Salleh, S. S., & Rahman, N. A. (2014) Synthesizing Asli Malay Song: Transforming Spoken Voices into Singing Voices. *Proceedings of the 8th International Conference on Robotic, Vision, Signal Processing & Power Applications* (pp. 303-310). Springer Singapore.

Za'ba, N., Jamil, N., Salleh, S. S., & Rahman, N. A. (2011). Investigating ornamentation in Malay traditional, Asli Music. Proceedings of the WSEAS International Conference/EURO-SIAM/Europment, 46-52.