An Optimization of Fundamental Frequency and Length of Syllables for Rule-Based Speech Synthesis

Kyawt Yin Win and Tomio Takara

Department of Information engineering, University of the Ryukyus, Okinawa, Japan
win@iip.ie.u-ryukyu.ac.jp, takara@ie.u-ryukyu.ac.jp

Abstract. In this paper an optimization method has been proposed to minimize the differences of fundamental frequency (F_0) and the differences of length among the speakers and the phonemes. Within tone languages use pitch variation to construct meaning of the words, we need to define the optimized fundamental F_0 and length to obtain the naturalness of synthetic sound. Large variability exists in the F_0 and the length uttered by deferent speakers and different syllables. Hence for speech synthesis normalization of F_0 and lengths are important to discriminate tones. Here, we implement tone rule by using two parameters; optimized F_0 and length. As an advantage in the proposed method, the optimized parameters can be separated to male and female group. The effectiveness of the proposed method is confirmed by the distribution of F_0 and length. Listening tests with high correct rates approve intelligibility of synthetic sound.

Keywords: Speech, Optimization, Normalization, Myanmar tone, Rule-based synthesis.

1 Introduction

There are some researches on optimal unit selection algorithm for corpus-based TTS system [1]. In our former research, we introduced Rule-based Myanmar speech synthesis system [2-3]. In that system fundamental speech units are demi-syllables with level tone. To construct the TTS system, monosyllabic words are analyzed and the parameters are obtained for synthesis of tones. Tone rules were F_0 linear pattern.

Within tone languages that use pitch variations to contrast meaning [4]. For example, Myanmar is a tonal language comprising four different lexical tones. Fig.1 shows an example of F_0 contour of the four Myanmar tones with the syllable /ma/ uttered by a male native speaker. Also Mandarin Chinese has four different lexical tones. The exact nature of the F_0 characteristics of Mandarin words is highly variable across utterances and speaker. Four lexical tones in isolated syllables can be characterized to mainly in terms of the shape of their F_0 contour. Therefore F_0 contour is the most crucial characteristic of tone. Furthermore duration of tones is also important [5]. Even rule-based speech synthetic system with linear F_0 pattern is very simple, it is important to define reliable value of F_0 and syllables length to implement synthesis rule. The acoustic of speech are notoriously variable across speakers. Large variability exists in the F_0 height and the length of syllables uttered by deferent

T.-h. Kim et al. (Eds.): FGIT 2010, LNCS 6485, pp. 114–124, 2010.

speakers and different syllables [4]. Hence for speech synthesis optimization of F_0 and lengths are important and necessary to discriminate tones.

Standard Myanmar is used by 8 main races and sub-races as an official language. It is spoken in most of the country with slight regional variations. In addition, there are other regional variants that differ from standard Myanmar in pronunciation and vocabulary [6, 7, 8]. Accordingly a large variability exists in the F_0 and lengths among the speakers. Beside in Myanmar, however, tones are unique in their simplistic pattern not only related to F_0 but also more specifically and importantly in terms of length. Myanmar tones have different lengths between short-tone and long-tone groups. This is the basis for the proposed linear pattern for tone rule using optimized F_0 and optimized lengths.

In our former research, tone rule is implemented with linear pattern using the average F_0 and the averages of syllable's length which are normalized value. Even though, the reasonable high intelligibility of synthesized tone was confirmed through listening tests of synthesized words, there are some errors between male and female speech parameters.

In this paper we normalize F_0 and length of each tone, so that the square-sum of each difference between F_0 and its arithmetical average was minimized by using optimization method. The average F_0 for each word is selected from the frames at the center of syllable. The synthetic speeches are evaluated by listening tests. The results show that our proposed method gives high intelligibility of synthetic sound comparing with other tone synthesis rule with F_0 linear pattern, such as VieTTS [9].

The organization of the paper is as follows.

Section 1: Introduction
Section 2: Background of Speech Synthesis System
Section 3: Tone Synthesis Procedure using optimized F_0 and length
Section 4: Results and Discussion
Section 5: Conclusion.

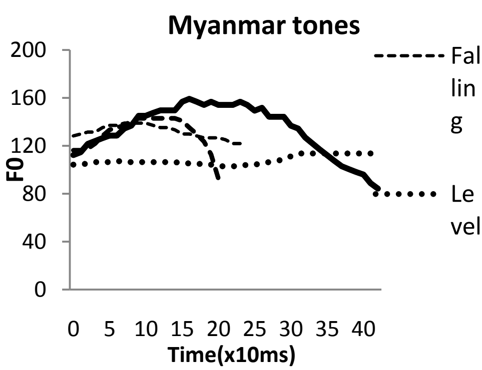


Fig. 1. Example of four tones of Myanmar syllable /ma/

2 Background of Speech Synthesis System

2.1 Speech Analysis and Synthesis

2.1.1 Speech Analysis

The analysis part of our speech synthesis system is designed using cepstral analysis. The frame length is 25.6ms and the frame shifting time is 10ms. As the window function for speech analysis, a time–domain Hamming window is used with a length of 25.6ms. The cepstral coefficient or cepstrum is defined as the inverse Fourier transform of the short-time logarithmic amplitude spectrum [10]. The special feature of the cepstrum is that it allows separating representation of the spectral envelope and excitation. The resulting parameter of speech units include the number of frames and, for each frame, voiced/unvoiced (V/UV) decision, pitch period and cepstral coefficients $c(m)$, $0 \leq m \leq 29$.

2.1.2 Speech Synthesis

Under the control of the synthesis rule, the speech synthesis sub-system generates speech from pre-stored parameters. The source-filter model [11] is used as the speech production model. Fig.2 shows the structure of the speech synthesis sub-system in MyanmarTTS. The synthetic sound is produced using the Log Magnitude Approximation (LMA) filter, which has been introduced by Imai [12]. It presents the vocal tract characteristics. The spectral envelope is represented by the cepstral coefficients of 30 lower-order frequency elements. The LMA filter is a pole–zero filters that is able to represent efficiently the vocal tract features for all speech sounds.

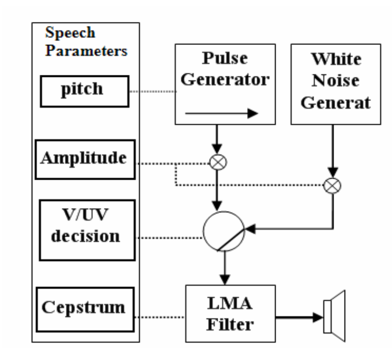


Fig. 2. MyanmarTTS speech synthesis sub-system

3 Tone Synthesis Procedure Using Optimized F_0 and Length

3.1 Tone Synthesis

The four Myanmar tones are analyzed to extract F_0 patterns. The data set is prepared as voiced sounds and meaningful words. We select consonant-vowel (CV) form with voiced consonants /b/, /m/, /l/ and three typical vowels /a/, /i/ and /u/. In total, 180 words (= 3 consonants x 3 vowels x 4 tones x 5 speakers) are used for tone analysis. After analyzing, four tones are distributed as shown in Fig.3. We find that the four tone groups overlapped and are not clearly discriminated. In our former research, we normalized F_0 and length to obtain relative values among the tones. The normalized parameters of tones using one syllable word were plotted in the distribution [3]. In this paper the normalized parameters by former normalization method using three syllables are shown in Fig.4.

3.2 Proposed Optimization Method

Lagrange's optimization method [13-14] is used for normalization. In this study we use 36 words of F_0 patterns by utterance of five native speakers. The words include three typical vowels "a", "i" and "u" with voiced consonants "b", "m" and "i". We select F_0 from three frames at the center of syllable word for each tone. The average F_0 values are selected from the middle frames of F_0 contours.

To minimize large differences of F_0 and differences of lengths among the speakers by means of tones, optimization method is carried out. The average of F_0 contours for each tone is given by

$$f_{ij}=1/n\sum_{k=1}^{n} f_{ij}^{k} \tag{1}$$

where n is number of F_0 contour. f_{ij} is F_0 at the center of syllable of i^{th} tone and j^{th} speaker.

Similarly, the average of tones is defined as A_j and the average of all speakers is defined as A.

To normalize f_{ij}, Lagrange's optimization technique is utilized in this paper. For convenience, we define $U_{ij}{}^{0}$ and R_{ij} such as

$$R_{ij} = A\text{-}A_j \tag{2}$$

$$U_{ij}{}^{0}=f_{ij}{}^{0}\text{-}f_{ij} \tag{3}$$

where, $f_{ij}{}^{0}$ are normalized values of f_{ij}.

Then, in our problem, concentration of $f_{ij}{}^{0}$ around A is accomplished by minimizing

$$W(f_{ij}^{0}) = \sum_{j=1}^{s}\left(A-f_{ij}^{0}\right)^{2} \tag{4}$$

under the constraints

$$U_{ij}{}^{0} = \alpha_{ij}R_{ij} \tag{5}$$

where, α_{ij} are scale numbers and s is numbers of speaker.

Thus, normalized $f_{ij}{}^{0}$ are given by minimizing Lagrange's function L ($f_{ij}{}^{0}$)

$$\mathrm{L}\,(f_{ij}^{0}) = \mathrm{W}(f_{ij}^{0}) + \sum_{j=1}^{s}\lambda_j\left(U_{ij}-\alpha_{ij}R_{ij}\right) \tag{6}$$

For Eq. (6), we have

$$\frac{\partial \mathrm{L}}{\partial f_{ij}^{0}} = 2\left(f_{ij}^{0} - A\right) + \lambda_j = 0 \tag{7}$$

$$\frac{\partial \mathrm{L}}{\partial \lambda_j} = U_{ij} - \alpha_{ij} R_{ij} = 0 \tag{8}$$

Solving Eqs. (7), (8) gives

$$f_{ij}^{0} = f_{ij} + \alpha_{ij} R_{ij} \tag{9}$$

$$\lambda j = 2(A - f_{ij} - \alpha_{ij} R_{ij}) \tag{10}$$

According to Eqs.(2) and (3), equation (5) indicates that if $\alpha_{ij}=1$, f_{ij} around A_j, i.e., $f_{ij} - A_j$ is shifted to f_{ij}^{0} around A, i.e., $f_{ij}^{0} - A$, while $\alpha_{ij} = 0$, i.e., $f_{ij}^{0} = f_{ij}$ which doesn't give normalization. When male and female speakers intermix, average A behaves as a center of A_j for male and A_j for female.

On the other hand, the minimum value of L is derived as follows:

$$\mathrm{L}_{\mathrm{min}} = \sum_{j=1}^{s} (A - f_{ij} - \alpha_{ij}^{0} R_{ij})^2 \tag{11}$$

which leads

$$\alpha_{ij}^{0} = (A - f_{ij}) / R_{ij} \tag{12}$$

because $\mathrm{L}_{\mathrm{min}} \geq 0$.

$$(A - f_{ij}) / R_{ij} > 0 \tag{13}$$

Hence, f_{ij} and Aj are always the same side of A.

Then, we have the relation

$$0 \leq \alpha_{ij} \leq \alpha_{ij}^{0} \tag{14}$$

From Eqs.(3) and (5),we get general equation

$$f_{ij}^{0} = f_{ij} + \alpha_{ij} R_i \tag{15}$$

For the sake of convenience, we may simply choose α_{ij} in this paper, such that

$$\alpha_{ij} = \alpha = 1/2 \tag{16}$$

In this way f_{ij} is normalized. The normalized value f_{ij}^{0} is given by,

$$f_{ij}^{0} = f_{ij} + \alpha R_{ij} \tag{17}$$

The optimized results are plotted in Fig. 5. Fig.5 (a), (b) show the distribution of four tones with optimized F_0 and optimized lengths, which are clearly discriminated in tone groups. From these figures we confirm that proposed method is an effective method to define the parameters for speech synthesis rule. Furthermore, as an advantage in the proposed method, the male and female can be distinguished.

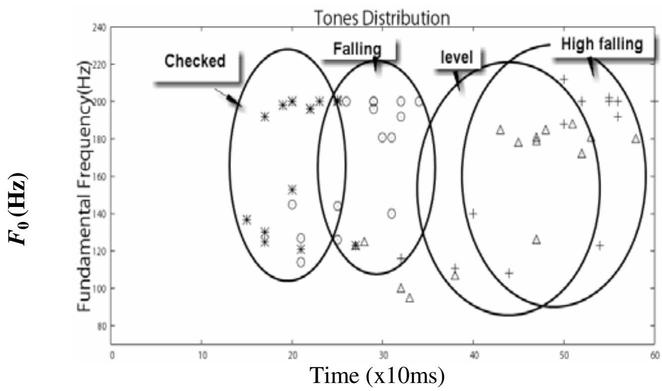


Fig. 3. Tones distribution of analysis-synthesis sounds by three female speakers and two male speakers before optimization

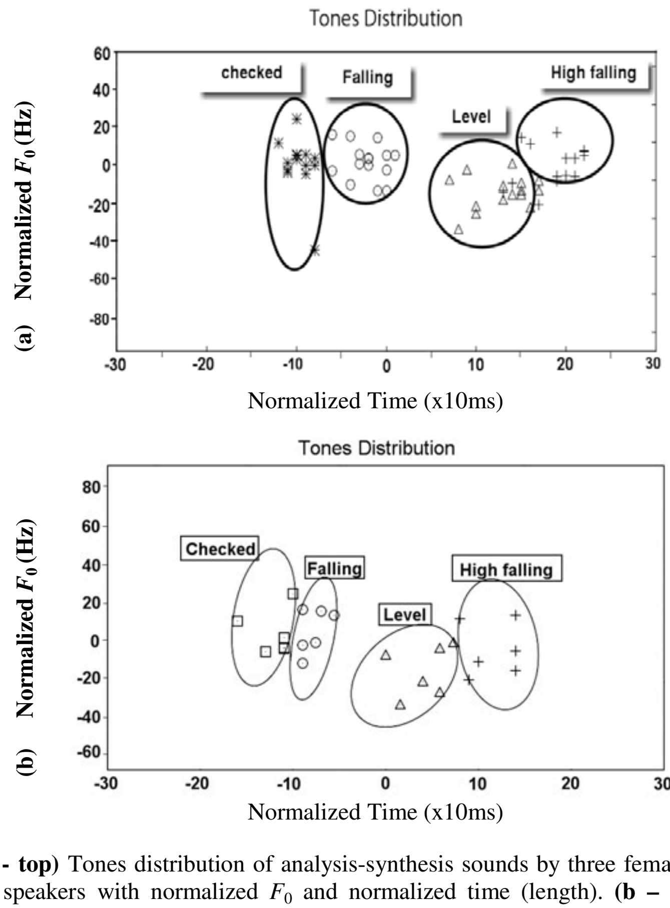


Fig. 4. (a - top) Tones distribution of analysis-synthesis sounds by three female speakers and two male speakers with normalized F_0 and normalized time (length). **(b – bottom)** Tones distribution of analysis-synthesis sounds uttered by two male speakers with normalized F_0 normalized time (length).

3.3 Tone Synthesis Rule with Linear F_0 Pattern

Myanmar tones are unique in their simplistic pattern not only related to F_0 but also more specifically and importantly in terms of length. Myanmar tones have different lengths between short-tone and long-tone groups. In accordance, after optimization we define tone rule employing two parameters; F_0 at the center of syllables and syllable's length as opposed to focusing on length alone. Tone rules are constructed with linear F_0 patterns.

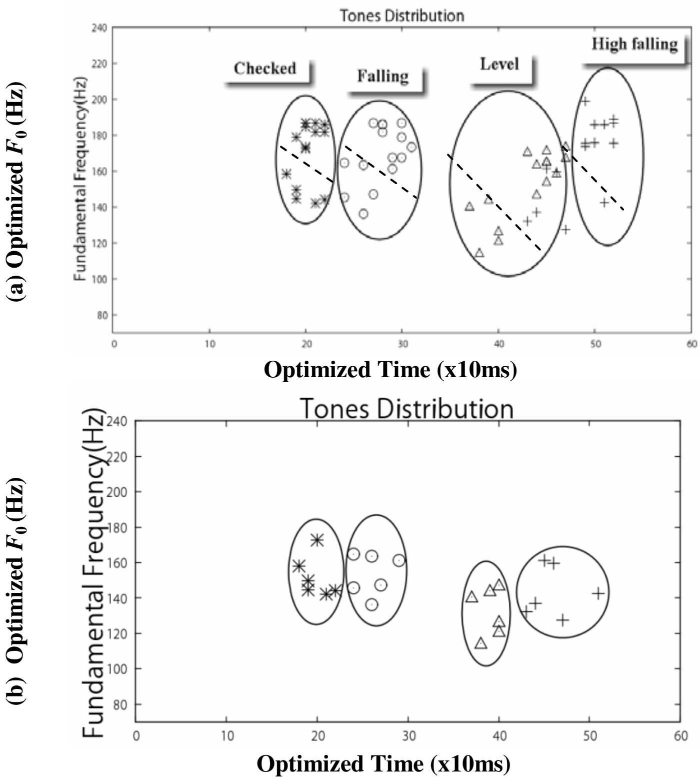


Fig. 5. (a - top) Tones distribution by three female speakers and two male speakers with optimized F_0, and optimized length. **(b - bottom)** Tones distribution by two male speakers with optimized F_0, and optimized length.

When we calculated the average frame length and average F_0 to make tone rules for male and female, we apply the concept of the center of gravity. As an example, Fig. 6 shows the calculation design of average F_0 and length using center of gravity. The tone rules are implemented based on optimized F_0 and optimized length of each tone as shown in Fig. 7.

We consider F_0 distribution as the mass distribution. We calculate average F_0 and length by using the concept of center of gravity x as follows:

$$x = \left(\sum_{i=1}^{n} x_i m_i\right)/M \tag{18}$$

$$M = m_1 + m_2 + m_3 + \ldots + m_n$$

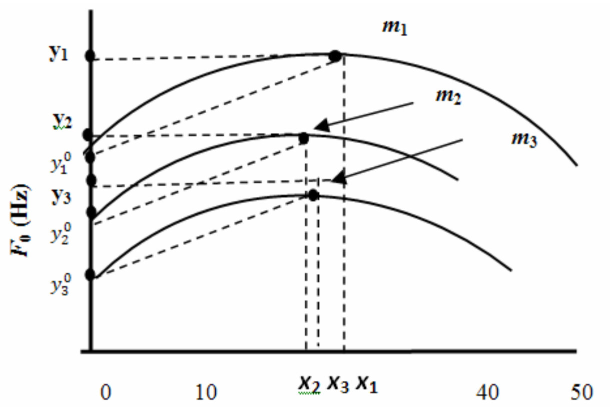


Fig. 6. The calculation design of average F_0 and length

Where m_i represents the weight of personal quality of F_0 of i^{th} speaker and x is average length of F_0 contour. Specifically, weight of personal quality of F_0 is different among the different speakers. As an example for three speakers, m_1, m_2 and m_3 are different values. In our experiments, all speakers are native and they have clear utterances and hearing ability. Therefore in this paper we consider their speech units have the same reliability. Then we have,

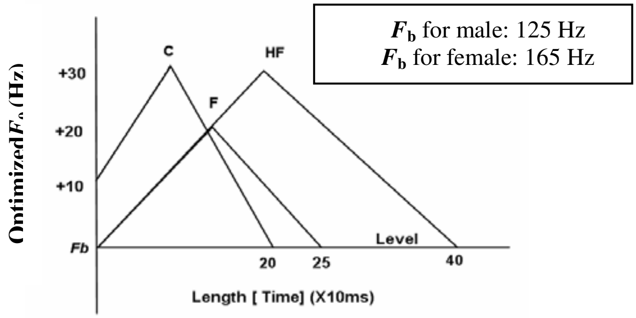
$m_1 = m_2 = m_3 = m$ (*Example*: for three speakers)

From Eq. (16) average F_0 value at the center of contour y is calculated as

$$y = \frac{m(y_1 + y_2 + y_3)}{3m} = \frac{(y_1 + y_2 + y_3)}{3} \tag{19}$$

Similarly the average length of time co-ordinate x is calculated as

$$x = \frac{(x_1 + x_2 + x_3)}{3} \tag{20}$$



***L*: Level tone, *F*: Falling tone, *Hf*: High falling tone, C: Checked tone,**

Fig. 7. The diagram of tone rule

Using these rules, we carried out the listening tests to evaluate intelligibilities of synthetic speech of syllables and to evaluate the effect of proposed method.

4 Results and Discussion

Results of these tests are shown Fig. 8. These results have been obtained by using listening test. The result of our tone synthesis system and effectiveness of optimization are discussed as follows:

- Proposed method elicits the highest correct rate 99.68% for male speakers and 98.75% for female speakers.
- From these results we can confirm that optimized F_0 and length are conducted natural synthetic speech. Since we defined the scale factors of relative values properly, the optimized values are obtained.
- In VieTTS system[9], the result for linear pattern is about 85% for male, whereas the result of our system for male is 95.8%, even though our listening tests were done using the speech sounds of multiple speakers and different genders. Consequently, we can show that our linear pattern for tone rule is more effective than VieTTS's corresponding one since we applied the optimization method by means of multiple speakers and multiple phonemes.
- As a discussion concerning with above mentioned comparison, we consider that the optimization gives the effective values for both male and female, since we defined the scale factors of relative values correctly.
- Consequently, the introduced optimization method is effective and applicable for other speech synthesis rule for other tonal languages.

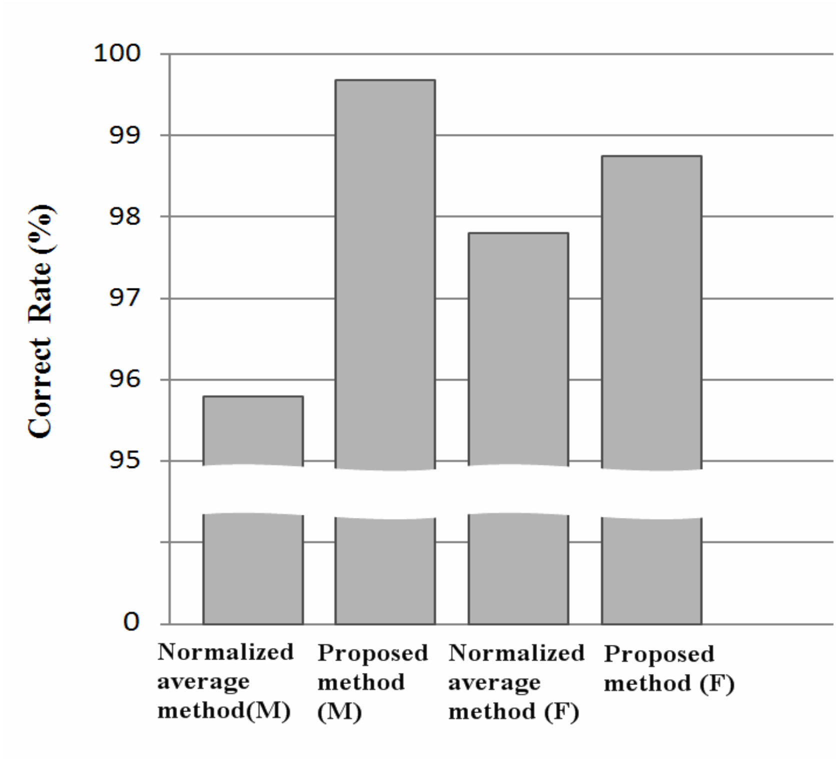


Fig. 8. The results of correct rate of perception of synthesized tone

5 Conclusion

An optimization method to define the parameters; F_0 and syllable's length for tone synthesis is introduced. We implemented tone rules of linear pattern based on two parameters, the optimized F_0 at the center of syllable and the optimized syllable's length. The effectiveness of the proposed method is confirmed by distribution of tones and the intelligibility scores of listening test. Although the high intelligibility of synthesized tone draws reasonably high correct rates in former research, the proposed method achieve the better results. Furthermore, in the proposed method, the optimized parameters can be separated into male and female groups. The introduced proposed method is applicable for other tone synthesis rule of other tonal languages.

References

1. Lee, M., Lopresti, D.P., Olive, J.P.: A Text-To-Speech Platform for Variable Length Optimal Unit Searching Using Perception Based Cost Function. International Journel Of Speech Technology 6, 347–365 (2003)
2. Win, K.Y., Takara, T.: Myanmar Speech Synthesis System Using Cepstral Method. In: The International Conference on Electrical Engineering (2008)

3. Win, K.Y., Takara, T.: Rule-based speech synthesis of Myanmar Using Cepstral Method. In: Proceeding of the 11th conference of Oriental-COCOSDA, NICT, Kyoto, Japan, November 25-27, pp. 225–229 (2008)
4. Huang, J., Holt, L.L.: General Perceptual Contributions to Lexical tone normalization. J. Acoust. Soc. Am. 125(6) (June 2009)
5. Zhang, S., Huang, T., Xu, B.: Tone Modeling for Contious Mandarin Speech Recognition. International Journel Of Speech Technology 7, 115–128 (2004)
6. Myanmar Language Committee, "Myanmar Grammar", Myanmar Language Committee, Ministry of Education, Myanmar (2005)
7. Thein Tun, U.: Some acoustic properties of tones in Burmese. In: Bradley, D. (ed.) Papers in South – East Asian Linguistics8: Tonation Canberra: Australian National University, pp. 77–116 (1982)
8. Wheatley, J.K.: Burmese. In: Cormier, B. (ed.) The World's Major Languages, pp. 834–845. Oxford University Press, New York
9. Do, T.T., Takara, T.: Vietnamese Text-To-Speech system with precise tone generation. Acoust. Sci. & Tech. 25(5), 347–353 (2004)
10. Noll, A.M.: Cepstrum Pitch Determination. Journal of the Acoustical Society of America 41(2), 293–309 (1967)
11. Furui, S.: Digital Speech Processing, Synthesis, and Recognition, 2nd edn., pp. 30–31. Marcel Dekker, Inc., New York (2001)
12. Imai, S.: Log Magnitude Approximation (LMA) Filter. Trans. IECE Jpn. J63-A, 886–893 (1980)
13. Xia, Y., Wang, J.: A General Methology for Desiging Globally Convergent Optimization Neural Networks. IEEE Transactions on Neural Networks 9(6) (November 1998)
14. Deng, L., Shaughnessy, D.O.: Speech Processing A dynamic and Optimization-Oriented Approach. Marcel Dekker, Inc., New York (2003)