Journal of Computer Science 6 (3): 330-335, 2010
ISSN 1549-3636

Structural Modeling of Fundamental Frequency Contour for Thai Expressive Speech

Suphattharachai Chomphan
Department of Electrical Engineering, Faculty of Engineering at Si Racha,
Kasetsart University, 199 M.6, Tungsukhla, Si Racha, Chonburi, 20230, Thailand

Abstract: Problem statement: Appropriate modeling of fundamental Frequency (F_0) contour for speech is a key factor to preserve the quality of speech prosody. One successful approach has been conducted for tonal language of Mandarin Chinese. It is based on the assumption that the behavioral characteristics of vocal-fold elongation in vibration could be approximated by those of a simple forced vibrating system. Therefore this approach has been applied to model Thai expressive speech with best-fit function. **Approach:** An approach of structural modeling of voice F_0 contours of Thai expressive speech utterances using an approximation by those of a simple forced vibrating system has been conducted. Nowadays, modeling of F_0 contours of Thai expressive speech is very important in an analysis of speech, which brings about the speech communication with more interesting and effective. Our speech database consists of male and female speech and each one contains 4 different speech styles including angry style, sad style and enjoyable style and reading style. We use 5 sentences for each speech style and each sentence includes 100 samples. The speech sample in each group is analyzed for an F_0 contour, subsequently a number of structural modeling parameters are extracted for each contour. Thereafter, the parameters are used to synthesis the F_0 contour and then the synthesized contour is compared with that of natural speech by calculating RMS error. **Results:** From the experimental analysis, it is observed that RMS error of each speech style is different from the others. It reveals that the mentioned structural modeling responses to each speech style differently. Moreover the reading style has the smallest error among all styles. **Conclusion:** From the finding, it is a definite evidence to apply the modeling approach to the speech synthesis systems with good preservation of speech prosody.

Key words: Thai expressive speech, fundamental frequency modeling, structural modeling, expressive speech, simple forced vibrating system

INTRODUCTION

Prosody is an inherently supra-segmental feature of human speech. The fundamental frequency of voice speech is the most important feature among all of the features known to carry prosodic information. The F_0 contours of an utterance convey the stress, intonation and rhythmic structures, which determine the naturalness and intelligibility of synthetic speech. As a result, the appropriate modeling of F_0 contour plays a significant role in the speech processing area, e.g., speech recognition, speech synthesis, speech analysis and speech coding. A number of modeling techniques in the former studies have been performed in various levels of speech units, e.g., utterance level (Saito and Sakamoto, 2002; Li *et al*., 2004; Tao *et al*., 2006), word and syllable levels (Fujisaki and Sudo, 1971; Tran *et al*., 2006). In Thai speech, Fujisaki's model has been successfully applied for modeling of utterances, tones and words (Hiroya and Sumio, 2002; Seresangtakul and Takara, 2002; 2003). In the Thai speech synthesis, the statistical modeling of F_0 contour has been conducted by Chomphan and Kobayashi in the implementation of both speaker-dependent and speaker-independent systems in 2007-2009 (Chomphan and Kobayashi, 2007; 2008; 2009; Chomphan, 2009). Lately, the Fujisaki's model has been applied within a speaker-independent system as extended modules. Moreover, it has also been exploited in the modeling of Thai expressive speech; i.e., sad, happy, angry styles (Chomphan, 2010). This study proposed another approach of F_0 modeling of Thai expressive speech using the structural model which is based on the assumption that the behavioral characteristics of vocal-fold elongation in vibration could be approximated by those of a simple forced vibrating system (Ni and Hirose, 2006). The RMS error calculation has been done for evaluation the modeling performance for all

speech styles including angry style, sad style and enjoyable style and reading style. This study is a preliminary study for the advanced research in an advanced speech synthesis with various speaking styles for Thai.

MATERIALS AND METHODS

Structural model: The voice F_0 contour is modeled in a logarithmic scale, as depicted in Fig. 1. The mathematical model has been applied (Ni and Hirose, 2006) by using a structural control consisting of placing a series of normalized F_0 targets along the time axis, which are also specified by transition time and amplitudes. The transitions between targets are approximated by connecting truncated second-order transition functions.

From the background knowledge that the physical factors to regulate the frequency of vocal-fold vibrations are the mass, length and tension of vibrating structures, all of which are dynamically controlled primarily by the intrinsic and extrinsic muscles of the larynx and secondly by the sub-glottal pressure (Ni and Hirose, 2006). Fujisaki explained that logarithmic fundamental frequency varies linearly with vocal-fold elongation x (Fujisaki, 1983), which can be represented in the following mathematical term:

$$\ln f_0 = \frac{b}{2}x + \ln(\sqrt{a}c_0) \quad (1)$$

where, a, b and c_0 are constant coefficients (Fujisaki, 1983).

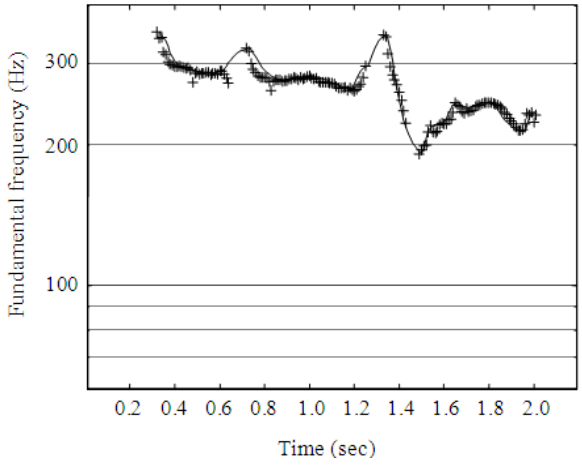

Fig. 1: F_0 contour with a trend line in a logarithmic scale

Assumption: The behavioral characteristics of vocal-fold elongation in vibration can be approximated by those of a simple forced vibrating system (Ni and Hirose, 2006).

Formulating the assumption, the behavioral characteristics of a simple forced vibrating system can be characterized by the amplifying coefficients of its vibrating amplitudes:

$$A\left(\frac{\omega_d^2}{\omega_f^2},\zeta\right) = \frac{1}{\sqrt{\left(1-\left(1-2\zeta^2\right)\frac{\omega_d^2}{\omega_f^2}\right)^2 + 4\zeta^2\left(1-2\zeta^2\right)\frac{\omega_d^2}{\omega_f^2}}} \quad (2)$$

where, ω_d and ω_f denote the natural angular frequencies of the driven system and the driving force, respectively, ζ is called damping ratio indicating how tightly the driving force and the driven system are coupled together. Subsequently, replacing ω_d^2/ω_f^2 (square frequency ratio) by λ and substituting $A(\lambda, \zeta)$ expressed in Eq. 2 for x of Eq. 1, as a result, the logarithmic fundamental frequency can be expressed as:

$$\ln f_0 = \frac{b\times C}{2}A(\lambda,\zeta) + \ln(\sqrt{a}c_0) \quad (3)$$

where, C is a constant coefficient.

Typically, a speaker has an individual vocal range. Let f_{0t} and f_{0b} denote the top and bottom frequencies of the vocal range of a speaker and λ_t and λ_b denote two λ values that are one-to-one mapped to f_{0t} and f_{0b}. The relationship between f_0 within the vocal-range frequency interval and its corresponding λ is shown as follows:

$$\frac{\ln f_0 - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\lambda,\zeta) - A(\lambda_b,\zeta)}{A(\lambda_t,\zeta) - A(\lambda_b,\zeta)} \quad (4)$$

Since f_{0t} and f_{0b} are the top and bottom frequencies of the vocal range, λ_t and λ_b shall be determined regardless of ζ.

Practically, f_0 and λ can be determined through $f_0 = T_{f0}(\lambda, \zeta)$ and $\lambda = T_\lambda(f_0, \zeta)$, where they can be derived from Eq. 4 as followings:

$$T_{f_0}(\lambda,\zeta) = \exp\left(\frac{A(\lambda,\zeta) - A(\lambda_b,\zeta)}{A(\lambda_t,\zeta) - A(\lambda_b,\zeta)} \times \ln\frac{f_{0_t}}{f_{0_b}} + \ln f_{0_b}\right) \quad (5)$$

and $T_\lambda(f_0, \zeta)$ can be obtained by searching λ from 1 step-by-step in small increments (e.g., 0.0001), given $\lambda_b > \lambda_t$, until λ satisfies the following conditions:

$$\ln\frac{f_0}{f_{0_b}} > \frac{A(\lambda+0.0001,\zeta)-A(\lambda_b,\zeta)}{A(\lambda_t,\zeta)-A(\lambda_b,\zeta)}\times\ln\frac{f_{0_t}}{f_{0_b}} \quad (6)$$

and

$$\ln\frac{f_0}{f_{0_b}} < \frac{A(\lambda-0.0001,\zeta)-A(\lambda_b,\zeta)}{A(\lambda_t,\zeta)-A(\lambda_b,\zeta)}\times\ln\frac{f_{0_t}}{f_{0_b}} \quad (7)$$

Allocation of pitch targets: By applying the structural model, the pitch targets on an F_0 contour is allocated in advance. Thereafter the parameter of the structural model will be approximated. For example, Fig. 2 shows a sparser specification of the F_0 contours shown in Fig. 1, the 'o' signs indicate the tonal peaks, the 'x' signs indicate the tonal valleys, and the square sign indicates a neutral target to reset an intonational phrase.

Let t_i and f_{0i} denote the timing and F_0 value of the i^{th} target, respectively. Table 1 lists the first eight target points (t_i, f_{0i}) and corresponding $\lambda_i = T_\lambda(f_{0i}, \zeta_0)$, given $\lambda_t = 1$; $\lambda_b = 2$; $f_{0b} = 120$Hz; $f_{0t} = 420$Hz and $\zeta_0 = 0.156$. The i^{th} local F_0 movement, i = 0, ..., 15, is defined as a scope extending from the i^{th} target (t_i, f_{0i}) to the next. If $f_{0i} \le f_{0i+1}$, the local F_0 movement is therefore rising; otherwise, falling. The measured F_0 contours are subsequently approximated by connecting the rising and falling transitions through these target points.

Table 1: Target values for sparsely specifying an F_0 contour

i	0	1	2	3	4	5	6	7
t_i (s)	0.32	0.46	0.57	0.72	0.89	1.00	1.18	1.25
f_{0i} (Hz)	351.00	296.00	287.00	323.00	275.00	280.00	266.00	296.00
λ_i	1.28	1.40	1.42	1.34	1.44	1.42	1.46	1.40

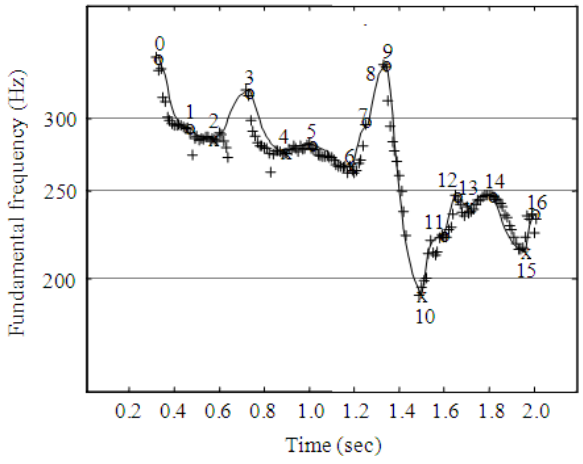

Fig. 2: Example of pitch target allocation on an F_0 contour

For an F_0 falling movement, say i = 0, first compute $\lambda(t)$ for $t_0 \le t \le t_1$ by using following equation:

$$\lambda(t) = \lambda_p + \Delta\lambda\left(1-\left(1+\frac{4.8}{\Delta t}t\right)e^{-\frac{4.8}{\Delta t}t}\right), \quad \text{for } t \ge 0 \quad (8)$$

with the following parameters (based on Table 1): $\lambda_p (= \lambda_0) = 1.28$; $\Delta\lambda = \Delta\lambda_0/0.95$ and $\Delta t = \Delta t_0/0.95$, where $\Delta\lambda_0 = (\Delta\lambda_1 - \lambda_0) = 0.12$; $\Delta t_0 = (t_1 - t_0) = 0.14$. Second, synthesize contour by using $f_0(t) = T_{f0}(\lambda(t), \zeta_0)$ of Eq. 5. The thick line between the 0^{th} and 1^{st} targets shown in Fig. 2 indicates the re-synthesized F_0 contour.

For an F_0 rising movement, say i = 2, first compute $\lambda(t)$ for $t_2 \le t \le t_3$ by using Eq. 8 with the following parameters: $\lambda_p = 2-\lambda_2 = 0.58$; $\Delta\lambda = \Delta\lambda_2/0.95$, where $\Delta\lambda = (2-\lambda_3)-(2-\lambda_2) = 0.08$ and $\Delta t = \Delta t_2/0.95$, where $\Delta t_2 = t_3 - t_2 = 0.15$. Then, synthesize contour by using $f_0(t) = T_{f0}(2-\lambda(t), \zeta_0)$ of Eq. 5. It has been noted that $A(\lambda, \zeta) = A(2-\lambda, \zeta)$ is applied to the computation. In Fig. 2, the thick line between the 2^{nd} and 3^{rd} targets indicates the re-synthesized F_0 contour.

A mathematical model: Subsequently, let $F_0(t)$ represent an F_0 contour as a function of time t in a vocal range $[f_{0b}, f_{0t}]$ (Ni and Hirose, 2006). Assume $\Lambda(t)$ to indicate a sequence of virtual tone graphs in λ-time space to specify the underlying lexical tone structures. Additionally, assume a latent scale $\zeta(t)$ to characterize the intonation components. Thus, the F_0 contour on the logarithmic scale of fundamental frequency is expressed as a scale transformation from $\Lambda(t)$ to $F_0(t)$, corresponding to the syllabic tones fitting themselves with sentence intonation in the vocal range:

$$\frac{\ln F_0(t) - \ln f_{0_b}}{\ln f_{0_t} - \ln f_{0_b}} = \frac{A(\Lambda(t),\zeta(t)) - A(\lambda_b,\zeta(t))}{A(\lambda_t,\zeta(t)) - A(\lambda_b,\zeta(t))}, \quad \text{for } t \ge 0 \quad (9)$$

Where:

$$A(\lambda,\zeta) = \frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2 + 4\zeta^2(1-2\zeta^2)\lambda}}, \quad \text{for } \lambda \ge 1 \quad (10)$$

Equation 9 and 10 jointly indicate a structural formulation of the control process of coupling the syllabic tones and sentence intonation together to form a final sentence melody. Equation 9 states that F_0 contour $F_0(t)$ is a transformation of a sequence of virtual tone graphs $\Lambda(t)$ on a latent scale $\zeta(t)$. $\Lambda(t)$ is expressed as a concatenation of n parametric bell-shaped patterns lining up in series on the time axis with the following definition:

$$\Lambda(t) = \Lambda_{r_1}(t) + \sum_{i=1}^{n-1} \min(\Lambda_{f_i}(t), \Lambda_{r_{i+1}}(t)) + \Lambda_{f_n}(t) \quad (11)$$

$\Lambda_{ri}(t)$ and $\Lambda_{fi}(t)$ indicate the rising and falling transitions of the i^{th} bell-shaped pattern, respectively. Their definitions are as follows:

$$\Lambda_{r_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{f_i}(1 - D_{r_i}(t_{p_i} - t)), & \text{for } t_{p_{i-1}} \le t < t_{p_i} \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

$$\Lambda_{f_i}(t) = \begin{cases} \lambda_{p_i} + \Delta\lambda_{f_i}(1 - D_{f_i}(t - t_{p_i})), & \text{for } t_{p_i} \le t < t_{p_{i+1}} \\ 0, & \text{otherwise,} \end{cases} \quad (13)$$

Where:

$$D_{x_i}(t) = \left(1 + \frac{4.8t}{\Delta t_{x_i}}\right) e^{-\frac{4.8t}{\Delta t_{x_i}}}, \quad x \in \{r, f\} \quad (14)$$

The model parameters in Eq. 9-14 are defined as follows:

- $[f_{0_b}, f_{0_t}]$ bottom and top frequencies of the vocal Range in hertz
- $[\lambda_b, \lambda_t]$ bottom and top values of normalized vocal ranges in λ
- $\zeta(t)$ latent scales
- n number of the bell-shaped patterns
- (t_{p_i}, λ_{p_i}) i^{th} peak coordinate in λ-time space; $t_{p_0} = 0$ and $t_{p_{n+1}} = \infty$
- Δt_{r_i} i^{th} rising transition time
- $\Delta\lambda_{r_i}$ i^{th} rising transition amplitude
- Δt_{f_i} i^{th} falling transition time and
- $\Delta\lambda_{f_i}$ i^{th} falling transition amplitude, $i = 1, ..., n$

Figure 3 shows an example of re-synthesis of F_0 contour by using the structural model. Figure 3a shows the F_0 contour extracted from the natural speech, while Fig. 3b shows corresponding value of λ for three different fixed damping ratios ζ (0.156, 0.02 and 0.9). Figure 3c shows the re-synthesized F_0 contour with the three damping ratios in Fig. 3b, while Fig. 3d compares the F_0 contour extracted from the natural speech and the re-synthesized F_0 contour with limited samples.

An experimental design: The flow chart in Fig. 4 shows the core process for our experiment. At first the speech corpus has been implemented. There is male and female speech in the corpus. Each of them has four speech styles including happy, sad, angry and reading styles.

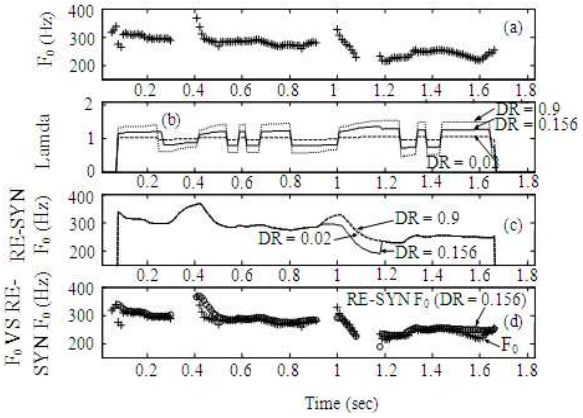

Fig. 3: An example of re-synthesis of F_0 contour by using the structural model (RE-SYN denotes "resynthesized", DR denotes "damping ratio")

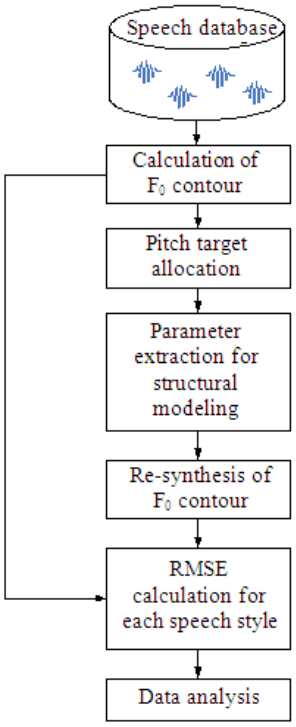

Fig. 4: Work flow for the experimental process

Each style consists of 5 sentences with 100 samples of utterances. Therefore our speech corpus contains 4,000 utterances. At the beginning, the F_0 values of an utterance have been calculated and then the pitch targets have been allocated by using local minimum/maximum criteria. In between any two adjacent pitch targets used as fixed points, an exponential function has been approximated to minimize the

difference between the approximated function and the F_0 contour. The corresponding parameters from all of the functions along the utterance will be used as its representatives. Subsequently, the resynthesis of F_0 contour from the parameters has been conducted. Thereafter, the RMS error between the natural F_0 contour and the resynthesized F_0 contour has been executed. Finally, we analyzed the summarized data from the previous stages.

RESULTS

From the RMS error calculation process, the experimental data can be summarized in the five following graphs (Fig. 5-9). The averaged RMS errors from five different sentences have been calculated.

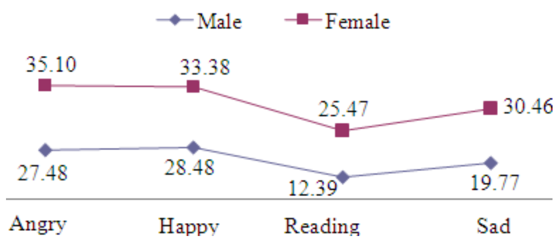

Fig. 5: Averaged RMS error for sentence "✦‾● ● – ● ୨● –4◌● –● ∎⊙∖–● ● ●● –★୨4" in IPA "have you finished your work?" in English

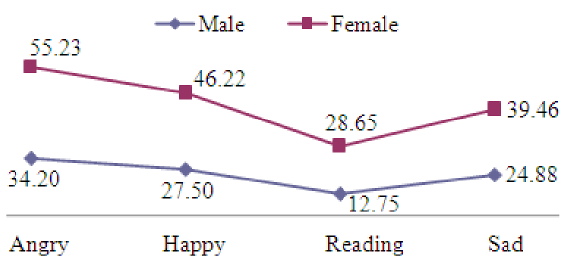

Fig. 6: Averaged RMS error for sentence "● ⁂● – ● ୨★–● ୨●★–● ◌● " in IPA "where have you been?" in English

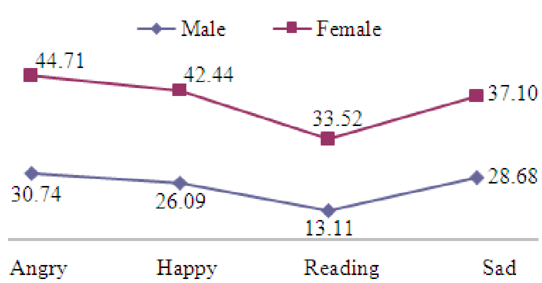

Fig. 7: Averaged RMS error for sentence "ɫ‾୨●● – ɫ‾୨⊙–✦● ◌⊙ɫ–ɫ୨+● ● –● ✶...● ● –● ୨..." in IPA "I will go back home." in English

Each graph represents one sentence and contains those of four speech styles including happy, sad, angry and reading styles. Moreover each graph contains 2 lines of male and female speech.

DISCUSSION

From the experimental results in Fig. 5-9, we found that the averaged RMS error of the angry speech is the highest level; meanwhile the averaged RMS error of the reading speech is the lowest level. The averaged RMS errors of the happy and sad speech are in the middle level. It can be obviously seen from all Fig. 5-9 that all 5 sentences have the corresponding results. When considering the differences between genders, we found that the averaged RMS error of female speech is above that of male speech. It can be seen that all sentences confirm this observation.

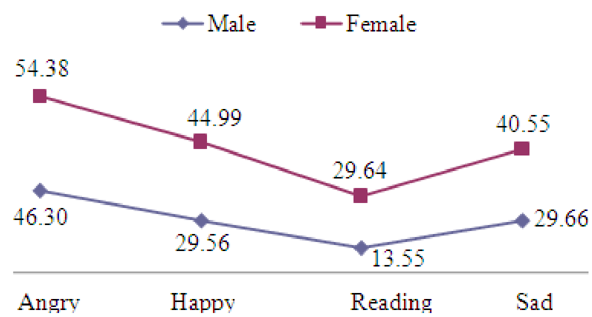

Fig. 8: Averaged RMS error for sentence "ɫ‾୨●● – ● ୨...✦–✦‾● ● –● ⁂+● –● ● ⊙● –● ୨★– ● ● +✦" in IPA "I love you most in the world." in English

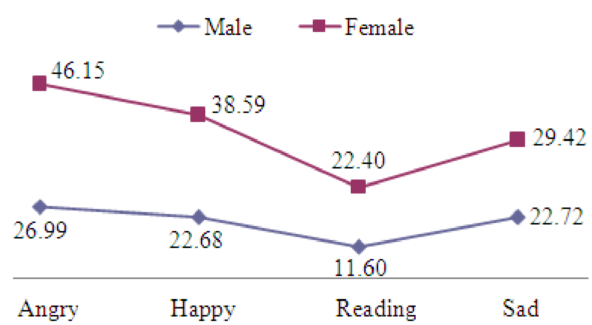

Fig. 9: Averaged RMS error for sentence "● ୨● – ● ୨+★–∖◌+★–● ୨★–● ⁂● +● –● ⁂୨...– ● ∎● " in IPA "We do not go to the sea." in English

CONCLUSION

This study proposes an approach of structural modeling of voice F_0 contours of Thai expressive speech utterances using an approximation by those of a simple forced vibrating system. Four different speech styles have been considered. It has been observed from the experimental analysis that the RMS error of each speech style is different from the others. The reading speech style can be modeled with best fit; meanwhile the angry speech style can be modeled with the highest RMS error. It can be concluded that the structural modeling responses to each speech style differently. For further study, this modeling approach is expected to apply for the speech synthesis systems to preserve the speech prosody with various speaking styles.

ACKNOWLEDGEMENT

The researcher is grateful to N. Sangkaew and A. Sricharoenchot for providing the speech database.

REFERENCES

Chomphan, S. and T. Kobayashi, 2007. Implementation and evaluation of an HMM-based Thai speech synthesis system. Proceeding of the 8th Annual Conference of the International Speech Communication Association, Aug. 2007, ISCA., Antwerp, Belgium, pp: 2849-2852. http://www.isca-speech.org/archive/interspeech_2007/i07_2849.html

Chomphan, S. and T. Kobayashi, 2008. Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis. Speech Commun., 50: 392-404. DOI: 10.1016/j.specom.2007.12.002

Chomphan, S. and T. Kobayashi, 2009. Tone correctness improvement in speaker-independent average-voice-based Thai speech synthesis. Speech Commun., 51: 330-343. DOI: 10.1016/j.specom.2008.10.003

Chomphan, S., 2009. Towards the development of speaker-dependent and speaker-independent hidden Markov model-based Thai speech synthesis. J. Comput. Sci., 5: 905-914. http://www.scipub.org/fulltext/jcs/jcs512905-914.pdf

Chomphan, S., 2010. Analytical study on fundamental frequency contours of Thai expressive speech using Fujisaki's model. J. Comput. Sci., 6: 36-42. http://www.scipub.org/fulltext/jcs/jcs6136-42.pdf

Fujisaki, H. and H. Sudo, 1971. A model for the generation of fundamental frequency contours of Japanese word accent. J. Acoust. Soc. Jap., 57: 445-452. http://ci.nii.ac.jp/naid/110003107854/en

Fujisaki, H., 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. In: The Production of Speech, Mac Neilage, P.F. (Ed.). Springer, New York, pp: 39-55.

Hiroya, F. and O. Sumio, 2002. A preliminary study on the modeling of fundamental frequency contours of Thai utterances. Proceedings of the International Conference on Signal Processing, Aug. 2002, IEEE Xplore Press, Beijing, China, pp: 516-519. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1181106

Li, Y., T. Lee and Y. Qian, 2004. Analysis and modeling of F_0 contours for cantonese text-to-speech. ACM Trans. Asian Language Inform. Process., 3: 169-180. DOI: 10.1145/1037811.1037813

Ni, J. and K. Hirose, 2006. Quantitative and structural modeling of voice fundamental frequency contours of speech in Mandarin. Speech Commun., 48: 989-1008. DOI: 10.1016/j.specom.2006.01.002

Saito, T. and M. Sakamoto, 2002. Applying a hybrid intonation model to a seamless speech synthesizer. Proceeding of the International Conference on Spoken Language Processing, Sept. 2002, Colorado, USA., pp: 165-168. http://www.isca-speech.org/archive/icslp_2002/i02_0165.html

Seresangtakul, P. and T. Takara, 2002. Analysis of pitch contour of Thai tone using Fujisaki's model. Proceeding of the International Conference on Acoustics, Speech and Signal Processing, May 2002, IEEE Xplore Press, Orlando, USA., pp: 505-508. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1005787

Seresangtakul, P. and T. Takara, 2003. A generative model of fundamental frequency contours for polysyllabic words of Thai tones. Proceeding of the International Conference on Acoustics, Speech and Signal Processing, Apr. 2003, IEEE Xplore Press Hong Kong, pp: 452-455. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1198815

Tao J., J. Yu and W. Zhang, 2006. Internal dependence based F_0 model for mandarin TTS system. Proceeding of the TC-STAR Workshop on Speech-to-Speech Translation, June 2006, Barcelona, Spain, pp: 171-174. http://www.elda.org/tcstar-workshop_2006/pdfs/tts/tcstar06_tao.pdf

Tran, D.D., E. Castelli, X.H. Le, J.F. Serignat and V.L. Trinh, 2006. Linear F_0 contour model for Vietnamese tones and Vietnamese syllable synthesis with TD-PSOLA. Proceeding of the International Symposium on Tonal Aspects of Languages, Apr. 2006, La Rochelle, France, pp: 137-142 http://www-mrim.imag.fr/publications/2006/XUA06/TAL2006SubmissionUpdate.pdf