Journal of Computer Science 6 (1): 36-42, 2010
ISSN 1549-3636

Analytical Study on Fundamental Frequency Contours of Thai Expressive Speech Using Fujisaki's Model

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: In spontaneous speech communication, prosody is an important factor that must be taken into account, since the prosody effects on not only the naturalness but also the intelligibility of speech. Focusing on synthesis of Thai expressive speech, a number of systems has been developed for years. However, the expressive speech with various speaking styles has not been accomplished. To achieve the generation of expressive speech, we need to model the fundamental Frequency (F0) contours accurately to preserve the speech prosody. **Approach:** Therefore this study proposes an analysis of model parameters for Thai speech prosody with three speaking styles and two genders which is a preliminary work for speech synthesis. Fujisaki's modeling; a powerful tool to model the F0 contour has been adopted, while the speaking styles of happiness, sadness and reading have been considered. Seven derived parameters from the Fujisaki's model are as follows. The first parameter is baseline frequency which is the lowest level of F0 contour. The second and third parameters are the numbers of phrase commands and tone commands which reflect the frequencies of surges of the utterance in global and local levels, respectively. The fourth and fifth parameters are phrase command and tone command durations which reflect the speed of speaking and the length of a syllable, respectively. The sixth and seventh parameters are amplitudes of phrase command and tone command which reflect the energy of the global speech and the energy of local syllable. **Results:** In the experiments, each speaking style includes 200 samples of one sentence with male and female speech. Therefore our speech database contains 1200 utterances in total. The results show that most of the proposed parameters can distinguish three kinds of speaking styles explicitly. **Conclusion:** From the finding, it is a strong evidence to further apply the successful parameters in the speech synthesis systems or other speech processing technologies.

Key words: Thai expressive speech, Fujisaki's model, fundamental frequency modeling, analysis of fundamental frequency, speech synthesis

INTRODUCTION

In speech processing area; including speech recognition, speech synthesis, speech analysis and speech coding, an appropriate modeling of F0 contour contributes the effectiveness of the implemented speech processing systems. The former study on F0 modeling has been considerably conducted in various speech units and several techniques such as utterance level (Fujisaki and Ohno, 1998; Fujisaki *et al*., 1990; Tao *et al*., 2006; Saito *et al*., 2002 Ni and Hirose, 2006; Li *et al*., 2004), word and syllable levels (Fujisaki *et al*., 1990; 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, 2003; Seresangtakul and Takara, 2002). In the Thai speech synthesis, Chomphan and Kobayashi implemented a speaker-dependent and speaker-independent systems in 2007-2009 (Chomphan and Kobayashi, 2007a; 2007b Chomphan and Kobayashi, 2008; Chomphan and Kobayashi, 2009), in which the F0 contour was modeled using statistical approach. Moreover, the speaker-independent system also used the Fujisaki's model in the extended modules. However, the expressive speech such as sad, happy, angry styles has not been considered. Therefore this study proposed an analysis of F0 modeling of Thai expressive speech using the Fujisaki's model which is a preliminary study for the advanced research in speech synthesis and recognition such as the expressive speech synthesis work in Japanese language (Tachibana *et al*., 2005; Tachibana, 2006).

MATERIALS AND METHODS

Fujisaki's model: The F0 contour is treated as a linear superposition of a global phrase and local-accent components on a logarithmic scale, as depicted in Fig. 1.

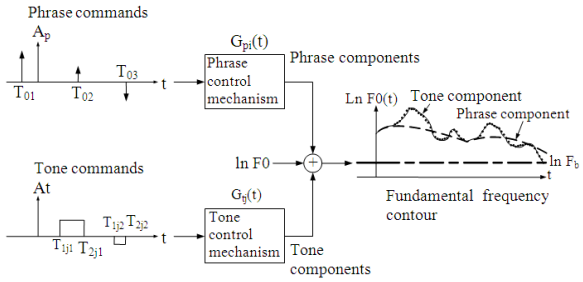

Fig. 1: An extension of Fujisaki's model for the generation of F0 contour

The phrase command produces a baseline component, while the accent command produces the accent component of an F0 contour. We use the two parameters of the Fujisaki's model as our phrase-intonation features including the baseline value of F0 and the magnitude of the phrase command, which complementarily reflect the global level of voicing frequency. Mathematically, the F0 contour of an utterance generated from an extension of the Fujisaki's model for tonal languages has the following expressions (parameters) (Seresangtakul and Takara, 2003):

$$\ln F0(t) = \ln F_b + \sum_{i=1}^{I} A_{pi}[G_{pi}(t - T_{0i})] + \sum_{j=1}^{J}\sum_{k=1}^{K(j)} A_{t,jk}[G_{t,jk}(t - T_{1jk}) - G_{t,jk}(t - T_{2jk})] \quad (1)$$

$$G_{pi}(t) = \begin{cases} (\alpha_i^2 t)\exp(-\alpha_i t) & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \quad (2)$$

$$G_{t,jk}(t) = \begin{cases} [1 - (1+\beta_{jk}t)\exp(-\beta_{jk}t)] & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \quad (3)$$

Where:

$G_{pi}(t)$ = Represents the impulse-response function of the phrase-control mechanism

$G_{t,jk}(t)$ = Represents the step-response function of the tone-control mechanism

The symbols in these equations denote that F_b is the smallest F0 value in the F0 contour of interest and A_{pi} and $A_{t,jk}$ are the amplitudes of the i-th phrases and of the j-th tone command. Here, T_{0i} is the timing of the i-th phrase command and T_{1jk} and T_{2jk} are the onset and offset of the k-th component of the j-th tone command. α_i and β_{jk} are time constant parameters, while I, J, K(j) correspond to the number of phrases, tones and components of the j-th tone contained in the utterance.

To find the optimal representative parameters, optimization is carried out by minimizing the mean squared error in the In F0(t) domain through the hill-climbing search in the space of model parameters (Seresangtakul and Takara, 2003).

By using this model, the parameters are extracted from our speech database, utterance by utterance. Subsequently, the derived parameters are computed are analyzed.

Derived parameters: From the conventional parameters, we proposed seven derived parameters which reflect the geometrical appearance of the F0 contour of an utterance as follows:

- Baseline frequency
- Numbers of phrase commands
- Numbers of tone commands
- Phrase command duration
- Tone command duration
- Amplitude of phrase command
- Amplitude of tone command

All of them have been extracted for three expressive speech styles of happiness, sadness and reading.

RESULTS

In our speech database, we use a sentence of "kʰid tʰɯŋ tɕaŋ lɤ : j" in IPA (means "Think of you so much" in English) for male and female genders. This sentence has been recorded in three expressive speech styles of happiness, sadness and reading. Each style contains 200 utterances of samples. Therefore we have 600 utterances of samples for each gender. The parameter extraction tools as used in (Mixdorff and Fujisaki, 1997) are applied in this study.

In each derived parameter, we analyzed the frequency distribution over its range and then the distributions of three expressive speech styles are plot in a graph to show the differences and similarities among those styles. The first seven graphs are of female speech (Fig. 2-8), while the next seven ones are of male speech (Fig. 9-15).

From all of these frequency distribution graphs, the first and second statistical moments (mean and standard deviation values) were subsequently calculated and shown in terms of the following comparative bar charts (Fig. 16-22). From these bar charts, we can also observe some differences between male and female speech.

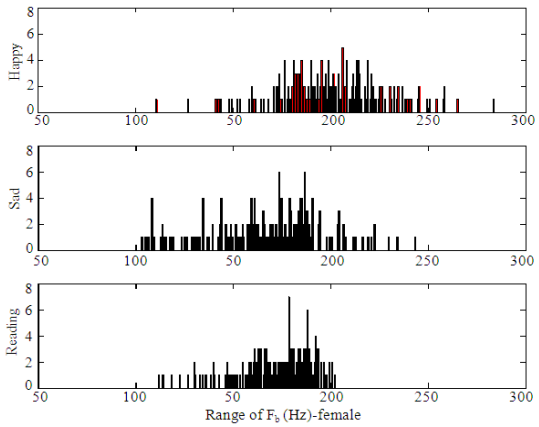

Fig. 2: Comparison of baseline frequency parameter distributions of three styles of female speech

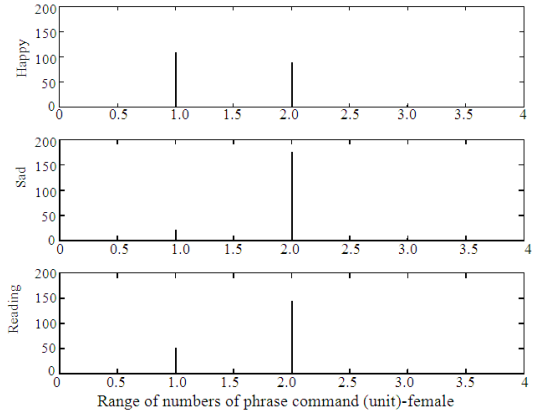

Fig. 3: Comparison of numbers of phrase commands parameter distributions of three styles of female speech

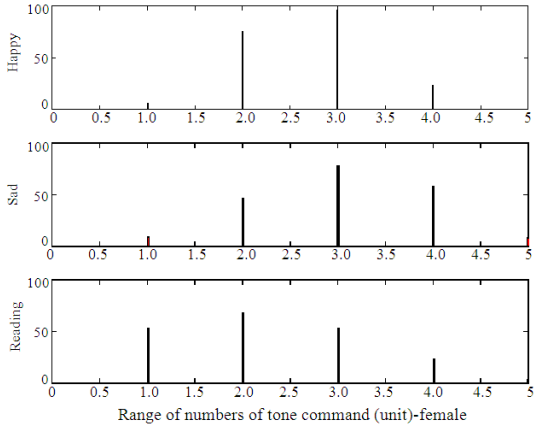

Fig. 4: Comparison of numbers of tone commands parameter distributions of three styles of female speech

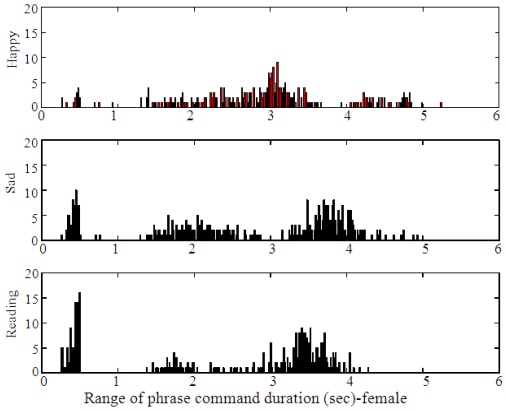

Fig. 5: Comparison of phrase command duration parameter distributions of three styles of female speech

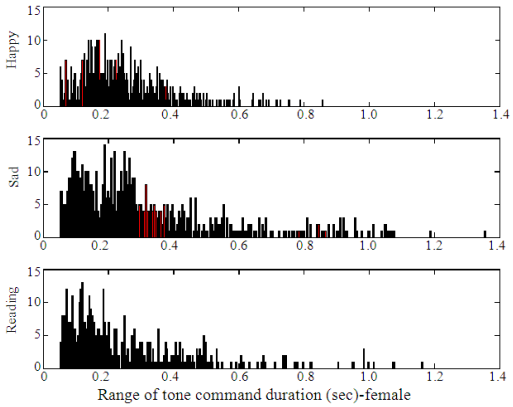

Fig. 6: Comparison of tone command duration parameter distributions of three styles of female speech

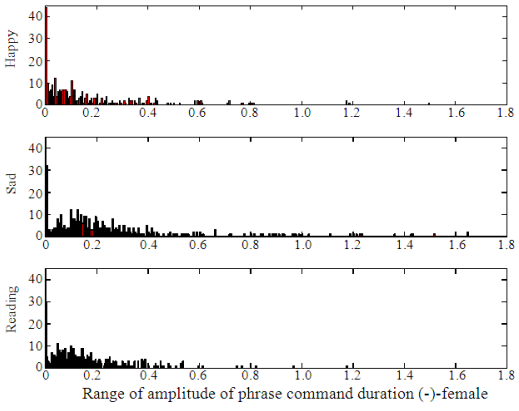

Fig. 7: Comparison of amplitude of phrase command parameter distributions of three styles of female speech

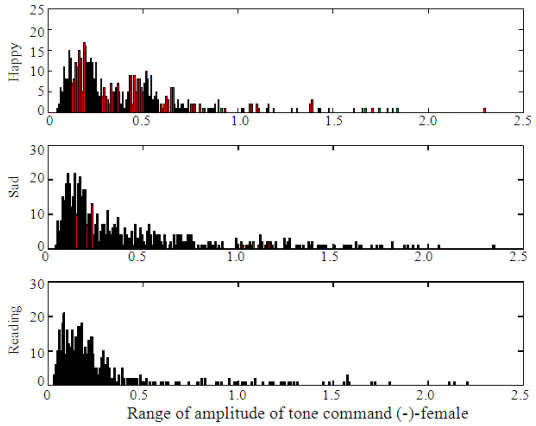

Fig. 8: Comparison of amplitude of tone command parameter distributions of three styles of female speech

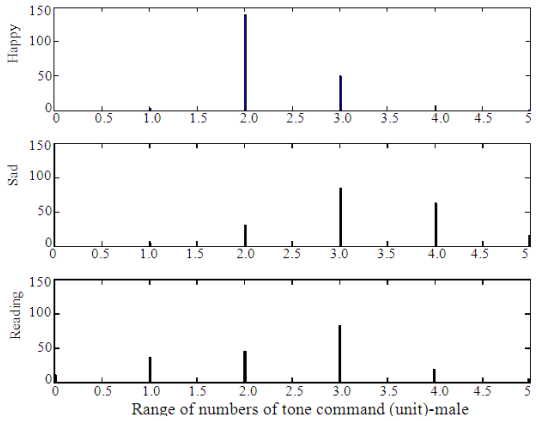

Fig. 11: Comparison of numbers of tone commands parameter distributions of three styles of male speech

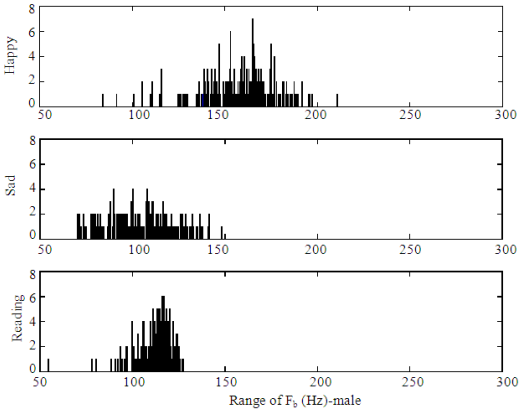

Fig. 9: Comparison of Baseline frequency parameter distributions of three styles of male speech

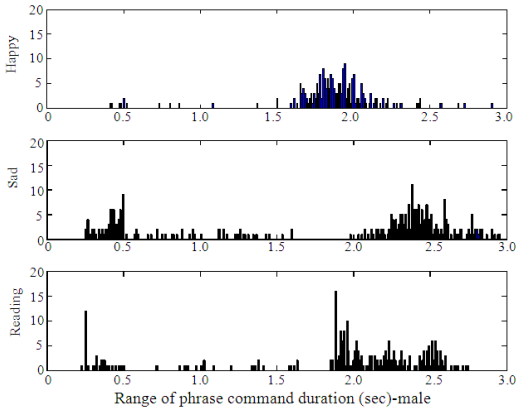

Fig. 12: Comparison of phrase command duration parameter distributions of three styles of male speech

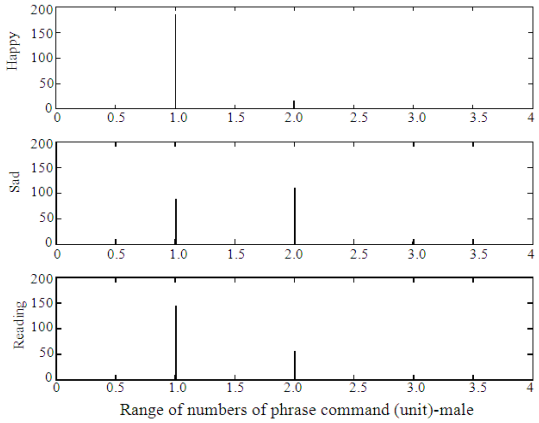

Fig. 10: Comparison of Numbers of phrase commands parameter distributions of three styles of male speech

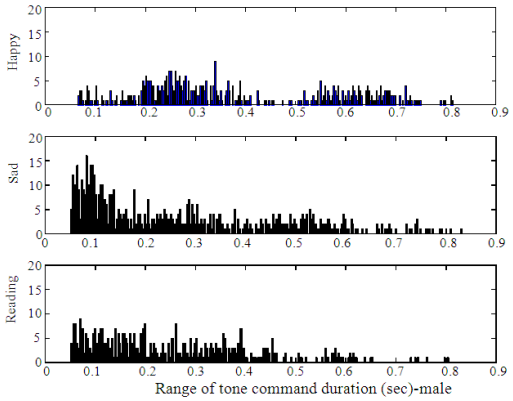

Fig. 13: Comparison of tone command duration parameter distributions of three styles of male speech

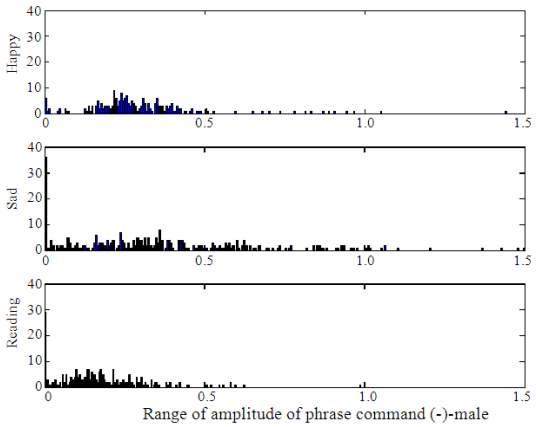

Fig. 14: Comparison of amplitude of phrase command parameter distributions of three styles of male speech

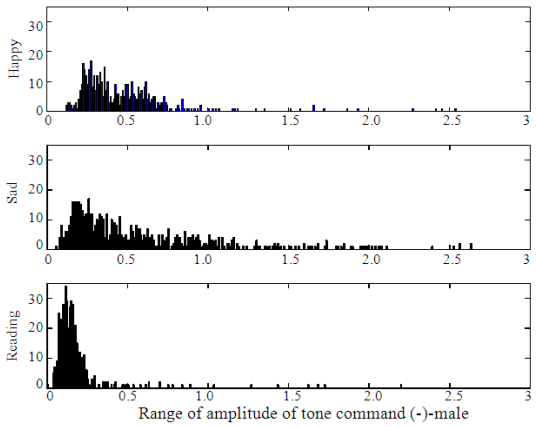

Fig. 15: Comparison of amplitude of tone command parameter distributions of three styles of male speech

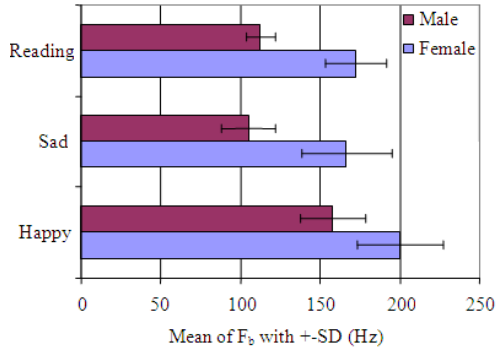

Fig. 16: Comparison of statistical figures of baseline frequency between male and female speech

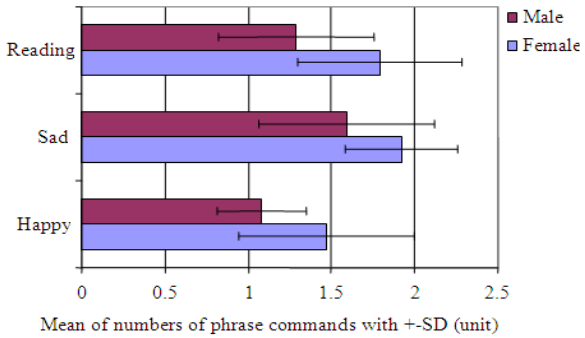

Fig. 17: Comparison of statistical figures of number of phrase commands between male and female speech

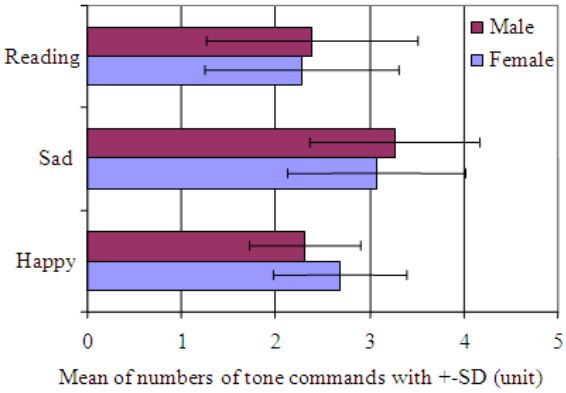

Fig. 18: Comparison of statistical figures of number of tone commands between male and female speech

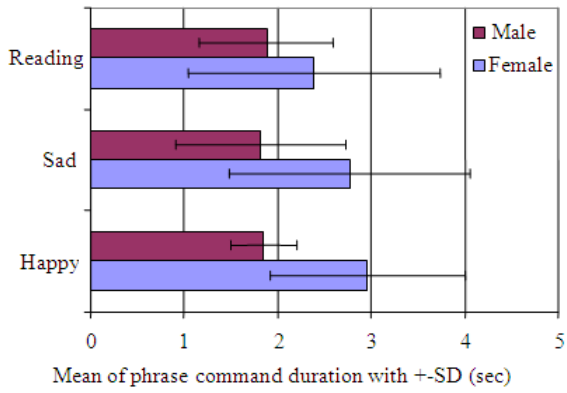

Fig. 19: Comparison of statistical figures of phrase command duration between male and female speech

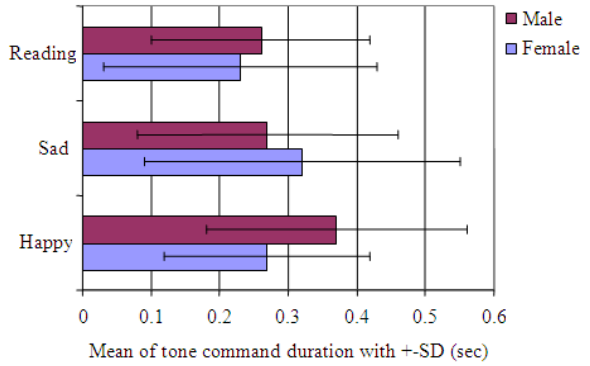

Fig. 20: Comparison of statistical figures of tone command duration between male and female speech

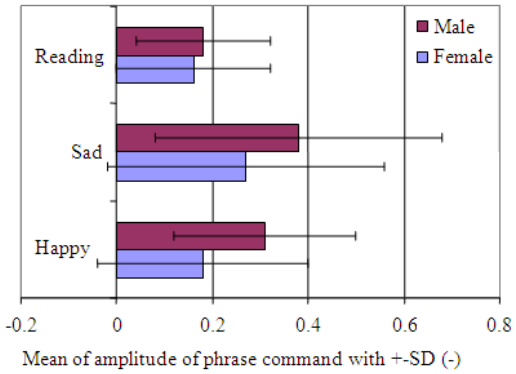

Fig. 21: Comparison of statistical figures of amplitude of phrase command between male and female speech

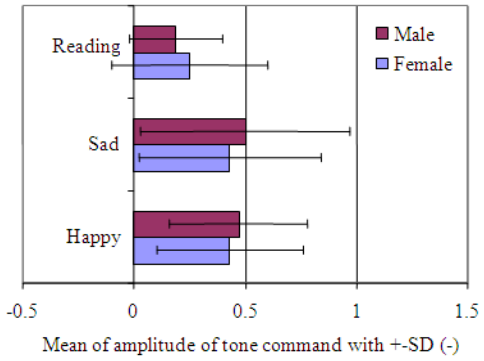

Fig. 22: Comparison of statistical figures of amplitude of tone command between male and female speech

DISCUSSION

From the frequency distribution graphs of male and female speech in Fig. 2-15, most results show that the three distributions of each speaking styles are significantly different. Except for only some cases, one distribution of speaking style is similar to another, i.e., in Fig. 8; the sad and reading styles of amplitude of tone command. It has been noted that some distributions have multi-modals, i.e., in Fig. 5 and 8; the phrase command duration. All in all, in nearly all of the frequency distribution graphs, three distributions of each speaking styles are distinguished from each others empirically.

From the statistical bar charts in Fig. 16-22, they represent the mean and standard deviation values for all seven parameters between male and female speech in comparison. In Fig. 16, 17 and 19; the parameters of baseline frequency, number of phrase commands and phrase command duration, it has been observed that the mean values of male speech for all speaking styles are less than that of female speech. In Fig. 18 and 21; the parameters of number of tone commands and amplitude of phrase command, it has been observed that the mean values of male speech for all speaking styles are higher than that of female speech. In Fig. 19; the parameter of phrase command duration, it has been seen that all speaking styles of male speech have the same level of mean values. However, the other parameters have different levels of mean values for different speaking styles. To distinguish one speaking style from the others, it is needed to use the derived parameters compositely.

From the experimental results, it is a strong evidence to further apply the derived parameters in the speech synthesis systems or other speech processing technologies. For examples, the parameters are expected to be applied in the tree-based context clustering in Thai speech synthesis (Chomphan and Kobayashi, 2007a) to categorize the speech units into groups. The data sharing in each of the speech unit clusters can consequently improve the efficiency of the overall speech synthesis system.

CONCLUSION

This study proposes an analysis of model parameters for Thai speech prosody with three speaking styles and two genders. The Fujisaki's model has been applied to model the F0 contour. The speaking styles of happiness, sadness and reading have been studied. Seven derived parameters from the Fujisaki's model are extracted. The results show that nearly most of the proposed parameters can distinguish three kinds of speaking styles explicitly. From this finding, the parameters are expected to apply in the speech synthesis systems in the future.

ACKNOWLEDGEMENT

The researchers are grateful to Sornthongkam and Inthornchai-eur for providing the speech database.

REFERENCES

Chomphan, S. and T. Kobayashi, 2007a. Design of tree-based context clustering for an HMM-based Thai speech synthesis system. Proceeding of the 6th ISCA Workshop on Speech Synthesis, Aug. 2007, Bonn, Germany, pp: 160-165. http://www.isca-speech.org/archive/ssw6/ssw6_160 .html

Chomphan, S. and T. Kobayashi, 2007b. 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, Antwerp, Belgium, pp: 2849-2852. http://www.isca-speech.org/archive/interspeech_20 07/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

Fujisaki, H. and S. Ohno, 1998. The use of generative model of F0 contours for multilingual speech synthesis. Proceeding of the International Conference on Spoken Language Processing, Dec. 1998, Sydney, Australia, pp: 714-717.

Fujisaki, H., K. Hirose, P. Halle and H. Lei, 1990. Analysis and modeling of tonal features in polysyllabic words and sentences of the standard Chinese. Proceeding of the International Conference on Spoken Language Processing, Nov. 1990, pp: 841-844.

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

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, 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 F0 contours for cantonese text-to-speech. ACM Trans. Asian Language Inform. Process., 3: 169-180. DOI: 10.1145/1037811.1037813

Mixdorff, H. and H. Fujisaki, 1997. Automated quantitative analysis of F0 contours of utterances from a German ToBI-labeled speech database. Proceeding of the Eurospeech, September 22-25, 1997, Rhodes, Greece, pp: 187-190. http://www.isca-speech.org/archive/eurospeech_1997/e97_0187.html

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, 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, Hong Kong, pp: 452-455. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1198815

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, Orlando, USA., pp: 505-508. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1005787

Tao J., J. Yu and W. Zhang, 2006. Internal dependence based f0 model for mandarin tts system. Proceeding of theTC-STAR Workshop on Speech-to-Speech Translation, Jun. 2006, Barcelona, Spain, pp: 171-174. http://www.elda.org/tcstar-workshop_2006/pdfs/tts/tcstar06_tao.pdf

Tachibana, M., J. Yamagishi, T. Masuko and T. Kobayashi, 2005. Speech synthesis with various emotional expressions and speaking styles. IEICE Trans. Inf. Syst., E88-D: 2484-2491. http://ci.nii.ac.jp/naid/110003501992

Tachibana, M., J. Yamagishi, T. Masuko and T. Kobayashi, 2006. A Style Adaptation Technique for Speech Synthesis Using HSMM and Suprasegmental Features. IEICE Trans. Inf. Syst., E89-D: 1092-1099. http://ci.nii.ac.jp/naid/110004719385

Tran, D.D., E. Castelli, X.H. Le, J.F. Serignat and V.L. Trinh, 2006. Linear F0 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: 1-4. http://www-mrim.imag.fr/publications/2006/XUA06/TAL2006SubmissionUpdate.pdf