Journal of Computer Science 6 (11): 1263-1271, 2010
ISSN 1549-3636

Fujisaki’s Model of Fundamental Frequency Contours for Thai Dialects

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 general, there are a number of rural dialects in Thai. However, four dialects are mainly spoken by Thai people residing in four core region including central, north, northeast and south regions. Recognizing and synthesizing Thai speech with different dialects are consequently difficult. **Approach:** 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. To treat the problem, the speech prosody is carefully preserved through modeling the fundamental frequency (F0) contours. The differences among the model parameters of four Thai dialects have been summarized. This study proposed an analysis of model parameters for Thai speech prosody with four regional dialects and two genders which is a preliminary work for speech recognition and synthesis. Fujisaki’s modeling; a powerful tool to model the F0 contour has been adopted. 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 regional dialect includes 200 samples of one sentence with male and female speech. Therefore our speech database contains 1600 utterances in total. The results showed that most of the proposed parameters can distinguish four kinds of regional dialects explicitly. **Conclusion:** By using the Fujisaki’s model, the results confirm that the proposed parameters can distinguish the regional dialects efficiently. In the future research, they were expected to be applied in the speech recognition and synthesis with various regional dialect characteristics.

Key words: Thai dialects, Fujisaki’s model, fundamental frequency

INTRODUCTION

An appropriate modeling of F0 contour contributes the effectiveness in speech processing, such as speech recognition, speech synthesis and speech coding. Fujisaki’s modeling of fundamental frequency for Thai expressive speech conducted in 2010 is proved to be effective for a limited-domain speech corpus (Chomphan, 2010). It has been found that the derived parameters can distinguish one style of speech from each other.

As for speech processing of Thai dialects, it has not been studied despite of a variety of the dialects spreading over four regions of Thailand. Beginning from the Northern region of Thailand, Thai dialect of “Lanna” or “Kammuang” is widely used, Lao-style Thai dialect is spoken in the North Eastern region, meanwhile South Thai dialect is spoken generally in the Southern part of Thailand.

By using the same way of Thai expressive speech (Chomphan, 2010), the study proposes an analysis of F0 modeling of four Thai dialects including standard Thai, Lanna or North dialect, Lao-style or North East dialect and South dialect. The extension of 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; 2006) is mainly used.

MATERIALS AND METHODS

Fujisaki’s model: The F0 contour of an utterance of speech is treated as a linear superposition of a global phrase and local accent components on a logarithmic scale, as depicted in Fig. 1 (Fujisaki and Sudo, 1971).

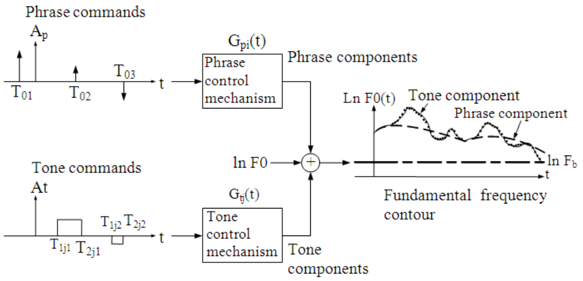


Fig. 1: An extension of Fujisaki’s model for the generation of F0 contour

Two commands generate the corresponding components of global phrase and local accent components. 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):

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

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

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

Where:

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

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

The symbols in the above three equations denote that Fb is the smallest F0 value in the F0 contour of interest and A_{pi} and At, jk are the amplitudes of the i-th phrases and of the j-th tone command. Here, T0i 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.

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

By using this generative 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 calculated 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 these derived parameters have been extracted for four regional Thai dialects including standard Thai, Lanna or North dialect, Lao-style or North East dialect and South dialect.

RESULTS

In our speech database, we use a sentence of “จินตนาการสำคัญกว่าความรู้” in IPA (means “Imagination is more important than Knowledge” in English) for male and female genders. This sentence has been recorded in four Thai dialects of standard Thai, Lanna Thai dialect, Lao-style Thai dialect and South Thai dialect. Each dialect contains 200 utterances of samples. Therefore we have 800 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 four Thai dialects are plot in a graph to show the differences and similarities among those dialects.

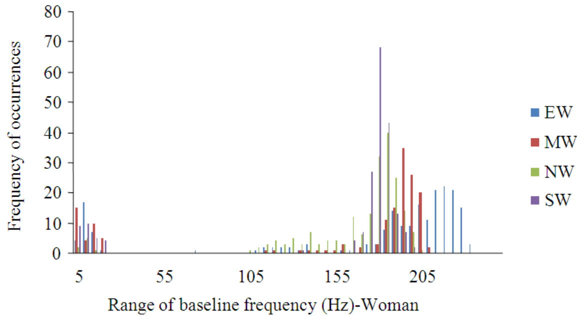


Fig. 2: Comparison of baseline frequency parameter distributions of four Thai female dialects

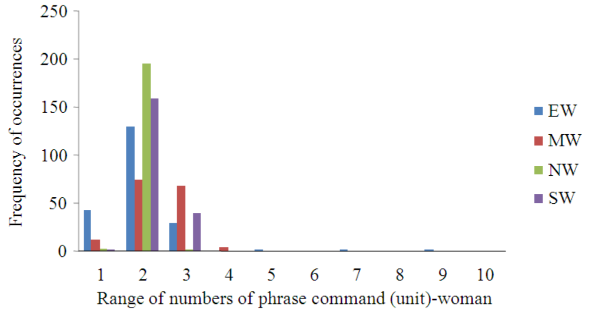


Fig. 3: Comparison of numbers of phrase commands parameter distributions of four Thai female dialects

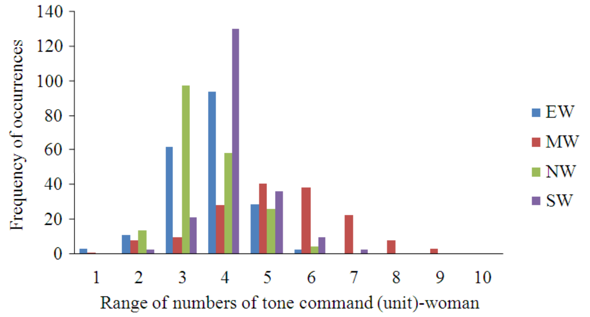
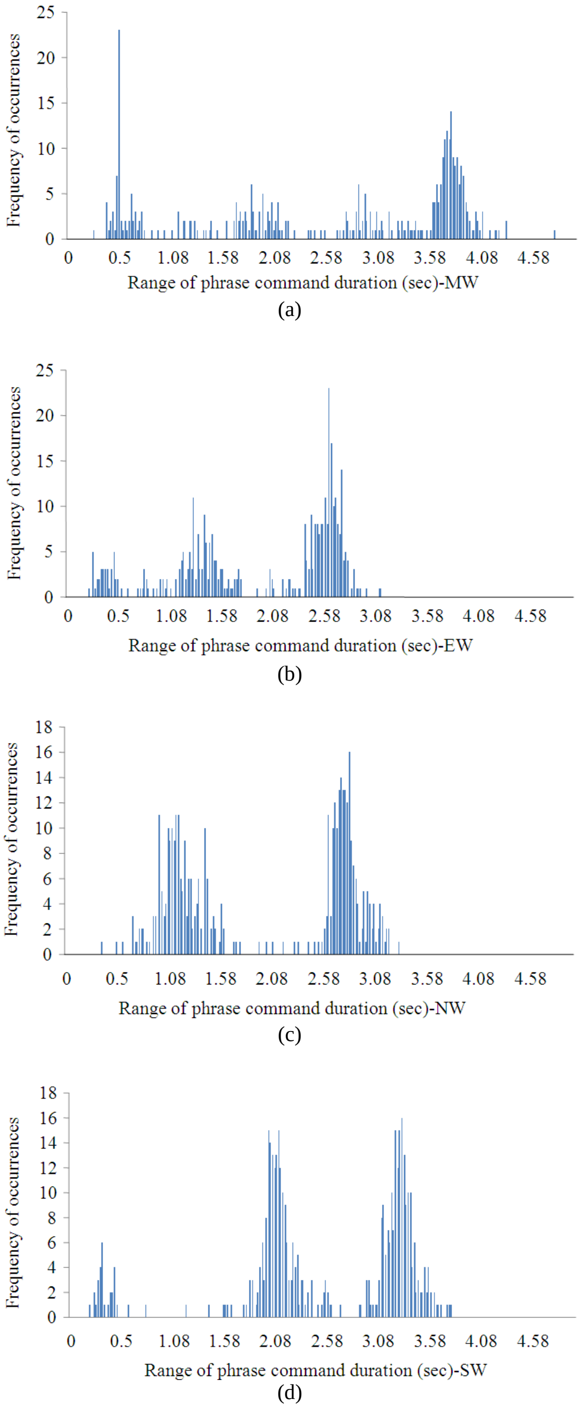


Fig. 4: Comparison of numbers of tone commands parameter distributions of four Thai female dialects

The first seven graphs are of female speech (Fig. 2-8), while the next seven ones are of male speech (Fig. 9-15). The following abbreviation are defined and used in most figure, EW, MW, NW and SW denote North East woman dialect, Standard Thai woman dialect, North woman dialect and South woman dialect, respectively, while EM, MM, NM and SM denote North East man dialect, Standard Thai man dialect, North man dialect and South man dialect, respectively.



(a)

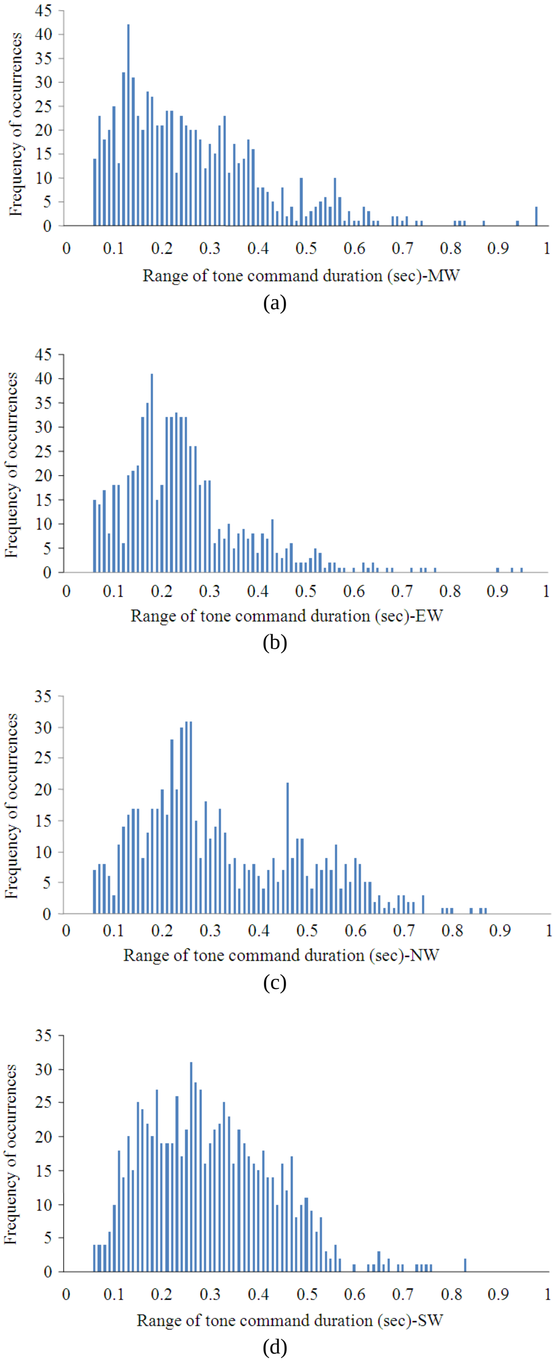
(b)

(c)

(d)

Fig. 5: Comparison of phrase command duration parameter distributions of four Thai female dialects

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.



(a)

(b)

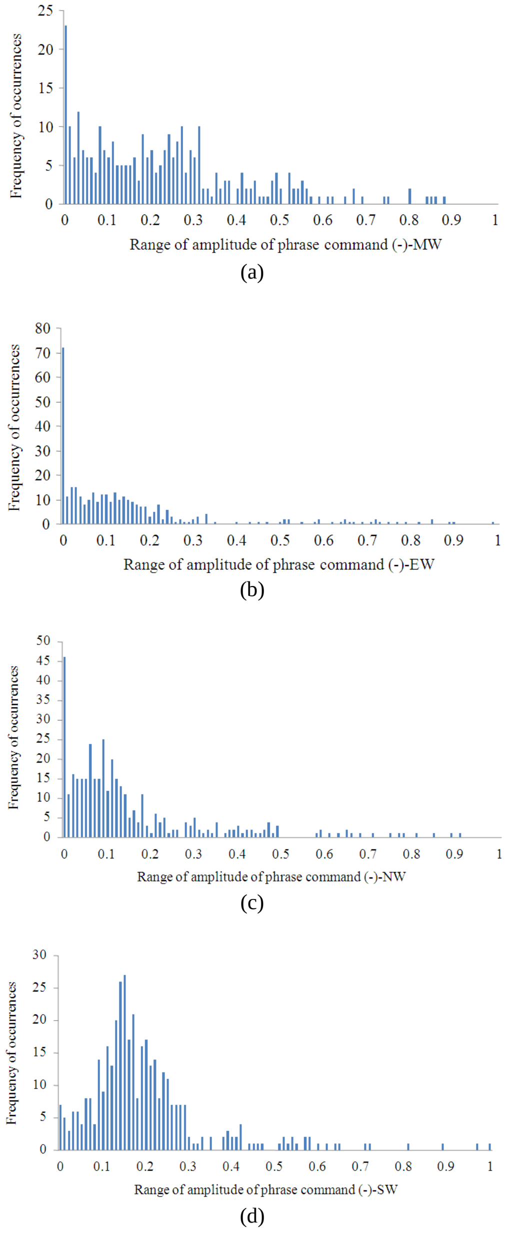
(c)

(d)

Fig. 6: Comparison of tone command duration parameter distributions of four Thai female dialects

DISCUSSION

From the frequency distribution graphs of male and female speech in Fig. 2-15, most results show that the four distributions of Thai dialects are significantly different. Except for only some cases, one distribution of Thai dialect is similar to another, i.e., in Fig. 8; the North East and North dialects of amplitude of tone command.



(a)

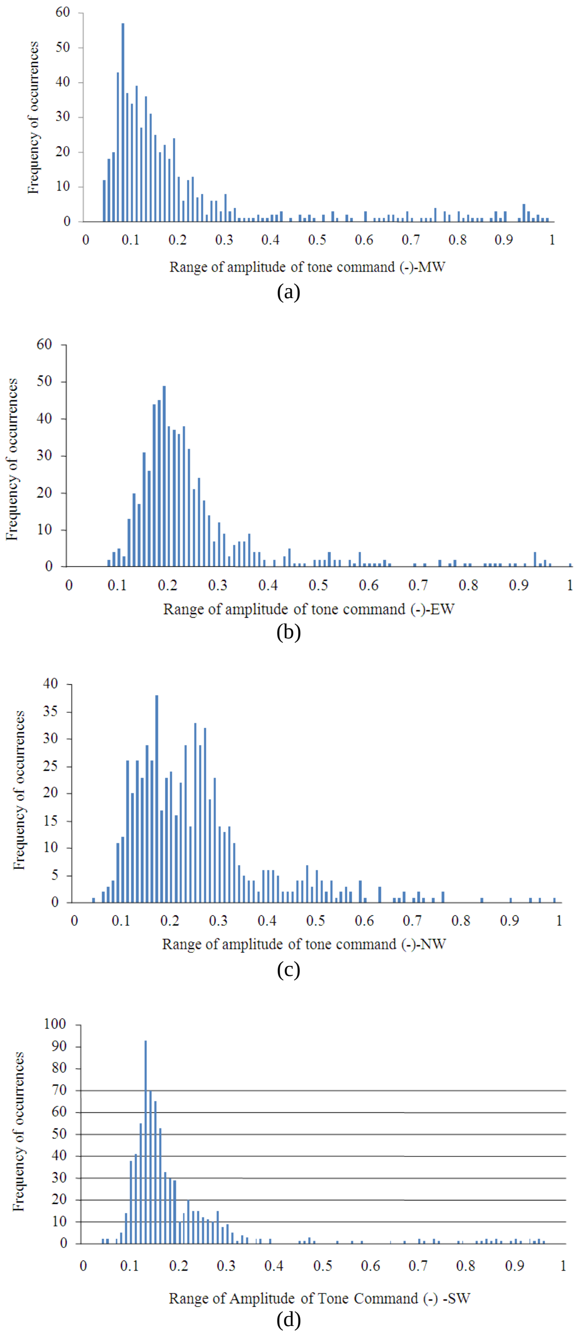
(b)

(c)

(d)

Fig. 7: Comparison of amplitude of phrase command parameter distributions of four Thai female dialects

It has been noted that some distributions have multi-modals, i.e., in Fig. 5 and 12; the phrase command duration. All in all, in nearly all of the frequency distribution graphs, four distributions of dialects are distinguished from each others empirically.



(a)

(b)

(c)

(d)

Fig. 8: Comparison of amplitude of tone command parameter distributions of four Thai female dialects

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. 16-18; the parameters of baseline frequency, number of phrase commands and number of tone commands, it has been observed that the mean values of male speech for all dialects are less than that of female speech.

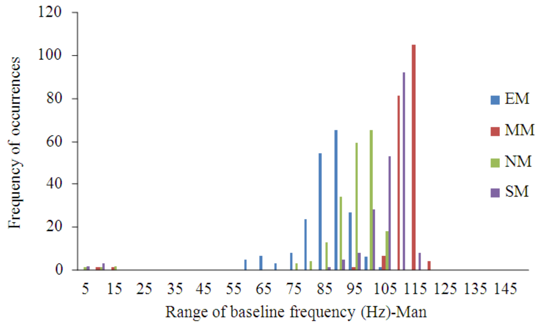


Fig. 9: Comparison of Baseline frequency parameter distributions of four Thai male dialects

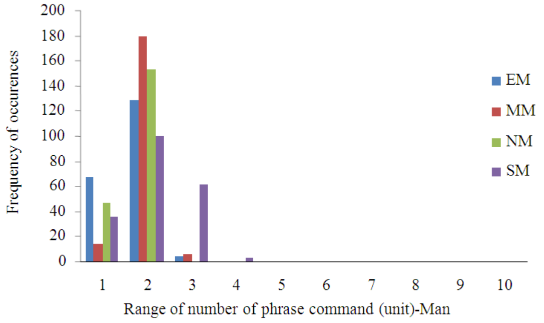


Fig. 10: Comparison of Numbers of phrase commands parameter distributions of four Thai male dialects

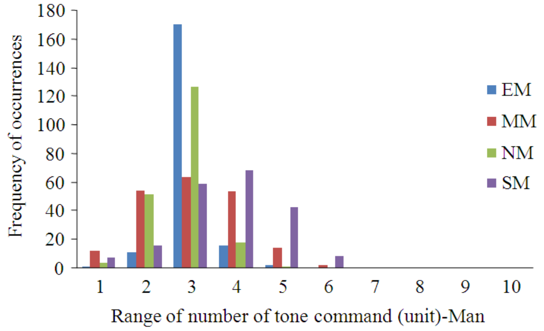
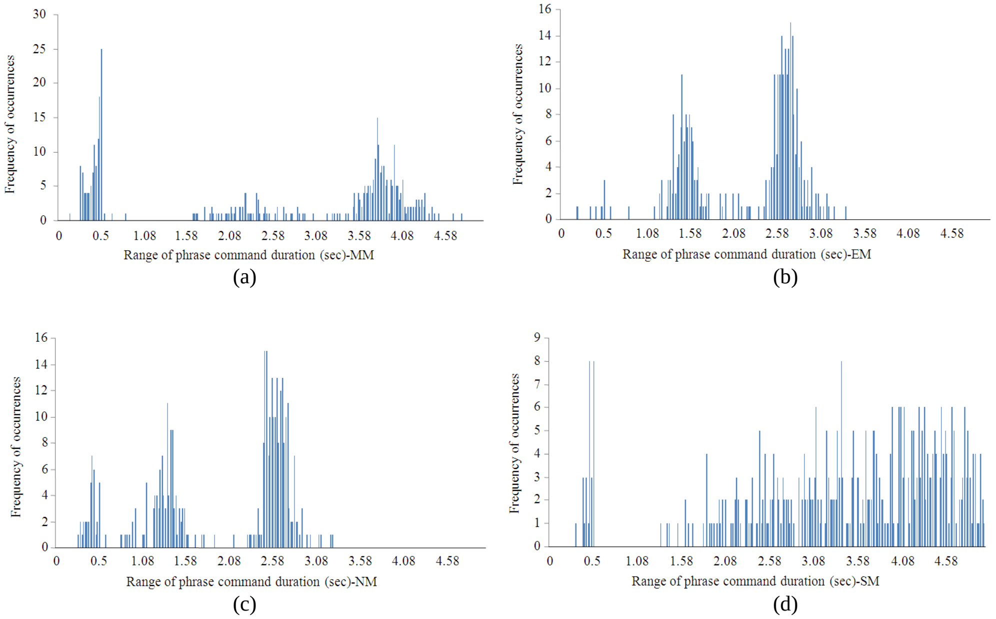


Fig. 11: Comparison of numbers of tone commands parameter distributions of four Thai male dialects

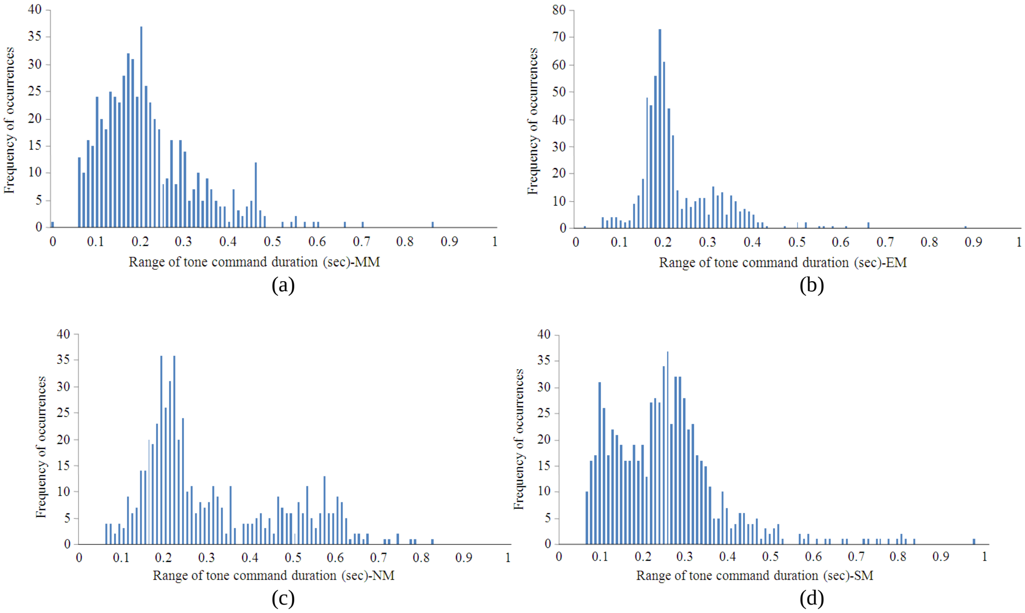
In Fig. 19 and 20; the parameters of phrase command duration and tone command duration, it has been observed that the mean values of male speech for dialects are quite similar to that of female speech. In Fig. 21 and 22; the parameter of amplitude of phrase command and amplitude of tone command, it has been observed that the standard Thai has a significant difference between male and female speech, while the others have some small differences. In the classification problem, one dialect can be distinguished from the others by using the derived parameters compositely.



(a) (b)

(c) (d)

Fig. 12: Comparison of phrase command duration parameter distributions of four Thai male dialects



(a) (b)

(c) (d)

Fig. 13: Comparison of tone command duration parameter distributions of four Thai male dialects

From the above experimental results, it is a strong evidence to further apply the derived parameters in the speech synthesis systems or other speech processing technologies. These experimental results correspond to the previous results conducted for Thai expressive speech (Chomphan, 2010). For examples, the parameters are expected to be applied in the tree-based context clustering process in the hidden Markov model based Thai speech synthesis (Chomphan and Kobayashi, 2007a; 2007b; 2008; 2009) to categorize the training speech units into groups. An appropriate data sharing in each of the speech unit clusters can consequently improve the efficiency of the overall speech synthesis system.

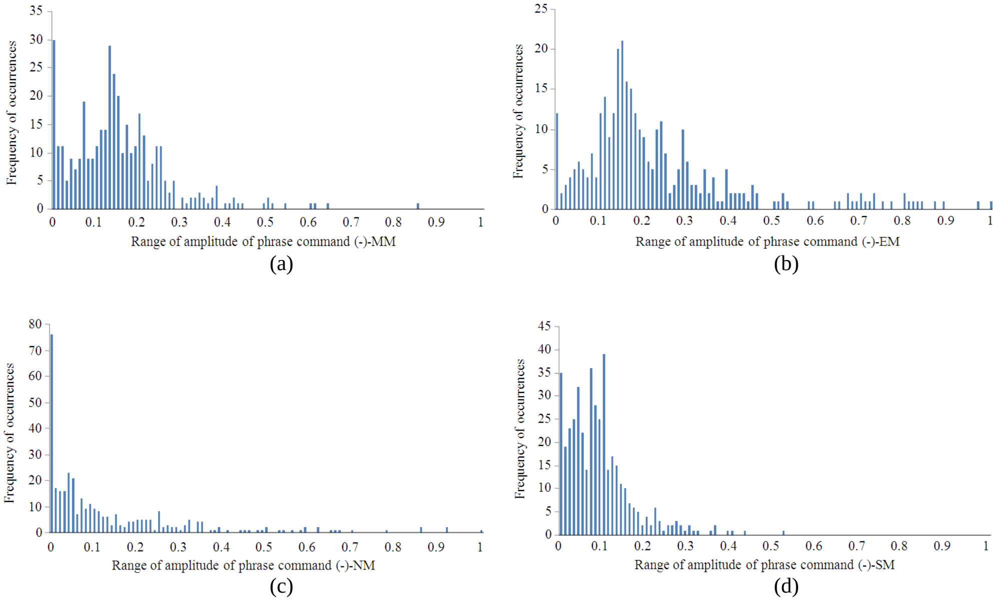


Fig. 14: Comparison of amplitude of phrase command parameter distributions of four Thai male dialects

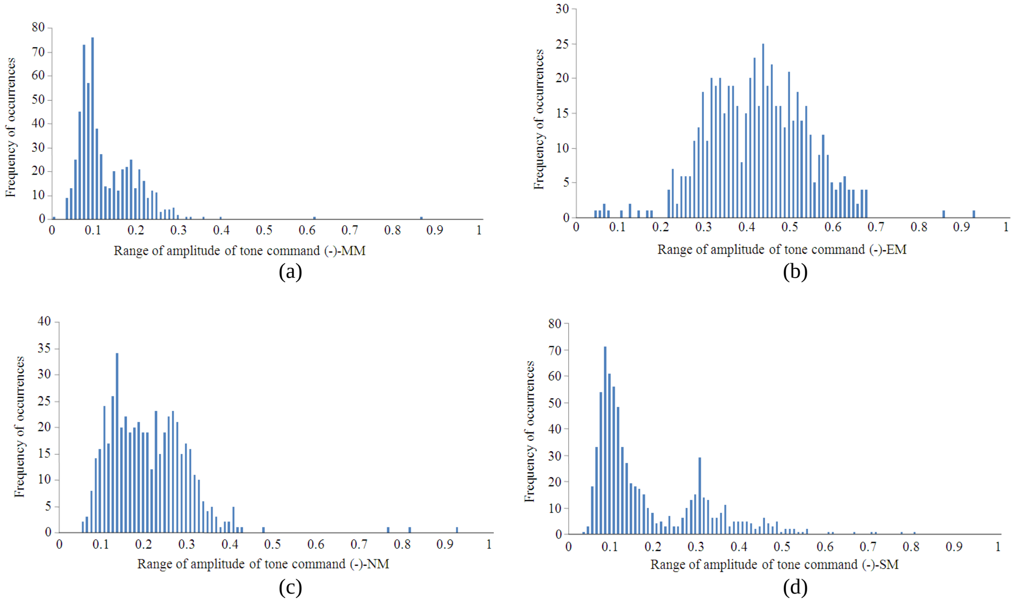


Fig. 15: Comparison of amplitude of tone command parameter distributions of four Thai male dialects

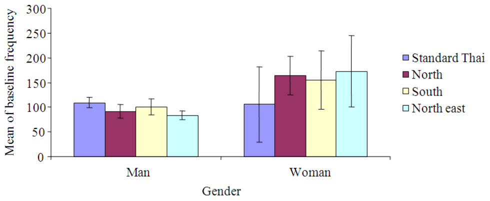


Fig. 16: Comparison of statistical figures of baseline frequency between male and female speech

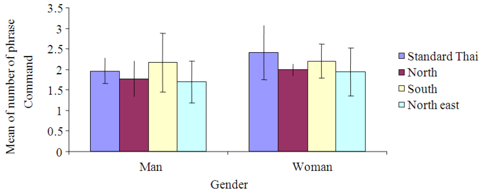


Fig. 17: Comparison of statistical figures of number of phrase commands between male and female speech

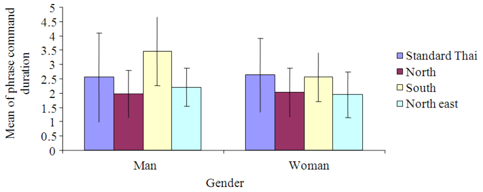


Fig. 18: Comparison of statistical figures of number of tone commands between male and female speech

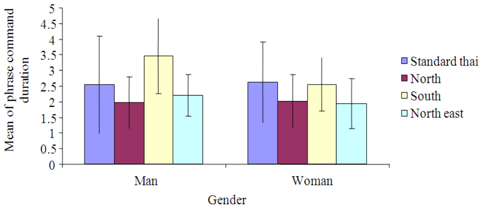


Fig. 19: Comparison of statistical figures of phrase command duration between male and female speech

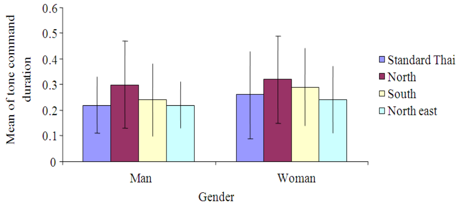


Fig. 20: Comparison of statistical figures of tone command duration between male and female speech

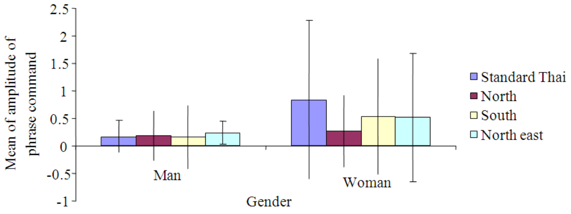


Fig. 21: Comparison of statistical figures of amplitude of phrase command between male and female speech

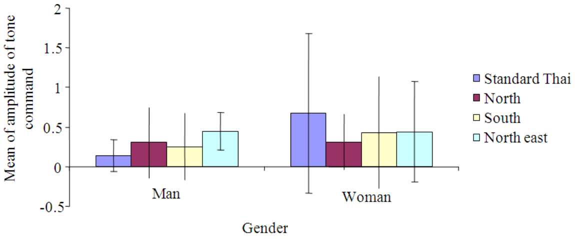


Fig. 22: Comparison of statistical figures of amplitude of tone command between male and female speech

CONCLUSION

An analysis of Fujisaki's model parameters for four Thai dialects has been performed in this study. The specified dialects include standard Thai, North dialect, North East dialect and South dialect, meanwhile the speech database covers both male and female genders. The Fujisaki's model has been applied to model the F0 contour of all dialects. Seven derived parameters from the Fujisaki's model are extracted. The experimental results show that most of the derived parameters can be used to distinguish all four Thai dialects explicitly. From this finding, the parameters are expected to further apply in the speech recognition and speech synthesis systems.

ACKNOWLEDGEMENT

The researchers are grateful to T. Tintanee and W. Leeparikon 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. 22.24, ISCA, 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. 27-31, 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., 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

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, Sept. 22-25, ISCA, Rhodes, Greece, pp: 187-190. http://www.isca-speech.org/archive/eurospeech_1997/e97_0187.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. 6-10, IEEE Xplore Press, Hong Kong, pp: 452-455. DOI: 10.1109/ICASSP.2003.1198815

Tachibana, M., J. Yamagishi, T. Masuko and T. Kobayashi, 2005. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing: Life-like agent and its communication. IEICE Trans. Inform. 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. Inform. Syst., E89-D: 1092-1099. http://ci.nii.ac.jp/naid/110004719385