American Journal of Applied Sciences 9 (3): 293-299, 2012
ISSN 1546-9239


# Thai Expressive Speech Processing Technology: A Review

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:** The studies on Thai expressive speech or emotional speech have been conducted for years. Most of them are expected to analysis the characteristics of Thai expressive speech. However, the conclusive reviews on these studies have not been conducted for further study on the speech technology or application of Thai expressive speech. **Approach:** The review of research on Thai expressive speech in various aspects has been performed. They include an analysis of fundamental frequency contours using Fujisaki's model, an analysis of fundamental frequency contours using structural model and speech compression with noisy environments. It has been noted that four speaking emotions include enjoyable, sad, angry and reading styles. **Results:** A comparison of two successful $F_0$ models has been reviewed. One approach is based on the Fujisaki's model which has been applied for many tonal and toneless languages. Another one is based on the structural model which has been conducted primarily for Mandarin Chinese. Moreover, a study of speech compression for noise-corrupted Thai expressive speech by using two coding methods of CS-ACELP and MP-CELP has been summarized. **Conclusion:** From the study, it can be seen that two mathematical models have been successfully applied to model the fundamental frequency contour of Thai expressive speech. As for speech compression, it can be seen that coding methods, types of noise, levels of noise, speech gender influence on the coding speech quality.

**Key words:** Linear Prediction (LP), fundamental frequency contours, speech gender influence, coding speech quality, expressive speech, fundamental frequency, rhythmic structures, Conjugate Structure Algebraic Code Excited Linear Predictive (CS-ACELP)

## INTRODUCTION

The expressive speech or emotional speech is the current challenge in the modern speech technology. The speech communication with implicit emotional information is in the frontier line of speech technology research. However, in Thai language, the expressive speech related research is in the beginning phase. The previous study in this issue is reviewed in this study.

The fundamental frequency of speech is the most important feature among all of the features known to carry prosodic information which is an inherently supra-segmental feature of human speech. The $F_0$ contours of an utterance or a sentence convey the stress, intonation and rhythmic structures, which indicates the naturalness and intelligibility of synthetic speech. Therefore, the appropriate modeling of $F_0$ contour plays an important role in the speech technology area, e.g., speech recognition, speech synthesis, speech analysis and speech coding. A number of modeling techniques in the previous studies have been conducted 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 (Hiroya and Hiroshi, 1971; Tran *et al*., 2006). In Thai, 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 area, the statistical modeling of $F_0$ contour has been achieved by Chomphan and Kobayashi (2007; 2008; 2009) and Chomphan (2009) in the implementation of speaker-dependent and speaker-independent systems during 2007-2009. Recently, the Fujisaki's model has been applied within a speaker-independent system as extended modules. Moreover, it has been applied in the modeling of Thai expressive speech; i.e., sad, happy, angry styles (Chomphan, 2010a; 2010c; 2010e). Moreover, another study has been conducted by using a structural model that 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; Chomphan, 2010e; 2010f). The RMS error calculation has been done for evaluation the modeling performance for both mentioned speech models and also for all speech

styles including angry style, sad style, enjoyable style and reading style. This study aims at comparing the Fujisaki's model and the structural model.

In the digital speech communication, low bitrate speech compression or coding is highly required to increase the channel capacity. The flexibility of coding rate is also needed to support the variety of the traffic occupancies depending on the type and number of users. In 1995, CS-ACELP coding was initially developed and standardized as ITU G.729 speech coding at the constant bitrate of 8 kbps. Few years later, MP-CELP coding has been developed to be a scalable coder. In the MP-CELP speech coder, it operates at various bitrates ranging from 4-12 kbps utilizing the flexibility in multi-pulse excitation coding (Chomphan, 2010b; 2010d; 2011a; 2011b; 2011c).

In this study, the review of research on Thai expressive speech in various aspects has been done. They include an analysis of fundamental frequency contours using Fujisaki's model, an analysis of fundamental frequency contours using structural model and speech compression with noisy environments using CS-ACELP and MP-CELP speech codes. Four speaking emotions of enjoy, sadness, anger and reading are selected in this study.

## MATERIALS AND METHODS

**Fujisaki's model:** The $F_0$ contour is treated as a linear superposition of a global phrase and local accent components on a logarithmic scale, as shown in Fig. 1

The phrase command generates a phrase component, while the accent command generates the accent component of the $F_0$ contour. We use two parameters of the Fujisaki's model as our phrase-intonation features including the baseline value of $F_0$ and the magnitude of the phrase command, which complementarily reflect the global level of voicing frequency. In mathematical formula, the $F_0$ contour of an utterance generated from an extension of the Fujisaki's model for tonal languages is represented as the following expressions Eq. 1-3 (Chomphan, 2010f):

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

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

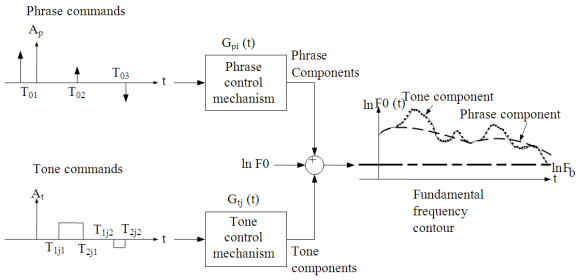


Fig. 1: An extension of Fujisaki's model for the generation of $F_0$ contour

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

Where:

$G_{pi}(t)$ = 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 these equations denote that $F_b$ is the minimum value in the $F_0$ contour of interest and $A_{pi}$ and $A_{t,jk}$ are the amplitudes of the i-thphrases and of the j-tr tone command. Moreover, $T_{0i}$ is the timing of the i-th phrase command and $T_{1jk}$ and $T_{2jk}$ are the onset and offset of the k-tr component of the j-tr tone command. While $a_i$ and $\beta_{jk}$ are time constant parameters, subsequently I, J, K (j) denote the number of phrases, tones and components of the j-tr tone of the utterance.

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

From the conventional parameters as described in earlier, seven derived parameters which reflect the geometrical appearance of the $F_0$ contour of an utterance are selected 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

These parameters have been extracted for four speech expressions of angry style, sad style, enjoyable

style and reading style. Subsequently, the extracted parameters are used to resynthesize the $F_0$ contour in the evaluation process.

**Structural model:** The $F_0$ contour is modeled in a logarithmic scale, as depicted in Fig. 2. The mathematical model has been applied (Ni and Hirose, 2006; Chomphan, 2010e; 2010f) 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 length, mass 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 × (MacNeilage, 1983), which can be formulated in the following mathematical term Eq. 4:

$$\ln f_0 = \frac{b}{2}x + \ln(\sqrt{a}c_0) \quad (4)$$

where, a, b and $c_0$ are constant coefficients.

The flow chart in Fig. 3 shows the main process for evaluating the structural model. At first the speech corpus has been implemented. It consisted of male and female speech utterances. Each of them has four speech styles including happy, sad, angry and reading styles. Each style consists of 5 sentences with 100 samples of utterances. Therefore the 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 have been used as its representatives. Thereafter, the resynthesis of $F_0$ contour from the parameters has been performed. Subsequently, the RMS error between the natural $F_0$ contour and the resynthesized $F_0$ contour has been calculated. Finally, the summarized data from the previous stages has been analyzed.

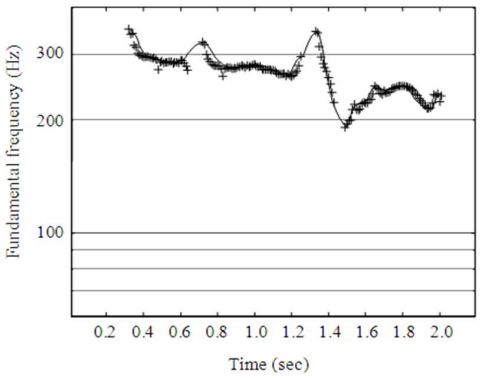


Fig. 2: $F_0$ contour with a trend line in a logarithmic scale

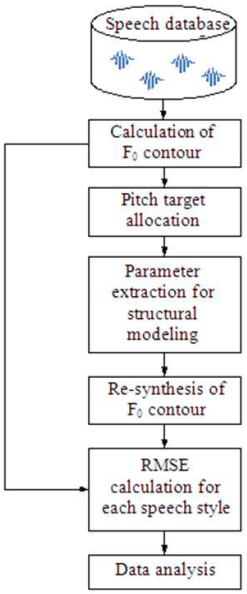


Fig. 3: Workflow for the experimental process

**CS-ACELP algorithm:** The CS-ACELP coder is improved from the conventional Code-Excited Linear Predictive (CELP) coding algorithm (Chomphan, 2011a; 2011c). The encoder extracts the speech features from the speech frames of 10 ms corresponding to 80 samples at a sampling rate of 8000 Hz. The extracted parameters of the CELP model include the linear-prediction filter coefficients, adaptive and fixed-codebook indices and gains. They are subsequently encoded and transmitted through the channel. When they arrive at the decoder, these parameters are used to retrieve the excitation sequence and the synthesis filter parameters. The synthesized speech is reconstructed by filtering this excitation sequence through the short-term synthesis filter based on a 10th order linear prediction filter and the long-term

synthesis filter using an adaptive codebook. Output from the synthesis filter is enhanced by filtering at a post processing unit.

The block diagram of CS-ACELP encoder is shown in Fig. 4. The input signal is high-pass filtered and scaled in the pre-processing unit. The Linear Prediction (LP) analysis is performed for the speech frame of 10-ms length. The obtained LP coefficients are subsequently transformed into Line Spectrum Pairs (LSP) and then quantized using predictive two-stage vector quantization. The excitation is chosen by applying an analysis-by-synthesis search procedure in which the error minimization between the original speech and the reconstructed speech is performed.

The block diagram of CS-ACELP decoder is shown in Fig. 5. At first, the parameter indices are extracted from the received bitstream. Subsequently, they are decoded to retrieve the coder parameters for every 10 m sec speech frame. The synthesized speech is reconstructed by filtering the excitation through the LP synthesis filter. The reconstructed speech signal is finally filtered at a post-processing unit which includes an adaptive post-filter, a high-pass filter and a scaling operation.

MP-CELP algorithm: The principle concepts for the bitrate scalable MP-CELP coder are explained in 2 parts of a core coder and a bitrate scalable tool (Chomphan, 2011a; 2011b; 2011c). The core coder obtains the high coding performance by applying a multi-pulse vector quantization as depicted in Fig. 6 (Ozawa *et al*., 1997; Taumi *et al*., 1996). The input speech of a 10 m sec frame length is analyzed at the LPC analysis module. The obtained LP coefficients are quantized in the LSP domain. The corresponding pitch delay is simultaneously encoded by using an adaptive codebook. The residual signal for LP and the pitch information is encoded by the multi-pulse excitation scheme. The multi-pulse excitation consists of several non-zero pulses. Their pulse positions are restricted in the algebraic-structure codebook and calculated by an analysis-by-synthesis scheme, e.g., (Laflamme *et al*., 1991). The pulse positions and signs are then encoded, while the gains for pitch predictor and the multi-pulse excitation are normalized by the frame energy and also encoded. Three stages of the bitrate scalable tools are conducted. It is embedded adjacently to the core coder as depicted in Fig. 7. A bitrate scalable tool encodes the residual signal from the core coder utilizing the multi-pulse vector quantization. An adaptive pulse position control is conducted to change the algebraic-structure codebook at each excitation-coding stage depending on the encoded multi-pulse excitation at the previous stage.

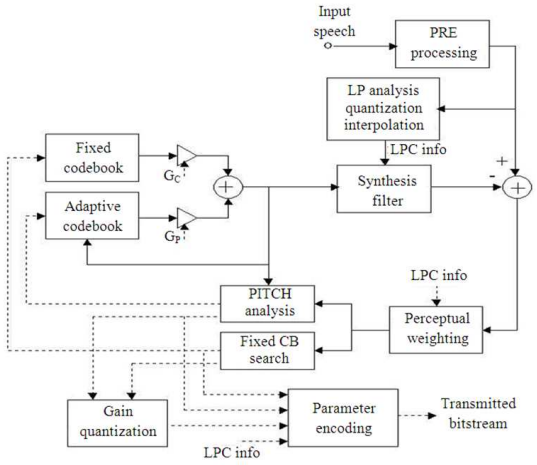


Fig. 4: Block diagram of CS-ACELP encoder

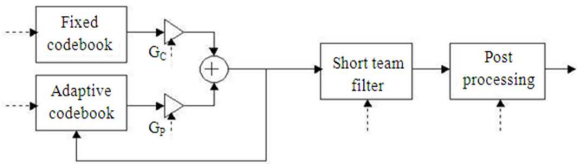


Fig. 5: Block diagram of CS-ACELP decoder

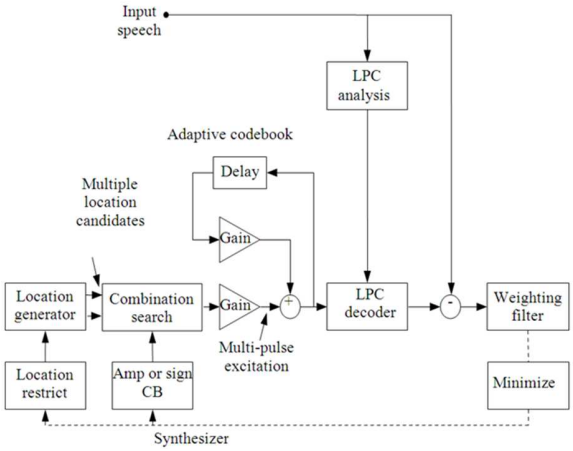


Fig. 6: Block diagram of MP-CELP core coder

The algebraic-structure codebook is adaptively controlled to prevent the occurrence of the same pulse positions as those of the multi-pulse excitation in the core coder or the previous stage. The pulse positions are chosen so that the perceptually weighted distortion between the residual signal and output signal from the scalable tool is minimized. The LP synthesis and perceptually weighted filters are the same as that of the core coder (Chomphan, 2011a; 2011b; 2011c).

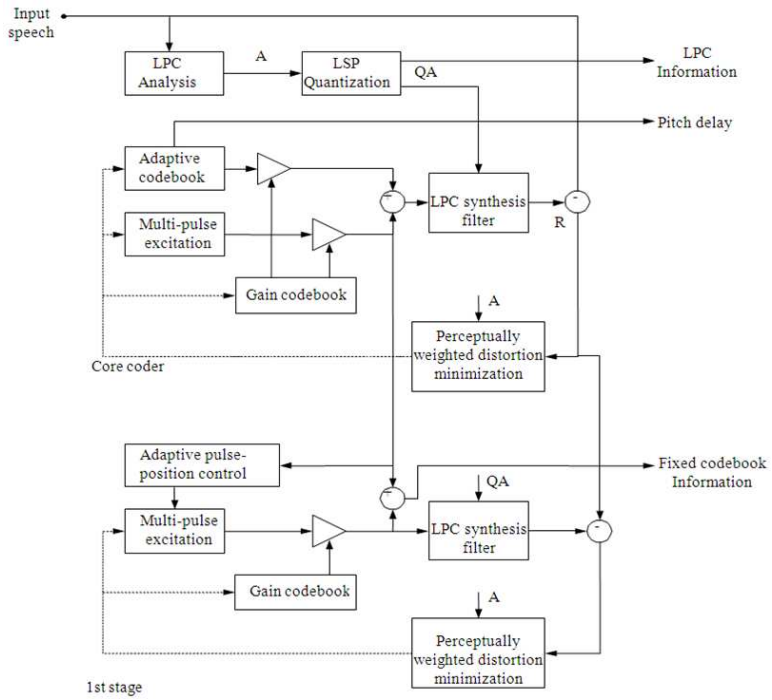


Fig. 7: Block diagram of one-stage bitrate scalable MP-CELP coder

## RESULTS

From the comparison of two successful $F_0$ models; Fujisaki's model and structural model, the applied speech database consists of male and female speech and each one contains 4 different speech styles including angry style, sad style, enjoyable style and reading style. Five sentences are applied for each speech style and each sentence includes 100 samples. It has been seen from the results that RMS error of each speech style is different from the others for both models. Moreover, it reveals that the RMS error of the Fujisaki's model is higher than that of the structural model for all speech styles. In other words, the structural model gives the better fit for modeling of the $F_0$ contour of the expressive speech than that of the Fujisaki's model (Chomphan, 2011d).

From the study of speech compression for noise-Corrupted Thai expressive speech by using two coding methods of CS-ACELP and MP-CELP, the former experimental results show that CS-ACELP gives the better speech quality than that of MP-CELP at all three bitrates (Chomphan, 2011b). When considering the levels of noise, the 20-dB noise gives the best speech quality, while 0-dB noise gives the worst speech quality. When considering the speech gender, female speech gives the better results than that of male speech. Finally, when considering the types of noise, the air-conditioner noise gives the best speech quality, while the train noise gives the worst speech quality (Chomphan, 2011b).

## DISCUSSION

From the comparison of two successful $F_0$ models; Fujisaki's model and structural model, it has been concluded from the previous study 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. When considering the differences between genders, it has been found that the averaged RMS error of female speech is above that of male speech. Moreover, the RMS error of the Fujisaki's model is mostly higher than that of the structural model for all speech styles. In other words, it can be concluded that the structural model gives the better fit for modeling of the $F_0$ contour of the expressive speech than that of the Fujisaki's model (Chomphan, 2011d).

From the study of speech compression for noise-corrupted Thai expressive speech by using two coding methods of CS-ACELP and MP-CELP, it is said that CS-ACELP gives the better speech quality than that of MP-CELP at all three bitrates of 6000, 8600 and 12600 bps. When considering the levels of noise, the 20-dB

noise gives the best speech quality, while 0-dB noise gives the worst speech quality. When considering the speech gender, female speech gives the better results than that of male speech. Finally, when considering the types of noise, the air-conditioner noise gives the best speech quality, while the train noise gives the worst speech quality (Chomphan, 2011b).

## CONCLUSION

This study reviews the study on Thai expressive speech. From the study of $F_0$ modeling; Fujisaki's model and structural model, it can be concluded that two mathematical models have been successfully applied to model the fundamental frequency contour of Thai expressive speech. From the study of speech compression for noise-corrupted Thai expressive speech by using two coding methods of CS-ACELP and MP-CELP, it can be concluded that coding methods, types of noise, levels of noise, speech gender influence on the coding speech quality.

## ACKNOWLEDGEMENT

The researchers are grateful to Kasetsart University at Si Racha campus for the research scholarship through the board of research.

## REFERENCES

Chomphan, S. and T. Kobayashi, 2007. Implementation and evaluation of an HMM-based Thai speech synthesis system. Proceedings of the 8th Annual Conference of the International Speech Communication Association, Aug. 27-31, ISCA Archive, Antwerp, Belgium, pp: 2849-2852.

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. DOI: 10.3844/jcssp.2009.905.914

Chomphan, S., 2010a. Analytical study on fundamental frequency contours of Thai expressive speech using Fujisaki's model. J. Comput. Sci., 6: 36-42, 10.3844/jcssp.2010.36.42

Chomphan, S., 2010b. Multi-pulse based code excited linear predictive speech coder with fine granularity scalability for tonal language. J. Comput. Sci., 6: 1288-1292. DOI: 10.3844/jcssp.2010.1288.1292

Chomphan, S., 2010c. Fujisaki's model of fundamental frequency contours for thai dialects. J. Comput. Sci., 6: 1263-1271. DOI: 10.3844/jcssp.2010.1263.1271

Chomphan, S., 2010d. Performance evaluation of multi-pulse based code excited linear predictive speech coder with bitrate scalable tool over additive white Gaussian noise and Rayleigh fading channels. J. Comput. Sci., 6: 1438-1442. DOI: 10.3844/jcssp.2010.1438.1442

Chomphan, S., 2010e. Structural modeling of fundamental frequency contour for thai expressive speech. J. Comput. Sci., 6: 330-335. DOI: 10.3844/jcssp.2010.330.335

Chomphan, S., 2010f. Tone question of tree based context clustering for hidden markov model based thai speech synthesis. J. Comput. Sci., 6: 1474-1478. DOI: 10.3844/jcssp.2010.1474.1478

Chomphan, S., 2011a. Analysis of fundamental frequency contour of coded speech based on multi-pulse based code excited linear prediction algorithm. J. Comput. Sci., 7: 865-870. DOI: 10.3844/jcssp.2011.865.870

Chomphan, S., 2011b. Speech compression for noise-corrupted thai expressive speech. J. Comput. Sci., 7: 1565-1573. DOI: 10.3844/jcssp.2011.1565.1573

Chomphan, S., 2011c. Tonal language speech compression based on a bitrate scalable multi-pulse based code excited linear prediction coder. J. Comput. Sci., 7: 154-158. DOI: 10.3844/jcssp.2011.154.158

Chomphan, S., 2011d. Modeling of fundamental frequency contour of thai expressive speech using Fujisaki's model and structural model. J. Comput. Sci., 7: 1310-1317. DOI: 10.3844/jcssp.2011.1310.1317

Hiroya, F. and S. Hiroshi, 1971. A model for the generation of fundamental frequency contours of Japanese word accent. J. Acoust. Soc. Japan, 27: 445-452.

Hiroya, F. and O. Sumio, 2002. A preliminary study on the modeling of fundamental frequency contours of Thai utterances. Proceedings of the 6th International Conference on Signal Processing, Aug. 26-30, IEEE Xplore Press, Beijing, China, pp: 516-519. DOI: 10.1109/ICOSP.2002.1181106

Laflamme, C., J.P. Adoul, R. Salami, S. Morissette and P. Mabilleau, 1991. 16 kbps wideband speech coding technique based on algebraic CELP. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Apr. 14-17, IEEE Xplore Press, Toronto, Ont., Canada, pp: 13-16. DOI: 10.1109/ICASSP.1991.150267

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

MacNeilage, P.F., 1983. The Production of Speech. 1st Edn., Springer-Verlag, New York, ISBN: 0387907351, pp: 302.

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

Ozawa, K., T. Nomura and M. Serizawa, 1997. MP-CELP speech coding based on multipulse vector quantization and fast search. Elect. Commun. Jap. Part III: Fundamental Elect. Sci., 80: 55-63. DOI: 10.1002/(SICI)1520-6440(199711)80:11<55::AID-ECJC6>3.0.CO;2-R

Saito, T. and M. Sakamoto, 2002. Applying a hybrid intonation model to a seamless speech synthesizer. Proceedings of the 7th International Conference on Spoken Language Processing, Sep. 16-20, ISCA Archive, Denver, Colorado, USA., pp: 165-168.

Seresangtakul, P. and T. Takara, 2002. Analysis of pitch contour of Thai tone using Fujisaki's model. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 13-17, IEEE Xplore Press, Orlando, USA., pp: 505-508. DOI: 10.1109/ICASSP.2002.5743765

Seresangtakul, P. and T. Takara, 2003. A generative model of fundamental frequency contours for polysyllabic words of Thai tones. Proceedings of the IEEE 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

Tao, J., J. Yu and W. Zhang, 2006. Internal dependence based $F_0$ model for mandarin tts system. Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation, Jun. 19-21, Barcelona, Spain, pp: 171-174.

Taumi, S., K. Ozawa, T. Nomura and M. Serizawa, 1996. Low-delay CELP with multi-pulse VQ and fast search for GSM EFR. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 7-10, IEEE Xplore Press, Atlanta, USA., pp: 562-565. DOI: 10.1109/ICASSP.1996.541158

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. Proceedings of the 2nd International Symposium on Tonal Aspects of Languages (ISTAL'06), La Rochelle, France.