American Journal of Applied Sciences 9 (9): 1415-1421, 2012
ISSN 1546-9239
© 2012 Science Publication

Effects of Noises on Fundamental Frequency Extraction Using Cepstral Analysis for Thai Dialects

[1, 2]Suphattharachai Chomphan
[1]Department of Electrical Engineering, Faculty of Engineering at Si Racha,
Kasetsart University, 199 M.6, Tungsukhla, Si Racha, Chonburi, 20230, Thailand
[2]Center for Advanced Studies in Industrial Technology, Kasetsart University,
50 Ngam Wong Wan Rd, Ladyaow, Chatuchak, Bangkok, 10900, Thailand

Abstract: Problem statement: The fundamental frequency (F0) of the human speech corresponds to the vibration frequency of the human vocal chords. To extract the F0 from a speech utterance, one approach is based on the Cepstral analysis. In Thai, there are four main dialects spoken by Thai people residing in four core region including central, north, northeast and south regions. Environmental noises are also playing an important role in corrupting the speech quality. It is needed to study of effects of noises on F0 extraction using the Cepstral analysis for Thai dialects. **Approach:** The Cepstral analysis is performed and some coefficients are used to determine the corresponding F0 values. Four types of environmental noises are simulated with different levels of power. The differences among the extracted F0 from clean speech and the extracted F0 from noise-corrupted speech are calculated in Root Mean Square (RMS) errors. **Results:** The selected noises are train, factory, car and air conditioner. Five levels of each type of noise vary from 0-20 dB. From the experimental results, it has been noticed that the effects of noises are different. The lowest effect is of air conditioner, meanwhile the noise level of 0 dB is of the highest effect. **Conclusion:** By using the Cepstral analysis, F0 values can be extracted from the noise-corrupted speech with different level of effects depending on the type and level of noises.

Keywords: Thai dialects, cepstral analysis, cepstrum, fundamental frequency, analysis of fundamental frequency, environmental noise, experimental results

INTRODUCTION

The cepstrum is defined as the inverse Fourier transform of the logarithm of the Fourier transform module. The cepstral analysis can bring about the F0 extraction from some corresponding coefficients. Cepstral coefficients are therefore very convenient to find the F0 without calculating a large iteration from the autocorrelation analysis approach (Ravichandran and Samy, 2006).

F0 estimation, also referred to as pitch detection, has been a popular research topic for several years and is still being investigated presently (Seresangtakul and Takara, 2003). The basic problem is to extract the F0 from a human speech signal, which is usually the lowest frequency component. In a periodic waveform, most partials are harmonically related, meaning that the frequency of most of the partials are related to the frequency of the lowest partial by a small whole number ratio. The frequency of this lowest partial is F0 of the waveform.

In noisy-environment speech communication, the noise causes the degradation of naturalness of the speech utterance. To develop the natural speech communication system, it is needed to know how the noise deteriorates F0 values. In this study, the ceptral analysis technique is applied for extracting the F0 values.

The speech corpus used in this study includes all four Thai dialects; standard Thai, Lanna or North dialect, Lao-style or North East dialect and South dialect. As for the environmental noises, four different types of noises are recorded and scaled to fit the determined power level.

MATERIALS AND METHODS

2.1. Cepstral Analysis

In the human speech production, a speech signal uttered is due to the input excitation emerged from the vocal chords and the trachea and also the response of the vocal tract system. From the signal processing point of view, the output of a system can be treated as the

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

convolution of the input excitation with the system response. To extract the voicing F0 of the output speech signal, it is needed to obtain the input excitation separately. The main objective of cepstral analysis of human speech is to separate the speech into its excitation source and system components (Ravichandran and Samy, 2006).

According to the theory of speech production, the voiced speech is produced by exciting the corresponding vocal tract system with periodic impulse sequence, meanwhile the unvoiced speech is produced by exciting the corresponding vocal tract system with a random noise sequence. The resulting speech can be determined as the convolution of input excitation and system response. It is denoted that e[n] is the input excitation and h[n] is the vocal tract system response, then the speech signal s[n] can be expressed in Eq. 1:

$$S\lfloor n \rfloor = e\lfloor n \rfloor * h\lfloor n \rfloor \qquad (1)$$

By using Discrete Fourier Transform (DFT), it can be represented in frequency domain in Eq. 2:

$$S(\omega) = E(\omega).H(\omega) \qquad (2)$$

The Eq. 2 indicates that the multiplication of excitation and system components in the frequency domain for the convolved sequence in the time domain. The speech signal has to be deconvolved into the excitation and vocal tract components in the time domain. The multiplication of the two components in the frequency domain has to be converted to a linear combination of the two components.

From the Eq. 2 the magnitude spectrum of given speech signal can be represented as:

$$|S(\omega)| = |E(\omega)|.|H(\omega)| \qquad (3)$$

The logarithmic representation is applied to linearly combine the E(ω) and H(ω) in the frequency domain. Therefore, the logarithmic representation of Eq. 3 can be represented as:

$$\log|S(\omega)| = \log|E(\omega)| + \log|H(\omega)| \qquad (4)$$

As indicated in Eq. 4, the logarithmic operation transforms the multiplication into a summation of these components. Each separation can be performed by taking the Inverse Discrete Fourier Transform (IDFT) of the linearly combined log spectra of excitation and vocal tract system components. It has been noted that the IDFT of log spectra transforms to quefrency domain or the cepstral domain which is closely similar to time domain as explained in Eq. 5:

$$\begin{aligned} c[n] &= \mathrm{IDFT}\left(\log|S(\omega)|\right) \\ &= \mathrm{IDFT}\left(\log|S(\omega)|\right) + \mathrm{IDFT}\left(\log|S(\omega)|\right) \end{aligned} \qquad (5)$$

The overall cepstral analysis procedure can be summarized in Fig. 1.

F0 extraction: In the quefrency domain, the vocal tract components are represented by the slowly varying components concentrated in the lower quefrency region (about lower than 0.002 s) and excitation components are represented by the fast varying components at the higher quefrency region (about higher than 0.002 s). To obtain the F0 value, it is conducted by take an inverse of the quefrency value at the first peak of the Cepstum at the higher quefrency region. From Fig. 2, for example, the Cepstrum has the first peak at the higher quefrency region at about 0.008 s. Therefore, the corresponding F0 value is 1/0.008 or 125 Hz (Chomphan, 2010a; 2010b).

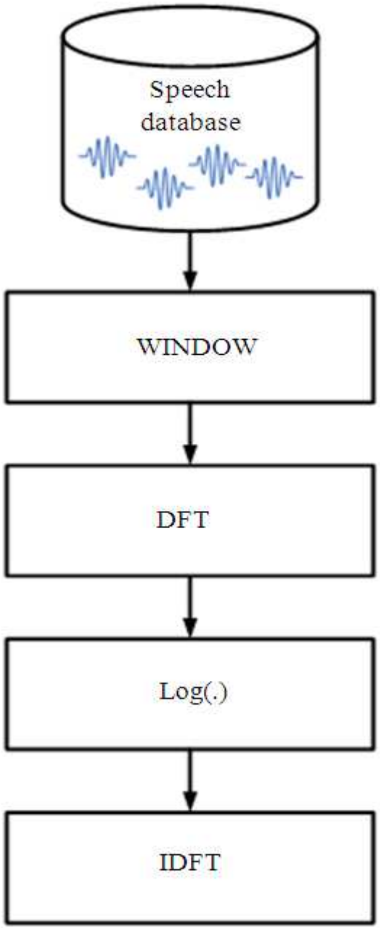

Fig. 1: Cepstral analysis procedure

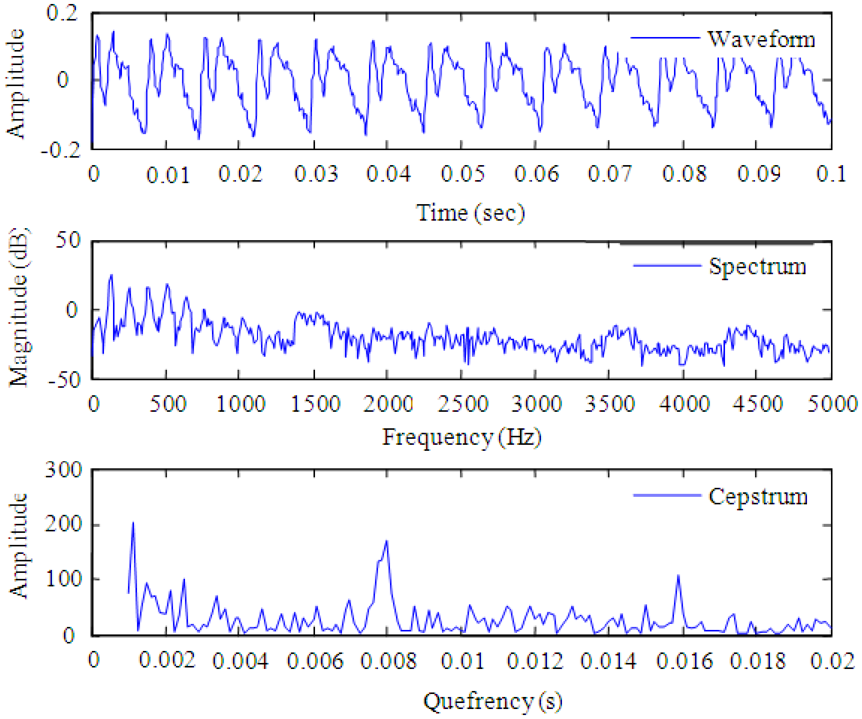

Fig. 2: Comparison of Spectrum and Cepstrum of an intercept of a speech utterance

Environmental noises: Four types of noises include train, factory, car and air conditioner. They are mixed directly with the recorded clean speech in the speech database. Before mixing noises with the clean speech, the noise volume or power are adjusted in several exact levels. As for the level variation of noises, the levels of each type of noise are varied from 0, 5, 10, 15, 20 dB, respectively (Chomphan and Kobayashi, 2008; 2009)

RESULTS

As for the speech material, ten sentences are selected as the common speech in Thai for male and female genders. The sentences have been recorded in four Thai dialects of standard Thai (Center-dialect), Lanna Thai dialect (North-dialect), Lao-style Thai dialect (Northeast-dialect) and South Thai dialect (South-dialect). The cepstral analysis has been performed and then the F0 contour for each speech utterance in the speech database has been extracted from the cepstral coefficients. (Mixdorff and Fujisaki, 1997; Fujisaki and Sudo, 1971; Chomphan and Kobayashi, 2007a; 2007b).

The resulting extracted F0s from cepstral analysis for all twenty samples of noise-corrupted speech are used to calculated RMS errors by comparing with that of the clean speech. The experimental results are presented in Fig. 3-6 for four dialects of female speech and Fig. 7-10 for four dialects of male speech.

DISCUSSION

From the experimental results of female speech as shown in Fig. 3-6, it can be noticed that the RMS error of air-conditioner noise is lowest, meanwhile RMS errors of the other types of noises are quite similar. When considering the level of noises in decibel, it can be empirically seen that the RMS errors at 0 dB are in the highest level and they move downward to the lowest level at 20 dB. From the experimental results of male speech as shown in Fig. 7-10, it can be noticed that the RMS error of air-conditioner noise is lowest, meanwhile RMS errors of the other types of noises are not much different. As for considering the level of noises in decibel, it can be obviously noticed that the RMS errors at 0 dB are in the highest level and they move downward to the lowest level at 20 dB.

When considering the gender of speech, it can be seen that the RMS errors of female speech are mostly higher than that of male speech. Due to the fast movement of F0 contour of female speech, its corresponding RMS error is larger.

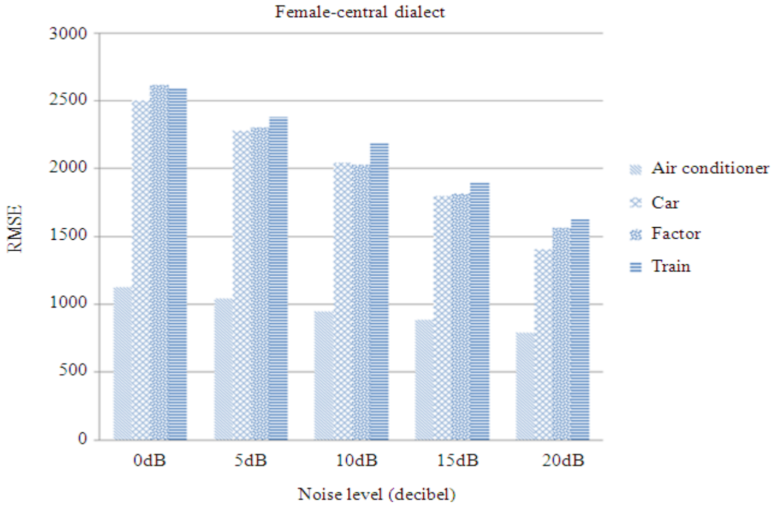

Fig. 3: Comparison of RMSE of female central dialect with four types of noises and five different levels of noise

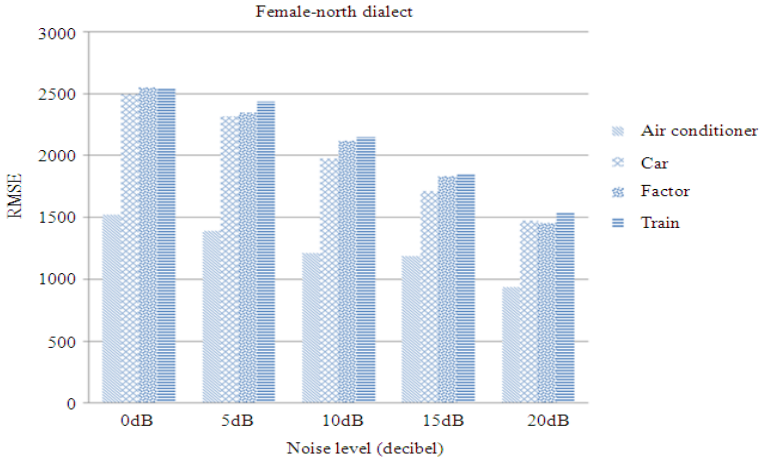

Fig. 4: Comparison of RMSE of female north dialect with four types of noises and five different levels of noise

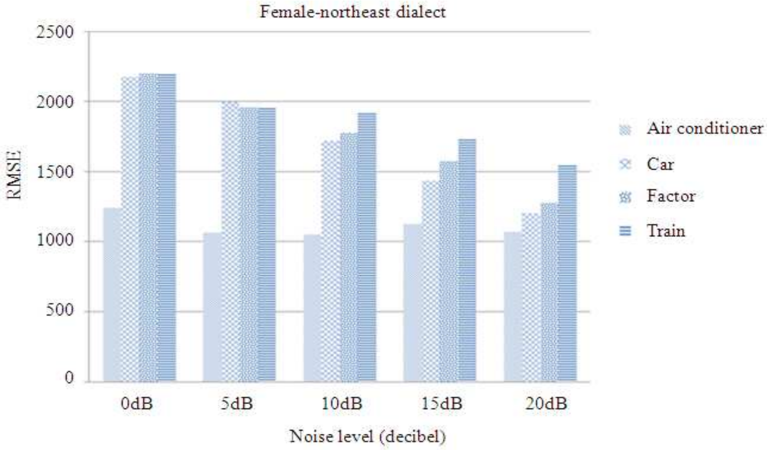

Fig. 5: Comparison of RMSE of female northeast dialect with four types of noises and five different levels of noise

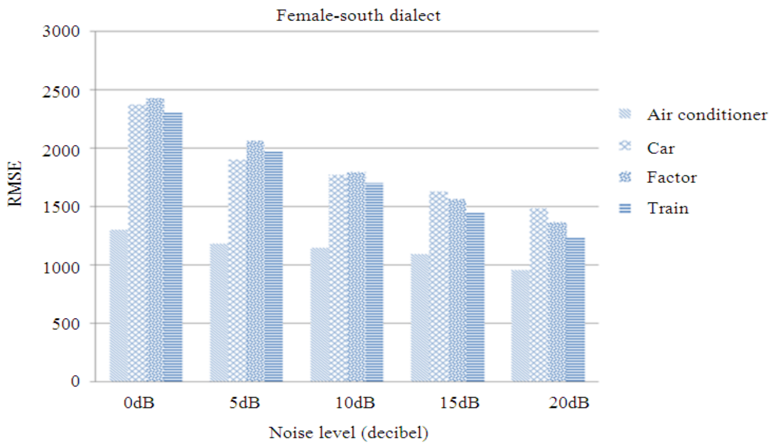

Fig. 6: Comparison of RMSE of female south dialect with four types of noises and five different levels of noise

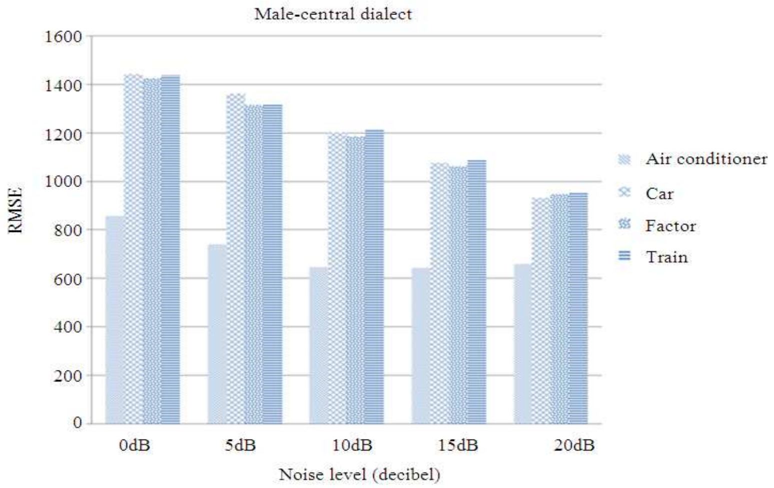

Fig. 7: Comparison of RMSE of male central dialect with four types of noises and five different levels of noise

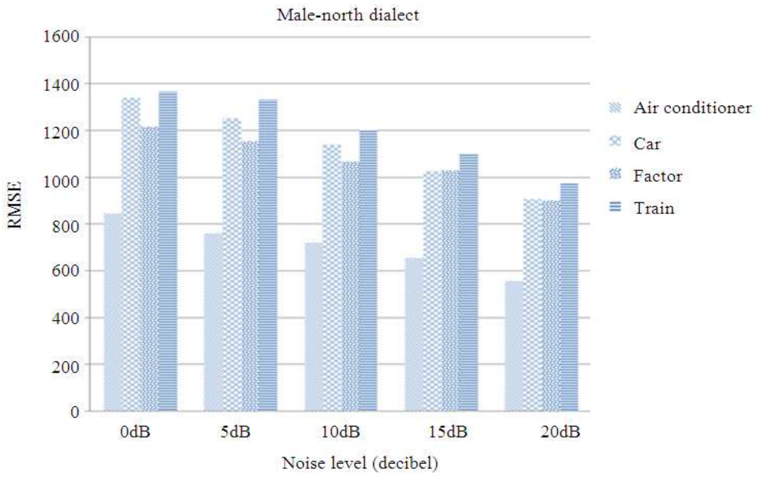

Fig. 8: Comparison of RMSE of male north dialect with four types of noises and five different levels of noise

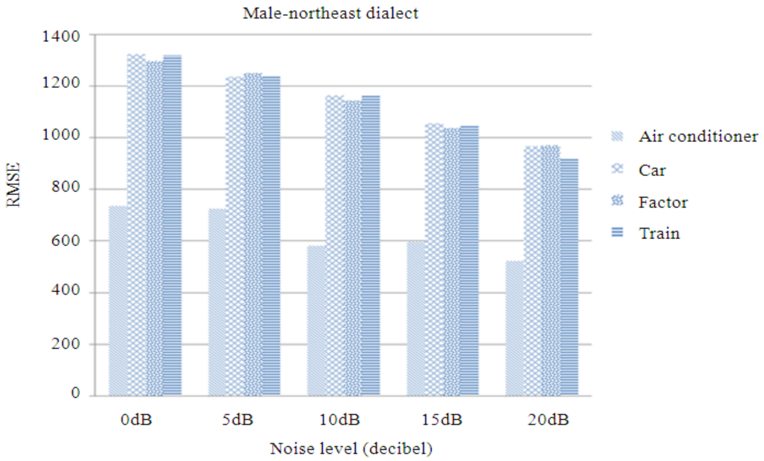

Fig. 9: Comparison of RMSE of male northeast dialect with four types of noises and five different levels of noise

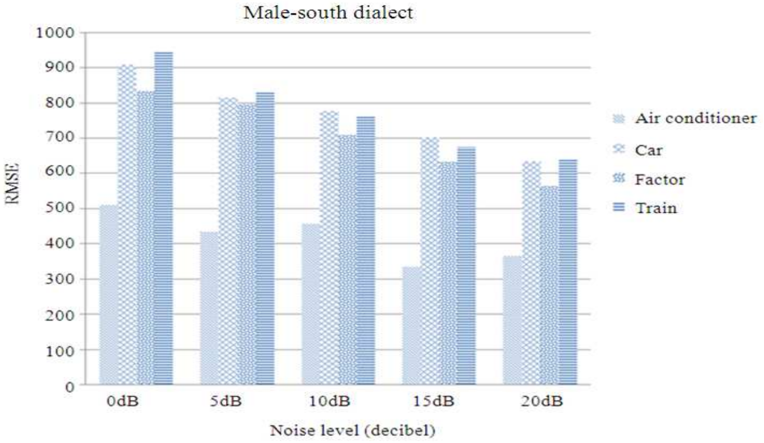

Fig. 10: Comparison of RMSE of male south dialect with four types of noises and five different levels of noise

CONCLUSION

This stduy presents a study of effects of noises on extraction of F0 contour by using cepstral analysis for Thai dialects. Four types of environmental noises are simulated with different levels of power. The comparisons of different types of noises and different levels of noise of four Thai dialects have been summarized. The experimental results show that the RMS error of air-conditioner noise is lowest for most of dialects. Moreover, the RMS errors of female speech are mostly higher than that of male speech. All in all, the environmental noises explicitly deteriorate the F0 contours from cepstral analysis.

ACKNOWLEDGEMENT

The author is grateful to Kasetsart University for the research scholarship through the Center for Advanced Studies in Industrial Technology.

REFERENCES

Chomphan, S. and T. Kobayashi, 2007a. Design of tree-based context clustering for an HMM-based Thai speech synthesis system. Proceedings of the 6th ISCA Workshop on Speech Synthesis, Aug. 22-24, Bonn, Germany, pp: 160-165.

Chomphan, S. and T. Kobayashi, 2007b. Implementation and evaluation of an HMM-based Thai speech synthesis system. Proceedings of the 8th Annual Conference of the International Speech Communication Association, (ISCA’ 07), ISCA, 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., 2010a. Analytical study on fundamental frequency contours of Thai expressive speech using Fujisaki's model. J. Comput. Sci., 6: 36-42. DOI: 10.3844/jcssp.2010.36.42

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

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.

Mixdorff, H. and H. Fujisaki, 1997. Automated quantitative analysis of F_0 contours of utterances from a German ToBI-labeled speech database. Proceedings of the Eurospeech, Sept. 22-25, Rhodes, Greece, pp: 187-190.

Seresangtakul, P. and T. Takara, 2003. A generative model of fundamental frequency contours for polysyllabic words of Thai tones. Proceedings of the International Conference on Acoustics, Speech and Signal Processing, April 6-10, IEEE Xplore Press, Japan, pp: 452-455. DOI: 10.1109/ICASSP.2003.1198815

Ravichandran, T. and K.D. Samy, 2006. Performance enhancement on voice using VAD algorithm and cepstral analysis. J. Comput. Sci., 2: 835-840. DOI: 10.3844/jcssp.2006.835.840