10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Physiological Pitch Range Estimation from a Brief Speech Input: A Study on a Bilingual Parallel Speech Corpus

ZHANG Wei[1,2], *XIE Yanlu*[1], *ZHANG Jinsong*[1]

[1]Institute of International Chinese Language Education, Beijing Language and Culture University, Beijing, China
[2]Department of Linguistics, McGill University, Montreal, Canada
wei.zhang16@mail.mcgill.ca, {xieyanlu,jinsong.zhang}@blcu.edu.cn

Abstract

The range of pitch that a speaker can maximally produce is constrained by the physiological characteristics. Unlike the speaking pitch range in speech samples, this kind of 'physiological pitch range' is independent from the content of the speech, but can be partly estimated by human listeners from even a brief speech, employing not only the fundamental frequency (F0) but also the spectral features. In our previous work, we proposed a spectrum-based algorithm for estimating physiological pitch range from a brief speech, which outperformed the traditional F0 analysis method when the speech input was as short as 300ms. The present study continued to test the algorithm on a Japanese-Chinese parallel speech corpus uttered by a group of native speakers of Japanese who spoke Mandarin as a second language. For each speaker, the proposed algorithm obtained almost the same pitch range from his/her L1 and L2 speech data, whereas the traditional method gave two estimations with a larger difference. The results verified that the proposed algorithm was more capable of estimating a speaker's physiological pitch range from a brief speech.

Index Terms: physiological pitch range, brief speech input, parallel speech corpus

1. Introduction

In speech communication, pitch conveys important information, and it is normalized automatically by the listener through a subjective estimation of the speaker's pitch range, which refers to pitch variations between the upper and lower extremes that the speaker produces [1]. Pitch range is potentially determined by speakers' organic factors and vocal characteristics [2, 3]. Gender, age, weight and pathology are among the factors that are most reported to influence the vocal characteristics of an individual [4-12], and physiological changes on these factors will cause the pitch range variation [13]. Therefore, for a certain speaker, his/her physiological structure determines the intrinsic pitch range variation, namely *physiological pitch range* [14, 15]. Physiological pitch range reflects the maximum range of pitch that the speaker can produce, and it is a speaker-specific quantity and does not change with the speech content [14, 15]. Many researches showed that a speaker's "pitch range" usually varied with mood or language [16-21], in the previous literature, for such cases it is called *speaking pitch range* [14]. Compared to physiological pitch range, speaking pitch range usually changes with communication need and speaking style, and is influenced by language proficiency [22-24].

Since pitch range differs across individuals, the absolute F0 value of a particular functional pitch realization such as tone in a tonal language varies with speakers. However, the phonological form can be decoded by human listeners in the context automatically and effectively. The interpretation of the pitch location, or the mechanism of pitch height normalization, has to depend on human listener's ability to estimate the speaker's overall pitch range [26, 28, 30].

Some studies further claimed that human listeners were able to identify the relative pitch height from a brief speech input of an unknown speaker, even when it is too short to contain sufficient F0 variation including high and low pitch targets [26-29]. For example, listeners can identify the Mandarin four tones (high-flatting, rising, low-dripping, falling) even when F0 is available only in the initial consonant and the first six glottal periods in the vowel [28]. It actually showed that listener's interpretation of the speaker's pitch range is very efficient. Besides the limited F0 information, when estimating pitch range listeners also use other potential cues in the spectral structure such as spectral tilt and F1 bandwidth [27, 29, 31-35], which contains the vocal tract characteristic of the speaker. Hence, the pitch range estimated by listeners here is more close to the concept of physiological pitch range.

The ability of pitch range estimation plays an important role not only in human speech communication but also in intelligent machine systems [25]. For the algorithms of automatic pitch range estimation, there are two main types of measures. One is based on "long-term distributional" (LTD) method, and the other is based on specific landmarks in speech that are linguistic in nature ("linguistic" measures) [3]. They are both aiming at calculating the maximum and minimum F0 values from a speaker's lengthy speech input. Thus, from previous literature we can conclude that the two measures are all calculating the "speaking pitch range" actually. However, in the case of short speech input, for example, when only one or two syllables are available, these methods will be vulnerable for reason of the insufficient F0 data.

Our previous work aimed at pitch range estimation using a brief speech input, by mapping the spectral structure and the pitch range directly using machine learning techniques [36]. More exactly, it estimated an individual's vocal characteristic, namely physiological pitch range, from the spectral information. It has been widely accepted that a speaker usually has a different speaking pitch range in L2 speech comparing to their native speech (narrower than L1) [22; 37-40]. While the physiological pitch range is a stable individual quality and do not easily change even in different speaking circumstance. In this paper, we gave a further verification that the proposed

10.21437/SpeechProsody.2020-196

method was more capable of estimating a speaker's physiological pitch range from a brief speech by using a Japanese and L2 Chinese parallel corpus, of which the speakers are all native Japanese.

2. Spectrum-based pitch range estimation model

Human listener's pitch range estimation is a real-time adaptive process incorporating a feedback mechanism. The estimation gets more accurate when more speech input feed in, and it converges when the speech input reaches a certain length. Recurrent neural network (RNN) has the similar property of online adaptation [41]. In our previous work, we adopted the RNN with long short-term memory (LSTM) cells, which gained a lot of success in speech engineering tasks [42-43], as the model for estimation.

In this experiment, we followed our previous proposed method, but we improved the modeling schemes and used larger and more diverse corpora.

2.1. Framework

The model framework is shown as Figure 1. The spectral features and calculated pitch range labels of training set are input to the machine learning algorithm, namely LSTM in this paper, and it makes mapping between features and labels. Then we get the trained regression model. For test stage, the spectral features of test set are fed into the trained model, then it will return the estimated pitch range values.

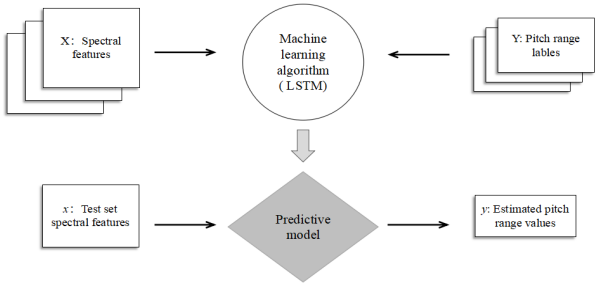

Figure 1: *Framework of spectral based physiological pitch range estimation model.*

2.2. Corpus

From the definition, physiological pitch range is the maximum range of the speaker's pitch. In the literature [14-15], physiological pitch range is obtained by elicited tone sweep, which the speakers are under a guidance of "Move gradually higher until you feel your voice break and go to a falsetto" and "Move gradually but quickly lower, letting your voice creak, until you can't go any lower". However, so far we don't have such a real physiological pitch range corpus which is large enough to train an LSTM model. Restricted to the resources, what we use now is still the collected speech, but we tried to make sure that each speaker's data is of as least 25 minutes to raise the probability of reaching the physiological pitch range.

In this model, we used the following three corpora:

- Chinese National Hi-Tech Project 863 Corpus, which included data of 110 hours from 166 speakers [44].
- Open-source Mandarin Speech AISHELL Corpus, which included data of 170 hours from 400 speakers [45].
- Open-source Mandarin Speech THCHS-30 Corpus, which included data of 30 hours from 50 speakers [46].

To balance the computation resource and robustness of the model, we only used a quarter of gender balanced data chosen randomly from each corpus. We took 10% of the data as the test set, the other as the training set. There was no overlap between the training and test sets at either the speaker level or utterance level.

2.3. Experimental setup

2.3.1. Features

We used 43-dimensional FBANKs and PITCH extracted from each frame (25ms window and 10ms shift) as the input features. We removed silence samples by voice activity detection. The features were extracted using KALDI toolkit [47].

2.3.2. LSTM models

In this experiment, we used 3 parameters, i.e., ceiling, floor and mean, to describe the pitch range and pitch level of a speaker. Inspired by [48], we trained the LSTM regression models in a multi-task method. For each speaker, the labels of ceiling, floor and mean were obtained from all his/her F0 values in the corpora. All labels were transformed into the logarithmic scale, which is more consistent with sensation experience of human listeners.

The LSTM regression models were trained using the KERAS toolkit [49]. Following the most economic seeting in previous work, the net depth was 30 time steps and window length was 1 frame (which means that the input are 300ms' speech segments), and the mean absolute percentage error (MAPE) criterion was chosen,

$$\mathrm{MAPE} = Num^{-1}\sum_{i=1}^{Num}\left|\frac{y-\tilde{y}}{y}\right|\times 100 \quad (1)$$

The net structure of the LSTM model is shown in Figure 2. It contained 2 LSTM layers, with 32 memory cells in each. For the purpose of regression, a dense layer containing one neuron serves as the output layer. The three pitch range parameters were trained together under the multi-task method.

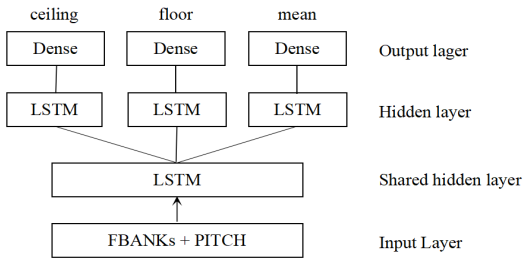

Figure 2: *Network structures for pitch range estimation model using multi-task method.*

2.3.3. Performance of the model

Table 1 describes the MAPE performance of the model on the test set. The *mean* and *floor* parameters are of the best and worst prediction performance.

Table 1: *MAPE Performance of the spectral based physiological pitch range estimation model in log scale and linear scale.*

	Ceiling	Mean	Floor
MAPE	2.19%	2.15%	2.52%

Our previous experiments showed that, using a 300ms-speech input, the proposed method performed better than the traditional LTD method. In the present work, we modified the pitch range modeling method. Previously, we used the mean and standard deviation as two parameters of pitch range, and the intended pitch range were represented by mean ± 2*standard deviation, indicating the upper and lower limits are equidistant to the mean. However, some studies showed that the perceptual cues for ceiling and floor pitches are different, and the ceiling is inherently more dynamic than the floor [50, 32]. Therefore, we modified it and used separate parameters, namely ceiling, mean and floor, to model the pitch range, which is supposed to be more accurate than before (the previous MAPE of mean is 2.3%).

3. Verification Experiment

The spectrum-based pitch range estimation model is trained in a way that simulates the human listener's perceptual mechanism when the speech input is brief. For the reason that in this process what plays an important role is the spectral information, which is supposed to contain the speaker's individual characteristic of his/her vocal tract. Thus, we suppose that the estimated pitch range from our method has the property of the physiological pitch range, which is a stable individual quality and do not change with the speech content. While the traditional LTD method of pitch range calculation is estimating the speaking pitch range, which would be affected by the speaking material. In this section, we conducted experiments to verify the physiological pitch range property using a Japanese-Chinese Parallel Corpus. We made the following hypothesis:

- For the same Japanese speaker, using his/her Japanese (L1) or Chinese (L2) speech, the spectrum-based pitch range estimation method gave almost the same pitch range results, while the traditional LTD method gave different ones.

3.1. Speech corpus

Eighteen native Japanese speakers (nine male, nine female) are recorded and all of them are L2 learners of Chinese in the last author's University [51]. Conversational Chinese 301 text was used for Chinese speech recording and it was translated into and slightly adapted for Japanese [37].

Speakers read both Chinese and Japanese materials and they were all told to read in a natural way. Then speech files were digitized at the sampling rate of 44.1 KHz with the quantization precision being 16 bits and were saved as wav format.

3.2. Two methods of pitch range estimation

3.2.1. Measure from spectrum based pitch range estimation model

For each speaker in the corpus, we chose 20 parallel utterances respectively from Chinese and Japanese material. Using the spectrum based pitch range estimation model, two sets of pitch range parameters were obtained, one is from the 20 Chinese utterances, the other is from the 20 Japanese utterances.

To fit the model, each utterance was divided into several 300ms-segments and fed into the well trained model. The final result was calculated by averaging the results of the segments from the 20 utterances.

3.2.2. Measure from LTD method

For comparison, the traditional LTD method of pitch range estimation was also tested. We calculated the mean and standard deviation of the F0 values, and then obtained the ceiling and floor parameters:

$$\text{ceiling} = mean + 2 \times sd \tag{2}$$

$$\text{floor} = mean - 2 \times sd \tag{3}$$

Here, *sd* represents the standard deviation.

F0 values were extracted with the PRAAT algorithm based on an autocorrelation method. All settings remained at the default. The tracking uses a time step of 10 ms.

For each speaker, two sets of pitch range parameters were calculated from the Chinese and Japanese speech corpora separately.

3.3. Results

Figure 3 and Figure 4 shows the pitch range results of the LTD method and the results of the spectrum based estimation model. They are showing the averages of pitch range values on 3 dimensions (i.e. ceiling, floor and mean) in different gender and spoken language for the 18 Japanese participants.

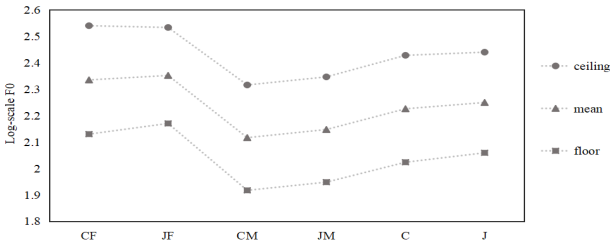

Figure 3: *Pitch range calculation Results of LTD method. (On horizontal axis, CF, JF, CM, JM represents Chinese spoken by Female participants, Japanese spoken by Female participants, Chinese spoken by male participants and Japanese spoken by male participants, respectively. C means Chinese spoken by all participants and similarly J means Japanese spoken by all participants).*

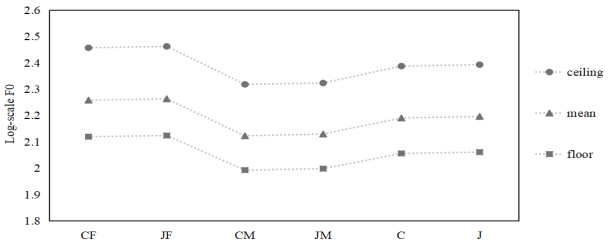

Figure 4: *F0 range calculation Results of spectrum-based pitch range estimation model.*

From Figure 3, in the LTD method, the F0 range parameters of Japanese participants are different when they speak Japanese and Chinese, both in male and female groups. Paired-Samples T tests show significant differences for floor ($t = 4.666$, $p < 0.001$), and mean ($t = 3.638$, $p = 0.002$). But no significant difference for ceiling ($t = 1.035$, $p = 0.315$). In Figure 4, from spectrum-based physiological pitch range estimation model, the F0 range parameters of Japanese participants are almost same when they are speaking Japanese and Chinese, both in male and female groups. Paired-Samples T tests show no significant differences for the ceiling, floor or mean.

For each method, we also analyzed the difference in pitch range parameters estimated from the Chinese and Japanese speech corpora. For example, for each participant, the ceiling difference is defined as the absolute difference between the ceiling values estimated from his/her Chinese and Japanese speeches, as shown below:

$$diff_{LTD} = \left|Para_{CH} - Para_{JP}\right| \quad (4)$$

$$diff_{SBM} = \left|Para_{CH} - Para_{JP}\right| \quad (5)$$

Here '$diff_{LTD}$' and '$diff_{SBM}$' represent the value differences of pitch range parameters from LTD method and Spectrum based pitch range estimation model (abbreviated as SBM), respectively. '$Para_{CH}$' and '$Para_{JP}$' represent the pitch range parameters estimated from Chinese and Japanese speech, respectively.

The results of Paried-Samples T Test between diff_LTD and diff_SBM are listed in Table2. For all the three pitch range parameters, the predicted differences of SBM is significantly lower than that of the LTD method ($p < 0.05$), which suggests that SBM is estimating the speaker-dependent physiological characteristic, and it is hardly affected by the content of the speech. These results proved the validity of the above hypothesis.

Table 2: *Results of Paired-Samples T Test on the pitch range difference between the two method from Japanese and Chinese speech. The 95% Confidence Interval of the Difference was used.*

		Mean	t	df	*p*-value
Ceiling	$Diff_{SBM}$ - $Diff_{LTD}$	-.019	-2.233	17	.039*
Mean	$Diff_{SBM}$ - $Diff_{LTD}$	-.025	-3.687	17	.002*
Floor	$Diff_{SBM}$ - $Diff_{LTD}$	-.014	-2.470	17	.024*

4. Discussion and Conclusion

Speech signals encode a lot of information. Spectral features like FBanks, describe the complex frequency distributions in the speech signal, are regarded to contain a large amount of information. It is unsurprising that the physiological pitch range can be predicted from the spectral features. The result that for all the pitch range parameters, the predicted differences from L1 and L2 speech of a same speaker of the proposed method is significantly lower than that of the LTD method, verified that the proposed spectrum-based method is more capable of estimating a speaker's physiological pitch range, which is irrelevant to the content he/she is producing, from a 300ms-speech.

As mentioned before, the labels of each speaker's pitch range in the present model is not precise enough to be the real upper and lower limit that the speaker is capable of producing. From the results, under this method, they can also show the stable property of the physiological pitch range.

In conclusion, as a follow-up of our proposed spectrum-based method of pitch range estimation from a brief speech input, this study modified the model by updating the modeling scheme and enriching the training data. More importantly, we tested the method on a Japanese-Chinese parallel speech corpus uttered by native Japanese speakers, and verified that the proposed method is capable of estimating a speaker's physiological pitch range.

5. Acknowledgement

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Science Foundation and Special Program for Key Basic Research fund of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities)(20YJ040002, 16ZDJ03), and the project of "Intelligent Speech technology International Exchange". Jinsong Zhang is the corresponding author.

6. References

[1] R.L. Trask, *A Dictionary of Phonetics and Phonology.* London: Routledge, 1996.

[2] J. Laver, *Principles of Phonetics.* Cambridge: Cambridge University Press, 1994.

[3] Mennen I, Schaeffler F, Docherty G. "Cross-language differences in fundamental frequency range: A comparison of English and German," *The Journal of the Acoustical Society of America*, 2012, 131(3): 2249-2260.

[4] Brockmann M, et al. "Voice loudness and gender effects on jitter and shimmer in healthy adults," *Journal of Speech, Language, and Hearing Research*, 2008, 51(5): 1152-1160.

[5] Van Dommelen W A, Moxness B H. “Acoustic parameters in speaker height and weight identification: sex-specific behavior,” *Language and speech*, 1995, 38(3): 267-287.

[6] Deliyski S A X D. “Effects of aging on selected acoustic voice parameters: Preliminary normative data and educational implications,” *Educational Gerontology*, 2001, 27(2): 159-168.

[7] Stathopoulos E T, Huber J E, Sussman J E. “Changes in acoustic characteristics of the voice across the life span: measures from individuals 4–93 years of age,” *Journal of Speech, Language, and Hearing Research*, 2011, 54(4): 1011-1021.

[8] Awan S N. “The aging female voice: acoustic and respiratory data,” *Clinical linguistics & phonetics*, 2006, 20(2-3): 171-180.

[9] Shipp T, Huntington D A. “Some acoustic and perceptual factors in acute-laryngitic hoarseness,” *Journal of Speech & Hearing Disorders*, 1965, 30(4):350.

[10] Hecker M H L, Kreul E J. “Descriptions of the speech of patients with cancer of the vocal folds. Part I: Measures of fundamental frequency,” *The Journal of the Acoustical Society of America*, 1971, 49(4B): 1275-1282.

[11] Cooper M. “Spectrographic analysis of fundamental frequency and hoarseness before and after vocal rehabilitation,” *Journal of Speech and Hearing Disorders*, 1974, 39(3): 286-297.

[12] Murry T, Doherty E T. “Selected acoustic characteristics of pathologic and normal speakers,” *Journal of Speech, Language, and Hearing Research*, 1980, 23(2): 361-369.

[13] Traunmüller H, Eriksson A. “The frequency range of the voice fundamental in the speech of male and female adults,” Unpublished manuscript, 1995.

[14] Baken R J, Orlikoff R F. “Clinical measurement of speech and voice,” *Cengage Learning*, 2000.

[15] Keating P, Kuo G. “Comparison of speaking fundamental frequency in English and Mandarin,” *The Journal of the Acoustical Society of America*, 2012, 132(2): 1050-1060.

[16] Altenberg E P, Ferrand C T. “Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women,” *Journal of Voice*, 2006, 20(1): 89-96.

[17] Mennen I, Schaeffler F, Docherty G. “Cross-language differences in fundamental frequency range: A comparison of English and German,” *The Journal of the Acoustical Society of America*, 2012, 131(3): 2249-2260.

[18] M L. Ng, Y. Chen, et. al. “Differences in vocal characteristics between Cantonese and English produced by proficient Cantonese-English bilingual speakers—a long-term average spectral analysis,” *Journal of Voice*, 2012, 26(4): e171-e176.

[19] M L. Ng, G. Hsueh, C S. Sam Leung. “Voice pitch characteristics of Cantonese and English produced by Cantonese-English bilingual children,” *International journal of speech-language pathology*, 2010, 12(3): 230-236.

[20] Y. Chan. “Acoustical differences in vocal characteristics between Cantonese and English produced by Cantonese-English bilingual adult speakers,” 2010, *Journal of Voice*, 15(2): 194-219.

[21] M. Bruyninckx. B. Harmegnies, et al. “Language-induced voice quality variability in,” *Journal of Phonetics*, 1994, 22: 19-31.

[22] M.G. Busà, M. Urbani. “A cross linguistic analysis of pitch range in English L1 and L2,” *Proc. 17th International Congress of Phonetic Sciences (ICPhS XVII)*, Hong Kong. 2011: 380-383.

[23] R. Hincks. “Processing the prosody of oral presentations,” *InSTIL/ICALL Symposium 2004*. 2004.

[24] C. Johns-Lewis. “Prosodic differentiation of discourse modes,” *Intonation in discourse*, 1986: 199-220.

[25] J. Edlund and M. Heldner, *Underpinning/nailon: automatic estimation of pitch range and speaker relative pitch, Speaker Classification II*. Berlin: Springer, 2007, pp. 229-242.

[26] J. Bishop and P. Keating, “Perception of pitch location within a speaker's range: fundamental frequency, voice quality and speaker sex,” *Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1100-1112, 2012.

[27] D. N. Honorof and D. H. Whalen, “Perception of pitch location within a speaker's F0 range,” *Journal of the Acoustical Society of America*, vol. 117, no. 1, pp. 2193-2200, 2005.

[28] C. Y. Lee, “Identifying isolated, multispeaker Mandarin tones from brief acoustic input: a perceptual and acoustic study,” *Journal of the Acoustical Society of America*, 2009, 125(2): 1125-1137.

[29] J. Kuang, M. Liberman, “Influence of spectral cues on the perception of pitch height,” *Proc. of ICPH*, 2015.

[30] E. Shriberg, D.R. Ladd, J. Terken, et al. “Modeling Pitch Range Variation Within and Across Speakers: Predicting F0 Targets When “Speaking Up”,” *Proc. of the 4th international conference on spoken language processing,* 1996.

[31] J. Kuang, Y. Guo, M. Liberman, “Voice quality as a pitch-range indicator,” *Proc. of Speech Prosody*, 2016, pp. 1061-1065.

[32] J. Kuang, M. Liberman, “Pitch-Range Perception: The Dynamic Interaction Between Voice Quality and Fundamental Frequency,” *Proc. of INTERSPEECH*, 2016, pp. 1350-1354.

[33] D. N. Honorof and D. H. Whalen, “Identification of speaker sex from one vowel across a range of fundamental frequencies,” *The Journal of the Acoustical Society of America*, vol. 128, n. 5, pp. 3095-3104, 2010.

[34] F. Y. Li. “A Study of Correlation Between Voice Quality Measures and Tonal F0 Parameters Based on Monosyllabic Chinese Corpora,” *Thesis of Beijing Language and Cultural University*, 2016

[35] J. Lin, Y. Xie, Y. Gao, et al, “Improving Mandarin tone recognition based on DNN by combining acoustic and articulatory features,” *Proc. of International Symposium on Chinese Spoken Language Processing. IEEE*, 2017.

[36] W. Zhang, Y. Xie, J. Zhang. “LSTM-Based Pitch Range Estimation from Spectral Information of Brief Speech Input,” *Proc. of ISCSLP. IEEE*, 2018, pp. 349-353.

[37] S. Shi S, J. Zhang, Y. Xie. “Cross-language comparison of F0 range in speakers of native Chinese, native Japanese and Chinese L2 of Japanese: Preliminary results of a corpus-based analysis,” *Proc. of ISCSLP. IEEE*, 2014:241-244.

[38] Aoyama, K., Guion, S.G. “Prosody in Second Language Acquisition: An Acoustic Analysis on Duration and F0 Range”. *The role of Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, Amsterdam: John Benjamins, 2007, pp. 281-297.

[39] Mennen, I., “Can language learners ever acquire the intonation of a second language?” *Proc. STILL, Marholmen,* Sweden, 1998, 17–20.

[40] Ullakonoja, R. “Comparison of pitch range in Finnish (L1) and Russian (L2),” *Proc. of the 16th International Congress of Phonetic Sciences*, 2007.

[41] C. Giles. “Dynamic Recurrent Neural Networks: Theory and Applications,” *IEEE Transactions on Neural Networks*, vol. 5, n. 2, pp. 153-156, 1994.

[42] S. Hochreiter, J. Schmidhuber. “Long short-term memory,” *Supervised Sequence Labelling with Recurrent Neural Network*, Berlin Heidelberg: Springer, 1997, pp. 1735-1780.

[43] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to Sequence Learning with Neural Networks,”, *Proc. of Advances in neural information processing systems*. 2014, pp. 3104-3112.

[44] S. Gao, B. Xu, et al, “Update progress of Sinohear: advanced Mandarin LVCSR system at NLPR,” *Proc. of 6th International Conference on Spoken Language Processing*, 2000.

[45] H. Bu, J. Du, X. Na, et al, “AIShell-1: An open-source Mandarin speech corpus and a speech recognition baseline,” *arXiv preprint arXiv:1709.05522*, 2017.

[46] D. Wang, X. Zhang. “Thchs-30: A free chinese speech corpus,” *arXiv preprint arXiv:1512.01882*, 2015.

[47] D. Povey, A. Ghoshal, G. Boulianne, et al, “The Kaldi speech recognition toolkit,” *Idiap*, 2011.

[48] Q. Zhang, C. Cao, et al. “Pitch Range Estimation with Multi features and MTL-DNN Model.” *Proc. of 14th IEEE International Conference on Signal Processing (ICSP)*, 2018.

[49] P.W.D. Charles, “KERAS”, *GitHub repository*, https://github.com/charlespwd/keras, 2013.

[50] C. Rolf, K. Elenius, M. Swerts, “Perceptual judgments of pitch range,” *Proc. of Speech Prosody*, Japan, pp. 173-176, 2004.

[51] Y. Kang, S. Lai. *Conversational Chinese 301*. Beijing language and culture University Press, 2007.