A Study on the Robustness of Pitch Range Estimation from Brief Speech Segments

Wenjie Peng, Kaiqi Fu, Wei Zhang, Yanlu Xie, Jinsong Zhang
Beijing Advanced Innovation Center for Language Resource
Beijing Language and Culture University
Beijing, China
{wenjayep, kaiq.fu}@gmail.com, wei_zhang_mail@163.com, {xieyanlu, jinsong.zhang}@blcu.edu.cn

***Abstract*—Pitch range estimation from brief speech segments is important for many tasks like automatic speech recognition. To address this issue, previous studies have proposed to utilize deep-learning-based models to estimate pitch range with spectrum information as input [1-2]. They demonstrated it could still achieve reliable estimation results when speech segment is as brief as 300ms. In this work, we further investigate the robustness of this method. We take the following situation into account: 1) increasing the number of speakers for model training hugely; 2) second-language(L2) speech data; 3) the influence of monosyllabic utterances with different tones. We conducted experiments accordingly. Experimental results showed that: 1) We further improved the accuracy of pitch range estimation after increasing the speakers for model training. 2) The estimation accuracy on the L2 learners is similar to that on the native speakers. 3) Different tonal information has an influence on the LSTM-based model, but this influence is limited compared to the baseline method. These results may contribute to speech systems that demanding pitch features.**

Keywords-component; pitch range estimation; LSTM; brief speech segments; L2; tone

I. INTRODUCTION

In human speech communication, pitch carries important information, and different people have different pitch range. Pitch range estimation could benefit many speech systems. Previous research showed that human beings could have an accurate estimation on pitch range from a brief (<50ms [3]) speech segment [4-7]. Inspired by these findings, W. Zhang et al. proposed to employ a LSTM model with spectrum information as input to estimate speaker's pitch range targets [1]. Experimental results showed that they could achieve a reliable pitch level estimation result with low (<2.5%) mean absolute error rate (MAPE) with speech segments as brief as 300ms (about 1~1.5 syllables). Different from [1], Q. Zhang et al. have utilized multi-feature and multi-task learning deep neural network (MTL-DNN) to estimate pitch range [2]. Both studies have demonstrated that pitch range estimation could make use of deep learning technique with spectrum information as input. These findings could benefit speech systems that demanding pitch features.

In this paper, we further investigate the robustness of the method in [1] under some challenging circumstances. One of them is the influence of big scale dataset, since they trained the model on a relative small dataset previously. From the view of speech production, when the number of speakers increased, pitch range will have a lager diversity due to its speaker-dependent nature, which may put more challenges to the estimation task. From the view of speech technology, deep-learning-based models often gain from large scale dataset. Considering these two aspects, the influence of a larger scale dataset on this model seems to be unclear.

Another scenario is applications of speech technology involved with pitch features, say computer-assisted pronunciation training(CAPT) system [8-9]. In CAPT system, it is often hard to obtain massive L2 data for model training. In addition to build a special L2 corpus (i.e. [10]), it is common to train models on a native corpus then evaluate them on the L2 data. The acoustic space between native speakers and L2 learners, however, differs hugely. For CAPT systems demanding pitch features, it is well worth investigating whether this method could still provide reliable estimation results in this situation.

Pitch range estimation is intrinsically difficult from brief speech segments especially when speech samples could not contain enough pitch variations. There are four basic lexical tones (high-level, mid-rising, low-dipping and high-falling) in Chinese. In Chinese monosyllabic utterances, the phenomenon of co-articulation between syllables will disappear while the influence of tonal information on F0 values will be much more prominent. [11] found that F0 range increases gradually from high-level to high-falling. Thus it is reasonable that speech samples with more high-fallings will result in a larger F0 range. Pitch range measures based on the distribution of F0 will be affected by different tones a lot. The LSTM-based method, however, takes spectrum information rather than F0 values as input. It is unclear whether or to what extent different tonal information affect the estimation results.

Considering the above three aspects, we conducted several experiments accordingly. We first increased speakers hugely compared to previous experiment for LSTM model training. To investigate the effect of L2 and different tonal information on the LSTM-based model, we then evaluated it on a Chinese L2 of Japanese corpus and a corpus with Chinese monosyllabic utterances respectively.

This paper is organized as follows. We first review the proposed method in section 2. In section 3 and section 4, we give details about the experiments for model training and test results. Two evaluation experiments will be presented in section 5 and section 6. Discussion and conclusions will be given in section 7.

II. METHOD REVIEW

A. Pitch Range

There is a common consensus that pitch range varies along two dimensions: pitch level and pitch span [12]. Pitch level refers to the overall pitch height of voice while pitch span represents the range within which pitch varies.

To quantify pitch level, many studies suggested using the mean F0 or median F0. As for pitch span, some long term distributional(LTD) measures have been adopted based on an analysis of F0 distribution within a speaker's voice, which includes the difference between the 95th and 5th percentile, difference between the 90th and 10th, maximum minus minimum F0, four standard deviations around the mean. In addition to LTD measures, 'linguistic' measures have also been proposed and adopted [13-14], which make use the specific landmarks in the F0 contour.

B. LSTM Model

Recurrent Neural Networks(RNNs) have been successfully applied in speech processing due to their ability to use the contextual information when mapping between the input and output [15]. Unfortunately, the range of context for standard RNN that can be accessed is very limited due to the influence of a given input on the hidden layer. This will affect the output, either decays or blows up exponentially as it loops through the whole network's recurrent connections.

To address the above issues, many attempts have been proposed. Among these methods, LSTM could provide an effective way to make output reliable by adding extra memory blocks. These blocks enable the model store and access information over long periods of time, thereby preventing the outputs from vanishing or exploding.

III. Experimental Setup

A. Speech Data

We used the open-source AISHELL-2 [16] Mandarin Corpus for model training. This corpus is split into three parts, namely training, evaluation and test. The details about the speakers and content of speech are shown in Table I. All the speech data was recorded in a quiet environment.

We first tried to train the model with all the speakers in the training dataset, but ended up with poor performance for pitch range estimation. After observing the data, we found that there exists an unbalanced distribution of pitch range targets in the training dataset.

To solve this issue, we did preprocessing on the dataset. We only chose part of the whole data for model training instead. For pitch level estimation, we first located the max and min pitch level targets, and then split speakers within the range (max-min) into 100 groups equally. We then chose speakers within each group randomly. Preprocessing for pitch span estimation was the same as pitch level, except that the number of speakers we chose within each group differed, at most 12 for pitch level and 10 for pitch span respectively. Finally, we got 899 speakers for pitch level training and 693 speakers for pitch span training with 100 utterances per speaker. For testing, we used the test part in AISHELL-2 corpus, including 10 speakers (5 male and 5 female) with 1000 utterances in total. There is no overlap between training and test dataset neither at the speaker-level nor the utterance-level.

At the stage of training, 20% of training data will be used as validation set, which is a common approach in machine learning.

TABLE I. Descriptions on the training dataset

Content of speech	voice commands, places of interest, entertainment, finance, technology, sports, English spellings and free speaking without specific topic
Speaker information	There are 845 males and 1146 females 1991 speakers in total in the training dataset. The age of speaker varies from 11 to 40. As for the accents, there are 678 speakers using Southern ones, 1293 using Northern ones and 20 speakers using other ones while recording.

B. Features

We first did intensity normalization on the raw audio files using Praat [17] with default settings. We then utilized Kaldi [18] toolkit to extract time-spliced-40-dimensional Fbank features per utterance with 25ms frame length and 10ms frame shift. The raw Fbank features were further processed using voice activity detection technique to remove the silent interval segments.

F0 tracking was performed using STRAIGHT [19] algorithm at 1ms interval with a specific range of F0. We set F0 range of 50-300Hz for male and 75-500Hz for female. In this study, we first transformed the extracted F0 into logarithmic domain with base 10 rather than the raw F0 values, then we calculated the mean and standard deviation of the transferred results at the speaker-level to represent pitch level and pitch span targets respectively.

We conducted the above operations all the same on the dataset we used in this work.

C. Evaluation criteria

We adopt MAPE to evaluate the performance of LSTM-based model. The formula of MAPE is defined as below:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(y_i - \widetilde{y_i})}{y_i}\right| * 100 \quad (1)$$

where N is the total number of input samples, y_i is the i-th pitch range target (pitch level or pitch span) to be estimated, and $\widetilde{y_i}$ is the corresponding model prediction. The less value MAPE is, the more coherence that pitch range is estimated with measurements from logF0.

D. LSTM model setting

We used Keras [20] toolkit for model training. Following previous model settings in [1], the main parameters of our model are listed as below:

TABLE II. Main parameter settings in LSTM

Parameter	Batch length	Batch size	Layer size	Hidden units
value	30	50	3	100

We kept the main parameter settings the same with [1], but increased the hidden size from 50 to 100 due to the increased size of training data. Each input to the model consists of 300ms speech sample, model will not give an output until the end of last time step. The gradient will update once per 50 samples. We trained two separate LSTM models with 3 LSTM layers for pitch level and pitch span estimation. The two models have the same parameter settings but differ in pitch range estimation targets, say the mean logF0 and the standard deviation of logF0.

IV. RESULTS

The results of evaluation on our test dataset are as below.

TABLE III. EVALUATION RESULTS OF PITCH RANGE ON AISHELL-2 TEST DATASET

	Pitch level	Pitch span
MAPE	1.09%	19.37%

It can be seen from Table III that we achieved a fairly low estimation error rate of 1.09% for pitch level, while 19.37% for pitch span. Although pitch level estimation result seems more promising than that in [1] (1.09% vs 2.3%), it does not make sense to compare these two directly due to the different datasets we used. As for pitch span, the error rate was much larger than that of pith level, which may suggest that spectral structure has higher correlation with pitch level than with pitch span [1].

To verify whether increasing training dataset could decrease the error rate, we split the above two training dataset into subsets to see the effect of training data size. We randomly split the above two training datasets into three subsets, namely 25%, 50%, 75% of the original ones. We trained models on these subsets with the same parameters settings. To get a more stable results, we repeated training on each subset 6 times, and calculated the average MAPE as the corresponding result. Table IV show the average of MAPE within each training set.

TABLE IV. MAPE OF PITCH RANGE ESTIMATION RESULTS ON DIFFERENT SIZE OF TRAINING DATASET

	25%	50%	75%	100%
Pitch level	1.35%	1.25%	1.17%	1.09%
Pitch span	24.53%	19.54%	19.25%	19.37%

It can be seen from Table IV that when training dataset increases, the estimation error rate will decrease especially for pitch level estimation. For pitch span estimation, the error rate drops at first, and get the lowest error rate with 75% of the whole training dataset. The error rate increase slightly when training on the whole dataset, which indicates that a relative large dataset is good enough for pitch span model training. Both results showed that large-scale dataset for model training will bring performance gain.

With the trained model on the whole dataset, we did two other evaluation experiments.

V. EVALUATION ON L2 CORPUS

A. Data

We first evaluated this method on Conversational Chinese 301 [21]. 19 native Japanese speakers (10 males, 9 females) were taken part in this study and they were told to read the Mandarin materials in a natural way. The average number of utterance per speaker is 301.

B. Evaluation Results and Analysis

To make the method comparable, we calculated the mean and standard deviation of logF0 within a speaker as the ground-truth labels for pitch level and pitch span respectively. We compared estimation results from the model with the ground-truth labels to investigate model performance under the effect of second-language, which are shown as below in Fig. 1 and Fig. 2 respectively.

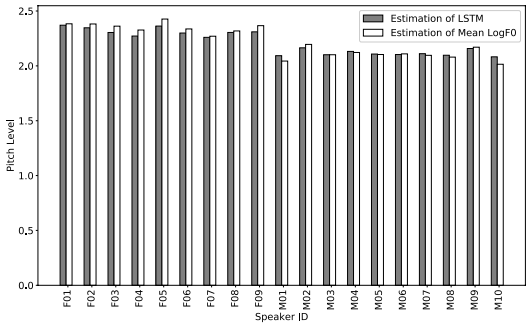

Figure. 1 Measurements of pitch level by LSTM and Mean logF0 (i.e. M02 means Japanese male with id 02, while F02 means Japanese female with id 02)

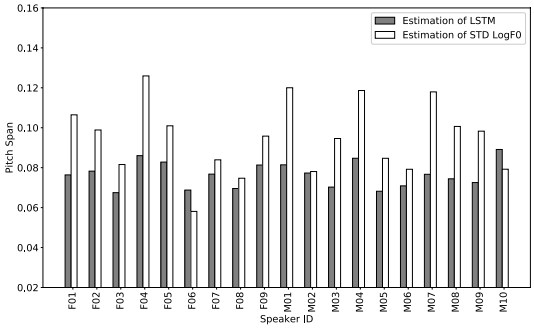

Figure.2 Measurements of pitch span by LSTM and Standard Deviation of logF0

It can be seen from Fig. 1 that there was a high coherence between model prediction and the measures of mean logF0. Pitch level varies across different speakers, but model's estimation results are very close to the ground-truth labels per speaker. Paired t-test shows that there is no significant difference between these two results (t=-1.476, df=18, p=0.157, Estimation of LSTM – Estimation of Mean LogF0). As for pitch span showed in Fig. 2, there seems to be a large difference between these two methods.

To quantify such difference, we calculated MAPE for pitch level and pitch span estimation respectively. And the MAPE for pitch level is 2.11% while 22.28% for pitch span, which is slightly higher than the error rate tested on the AISHELL-2 test dataset. The overall estimation MAPE of pitch level and pitch span tested on native Japanese

speakers seems to be close to that tested on the native Mandarin speakers. These results suggested that this method could still provide a relative reliable estimation results under the influence of second-language.

VI. Evaluation on Monosyllabic-utterance Corpus

A. Data

For this evaluation, we used part data from the BLCU-SAIT corpus [22]. We chose 35 native speakers (15 males, 20 females) with monosyllabic utterances per speaker for evaluation. The average number of monosyllabic utterances is 1085. The speech content covers the four basic tones in Mandarin (we removed some utterances with a neutral tone)

B. Baseline Method

We first established a baseline method to evaluate the performance of LSTM model under the effect of different tonal information. In this evaluation experiment, we used the mean and standard deviation of logF0 calculated at the monosyllabic-utterance-level as estimation results for baseline. We did not adopt 'linguistic' measures as baseline because it is hard to spot specific landmarks in monosyllabic utterances especially in utterances with Tone1.

C. Ground-truth Labels

In order to compare our model with baseline method, we then need to specify the ground-truth of pitch range per speaker. Unlike the ground-truth setup in the former evaluation experiment, we did this based on extra data that speaker produced. In practice, we did the calculation on two other part data belonging to that speaker, which includes 103 declarative sentences, 237 bi-syllable utterances together with the above monosyllabic utterances.

D. Evaluation Results and Analysis

Table V and Table VI show the MAPE results of pitch level and pitch span estimated from the baseline method and the LSTM-based method respectively.

TABLE V. Comparisons of MAPE on pitch level estimation between baseline and LSTM model

	Tone1	Tone2	Tone3	Tone4	Overall
Baseline	3.83%	1.25%	3.82%	1.71%	2.75%
LSTM	2.89%	1.74%	2.12%	2.19%	2.23%

TABLE VI. Comparisons of MAPE on pitch span estimation between baseline and LSTM model

	Tone1	Tone2	Tone3	Tone4	Overall
Baseline	78.75%	44.06%	38.68%	25.28%	47.04%
LSTM	32.79%	27.14%	22.99%	19.08%	25.59%

As is shown in Table V, LSTM-based method achieved a slightly lower error rate for pitch level estimation compared against baseline method (2.23% vs 2.75%). The top1 and the lowest error rate came from the cases of Tone1 and Tone2 respectively in both methods. These may be associated with the characteristics of 4 different Mandarin lexical tones, among which the production of Tone1 is at high level while Tone2 is at a relative median level.

As for pitch span showed in Table VI, LSTM-based model also outperformed the baseline method with overall MAPE of 25.59% compared against 47.04%. Besides, both methods achieved the lowest error rate in the case of Tone4. This may be due to the fact that Tone4 is involved with a larger range of F0 values compared to the rest three tones.

The varied estimation error rate among four tones indicated that different tonal information has an influence on the estimation results. Compared to the baseline method, however, LSTM-based model's lower variance of error rate among different tonal information suggested that such influence was relative limited than that in the baseline method. This difference may be due to the fact that LSTM-based model takes spectrum information as input rather than F0 values.

VII. Discussions And Conclusions

In this study, we did the following three attempts to further investigate the robustness of LSTM-based pitch range estimation method.

First, the large-scale data. We show large-scale data will decrease the estimation error rate especially for pitch level estimation. With the increased speakers, we further improved the accuracy of pitch range estimation.

Second, the effect of second-language. Evaluation results on the Mandarin L2 of Japanese speakers showed that LSTM-based method could still maintain a high accuracy for pitch level estimation under the influence of second-language. The MAPE for pitch range estimation on L2 data is similar to that on the native data.

Third, the influence of different tonal information. Experimental results of Chinese syllabic utterances showed that LSTM-based method could be affected under such effect, but the influence is relatively limited comparing to the baseline method.

These results further verified that the robustness of pitch range estimation by utilizing LSTM with spectrum information as input. The experimental results suggest that it could contribute to speech systems involved with pitch features. However, the MAPE tested on monosyllabic utterances is larger than that on AISHELL-2 native corpus, which may be due to the mismatch between different corpus. Besides, the error rate of pitch span estimation is much larger than that of pitch level in our experiment. Future work should focus on the above two issues.

Acknowledgement

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Fundamental Research Funds for the Central Universities (16ZDJ03, 18YJ030006), and the project of "Intelligent Speech Technology International Exchange". Jinsong Zhang is the corresponding author.

REFERENCES

[1] W. Zhang, et al. "LSTM-Based Pitch Range Estimation from Spectral Information of Brief Speech Input." *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018.

[2] Q. Zhang, et al. "Pitch Range Estimation with Multi features and MTL-DNN Model." *2018 14th IEEE International Conference on Signal Processing (ICSP)*. IEEE, 2018.

[3] C. Y. Lee, "Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study," *Journal of the Acoustical Society of America*, vol. 125, no. 2, pp.1125-1137, 2009.

[4] Moore, B. Corinne, and A. Jongman, "Speaker normalization in the perception of Mandarin Chinese tones." *The Journal of the Acoustical Society of America* 102.3: 1864-1877, 1997.

[5] D. N. Honorof and D. H. Whalen, "Perception of pitch location within a speaker's F0 range," *The Journal of the Acoustical Society of America*, vol. 117, pp. 2193–2200, 2005.

[6] 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.

[7] J. Kuang and M. Liberman, "Influence of spectral cues on the perception of pitch height," *Proceeding of ICPhS, Glasgow, UK*, 2015.

[8] D. Chun, Signal analysis software for teaching discourse intonation, Language Learning and Technology 2, 61-77, 1998, http://llt.msu.edu/vol2num1/ article4/index.html

[9] J. Kommissarchik, E. Komissarchik, (2000) Better Accent Tutor-Analysis and visualization of speech prosody, *Proceedings of InSTILL, Dundee, Scotland*, 86-89, 2000.

[10] N. F. Chen et al., "Large-Scale Characterization of Non-Native Mandarin Chinese Spoken by Speakers of European Origin: An Analysis on iCALL," *Speech Communication*, 2016.

[11] H. Liu, "The acoustic-phonetic characteristics of infant-directed speech in Mandarin Chinese and their relation to infant speech perception in the first year of life," pp. 3687-3687, 2003.

[12] D. R. Ladd, "Intonational Phonology," *Cambridge: Cambridge University Press,* 1996.

[13] I. Mennen, F. Schaeffler, and G. Docherty, "A methodological study into the linguistic dimensions of pitch range differences between German and English," *Proceedings of the 4th Conference on Speech Prosody, Campinas.* 2008.

[14] D. J. Patterson, "Linguistic approach to pitch range modelling," 2000.

[15] A. Graves, M. Abdel-rahman, and H. Geoffrey, "Speech recognition with deep recurrent neural networks," *2013 IEEE international conference on acoustics, speech and signal processing. IEEE*, 2013.

[16] J. Du, X. Na, X. Liu and H. Bu, "AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale," *arXiv preprint arXiv*, pp. 1808:10583, 2018.

[17] P. Boersma, "Praat: doing phonetics by computer." *http://www. praat. org/* (2006).

[18] D. Povey, et al. "The Kaldi speech recognition toolkit.", *No. CONF. IEEE Signal Processing Society*, 2011.

[19] H. Kawahara, A. Cheveigne and R. D. Patterson. "An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: revised TEMPO in the STRAIGHT-suite," *Fifth International Conference on Spoken Language Processing*, 1998.

[20] P. W. D. Charles, "KERAS," GitHub Repository, https://github.com/charlespwd/keras, 2013.

[21] Conversational Chinese 301[M]. Beijing Language and Culture University Press, 2005.

[22] W. Wang, X. Wei, J Yu, W. Wei, Y. Xie and J Zhang, "The BlCU-SAIT Speech Corpus Of Non-Native Chinese," *Oriental COCOSDA 2018, Miyazaki, Japan, May*, 2018.