Interspeech 2018
2-6 September 2018, Hyderabad

Improving Mandarin Tone Recognition using Convolutional Bidirectional Long Short-Term Memory with Attention

Longfei Yang[1], *Yanlu Xie*[2], Jinsong Zhang[2, *]

Beijing Advanced Innovation Center for Language Resources, Beijing Language and Culture University, Beijing 100083, China

yanglongfei908@gmail.com, {xyl, jinsong.zhang}@blcu.edu.cn

Abstract

Automatic tone recognition is useful for Mandarin spoken language processing. However, the complex F0 variations from the tone co-articulations and the interplay effects among tonality make it rather difficult to perform tone recognition of Chinese continuous speech. This paper explored the application of Bidirectional Long Short-Term Memory (BLSTM), which had the capability of modeling time series, to Mandarin tone recognition to handle the tone variations in continuous speech. In addition, we introduced attention mechanism to guide the model to select the suitable context information. The experimental results showed that the performance of proposed CNN-BLSTM with attention mechanism was the best and it achieved the tone error rate (TER) of 9.30% with a 17.6% relative error reduction from the DNN baseline system with TER of 11.28%. It demonstrated that our proposed model was more effective to handle the complex F0 variations than other models.

Index Terms: spoken language processing, Mandarin tone recognition, long short-term memory, attention mechanisms, deep learning

1. Introduction

As we know, Mandarin Chinese is a syllabic and tonal language. Each Mandarin word is made up of a basic syllable and the attached tone. There are four basic lexical tone patterns which are reflected in F0 contours (as shown in Fig. 1), i.e., Tone 1 (high-level), Tone 2 (middle-rising), Tone 3 (low-dipping) and Tone 4 (high-falling), and a neutral one. Tone 1-4 attached to the same syllable /xi/ have different meanings, e.g., xi1 ("*west*"), xi2 ("*learn*"), xi3 ("*lave*"), xi4 ("*thin*"), so lexical tones play an important role in distinguishing ambiguous words and syllables. Besides, tone recognition is also useful for language learners whose mother tongue is not tonal language to learning Mandarin Chinese. There are many important applications for the automatic recognition of tones: prosodic labeling (tones) of available speech databases, voice name dialing and computer aided language learning (CALL) systems for foreign Chinese learners, etc. [1].

It was found that high performance was easy to achieve in the tone recognition of isolated syllables or short words, but continuous speech presented difficulties that resulted in a much lower performance [1]. First of all, *tone sandhi* is a rule that causes mandarin tones to change in the certain situations [2]. For example, the compound syllables "ni3 hao3" may turn into "ni2 hao3", in which one half Tone 3 falls but does not rise. Secondly, tone patterns are context dependent [3]. The tone shape of one syllable is affected by the neighboring syllables and its F0 contour is different with its isolated pattern. This phenomenon is known as *tonal co-articulation* [4]. For instance, concatenated syllables with Tone 4+Tone 4 follow this rule. The maximum F0 value of the second Tone 4 is lower than that of the first Tone 4 and the minimum F0 value of the first Tone 4 is higher than its isolated pattern. It is because that human's articulatory organs cannot produce transient movements. Furthermore, F0 level and contour are also subject to other factors such as background noise, different speakers and speaking styles and topic-shift effects [3,5,6]. From the above, many variations cause troubles to tone recognition.

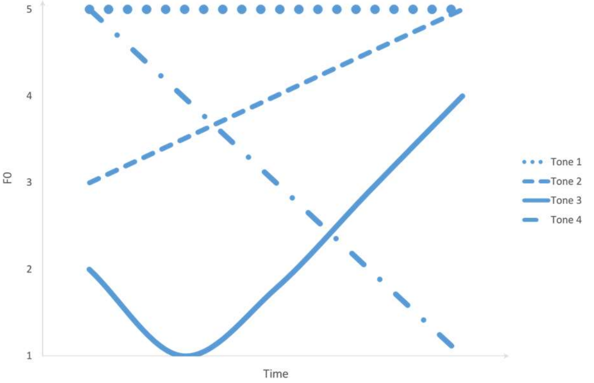

Figure 1: *Five-bar scale F0 patterns of the four basic lexical tones.*

Recently, automatic tone recognition has attracted some attentions and research. Motivated by the success of deep learning technology, some deep learning models have been applied to tone recognition. [7, 8] applies DNN to tone recognition on female corpus and some good results are achieved. More recently, [9] employs Convolutional Neural Network (CNN) for speech evaluation of the hearing-impaired population. However, feedforward neural networks like DNN and CNN are not designed to model time-series so that it is difficult to handle the F0 variations especially in continuous speech.

In this work, we applied Bidirectional Recurrent Neural Network with Long Short-Term Memory (BLSTM) units, which showed its capability of modeling temporal sequences in speech recognition task [10, 11], to tone recognition in continuous speech. The reason for adopting BLSTM is to deal with the bidirectional effect from neighboring tones. According to the previous researches on Mandarin Chinese tone, the neighboring tones interfere with each other extensively and this effect is bidirectional [4]. And not only

10.21437/Interspeech.2018-2561

the initial point and the termination point but also the entirety of the F0 curve are affected. We can make full use of BLSTM's ability to model temporal variation to handle this variation. In addition, attention mechanism was explored and it can guide the model to focus on relevant parts of the input sequence more than the irrelevant parts when doing a prediction task. Attention has also been applied to machine translation, computer vision, natural language processing and speech recognition areas and achieves a state-of-the-art performance [12-15]. According to [16], the load of tone is not distributed uniformly and the F0 contour can be divided into two parts, i.e., transition part and target part. And the target part, which is called *tone nucleus*, provides important clues for tone perception. According to the *tone nucleus* theory, we can introduce attention mechanism to guide the models to pay more attention to relevant parts for tone recognition.

In this paper, we proposed CNN-BLSTM model with attention mechanism. CNN extracted robust features to deal with spectral variation in speech signal by means of local connectivity and weight sharing and BLSTM handled the variation along the time axis. Attention mechanism was employed to capture appropriate context information.

2. Method

Our model was mainly constructed with bidirectional recurrent neural layers with LSTM unit (BLSTM) and attention layers. A part of model was shown in Fig. 2.

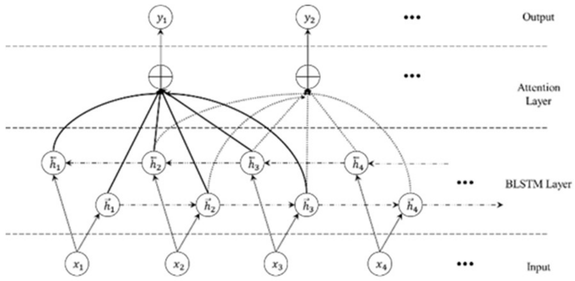

Figure 2: *Description of BLSTM layers with Attention mechanism.*

2.1. BLSTM

LSTM unit is proposed to overcome the *gradient vanishing* problem [17] and it consists of the characteristic units, *memory blocks,* in the recurrent hidden layers, and the *gate* mechanism. The *Memory cells,* components of the memory blocks, keep the temporal state of the network and the *gates* control the flow of activation information. The *input gate* controls the extent to which a new value flows into the cell, the *forget gate* controls the extent to which a value remains in the cell and the *output gate* controls the extent to which the value in the cell is used to compute the activation calculation of the LSTM.

Typically, given the sequence of input $X = (x_1, \dots, x_T)$, the LSTM layer computes activation information of the gates and cells from the time step $t = 1$ to T. The computation at the time step t can be demonstrated as (forward):

$$i_t = \sigma\left(W_{ix}x_t + W_{ih}\vec{h}_{t-1} + W_{ic}c_{t-1} + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_{fx}x_t + W_{fh}\vec{h}_{t-1} + W_{fc}c_{t-1} + b_f\right) \tag{2}$$

$$c_t = i_t\emptyset\left(W_{cx}x_t + W_{ch}\vec{h}_{t-1} + b_c\right) + f_tc_{t-1} \tag{3}$$

$$o_t = \sigma\left(W_{ox}x_t + W_{oh}\vec{h}_{t-1} + W_{oc}c_t + b_o\right) \tag{4}$$

$$\vec{h}_t = o_t\emptyset(c_t) \tag{5}$$

As for backward, we just compute $\overleftarrow{h}_t$ in the opposite direction. The next hidden layer concatenates the states of the forward and backward layers at time t:

$$h_t = \left[\vec{h}_t, \overleftarrow{h}_t\right] \tag{6}$$

Where i_t , f_t , o_t , c_t denote the output of four main components, i.e., input gate, forget gate, output gate, memory cell respectively. The W_{ix}, W_{ih}, W_{ic}, b_i are corresponding weight matrices between input gate i_t and inputs; the W_{fx}, W_{fh}, W_{fc}, b_f are that corresponding to forget gate f_t; the W_{ox}, W_{oh}, W_{oc}, b_o of output gate o_t to generate extents using current input x_t, the state h_{t-1} generated at previous time step and current state of cell c_{t-1}, which is known as *peephole connection*. It means that the current state of cell c_t take a decision to accept or forget an extent of inputs and memories kept before and output the state o_t as equal (4). It learns representations for each input x_t that depended on both the preceding and following context as equal (6).

2.2. Attention mechanism

As in Fig. 2, an attention layer is introduced. In detail, attention layer accepts the output of forward and backward layers and learns the extent to which should be paid more attention to than others in the sequence.

For each input vector h_i in a sequence of input h, the attention weights α_i can be computed as:

$$e_{ij} = \omega^{\mathrm{T}}\tanh(Ws_{i-1} + Vh_j + U\alpha_{i-1} + b) \tag{7}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{j=1}^{L}\exp(e_{ij})} \tag{8}$$

$$g_i = \sum_{j=1}^{L}\alpha_{ij}h_j \tag{9}$$

$$y_i = G(s_{i-1}, g_i) \tag{10}$$

Where α_{ij} is the attention weight. e_{ij} denotes attention scores which depends on the hidden state h_t, the $(i-1)$-th state s_{i-1} and the attention history α_{i-1}. g_i is called *glimpse*, which is the output of attention layer. $G(\cdot)$ stands for LSTM model that predicts state labels. In our experiment, at each time i, we made model consider only a subsequence through a window in the sequence h [15]. The width of window was explored.

3. Experiment setup

3.1. Dataset

3.1.1. Corpus

We adopted Chinese National Hi-Tech Project 863 corpus [18], which was used for LVCSR system development. There were 863 audio segments for each speaker, which were recorded with a 16000Hz sampling rate. Data from 148 speakers, including 74 females and 74 males, were used for training and that of 18 speakers, including 9 females and 9 males, for testing. There were no overlap of speakers and text between training and testing set. The detail was shown in Table 1.

Table 1: *Description of Training/Testing set.*

	Training set	**Testing set**
Duration	101h	6h
Speakers	148	18
Utterance (total)	42748	5625
Average length per utterance	12	

3.1.2. Data augmentation

To achieve a robust recognition system, data augmentation method proposed by [19] was employed. Vocal tract length perturbation (VTLP) with factors of {0.9, 1.0, 1.1}, tempo perturbation with factors of {0.9, 1.0, 1.1} and speaking rate perturbation with factors of {0.9, 1.0, 1.1} were adopted. And then perturbed data and the original were combined into the unified training set.

3.2. Model setup

A dataflow diagram of the proposed CNN-BLSTM with attention was shown in Fig 3.

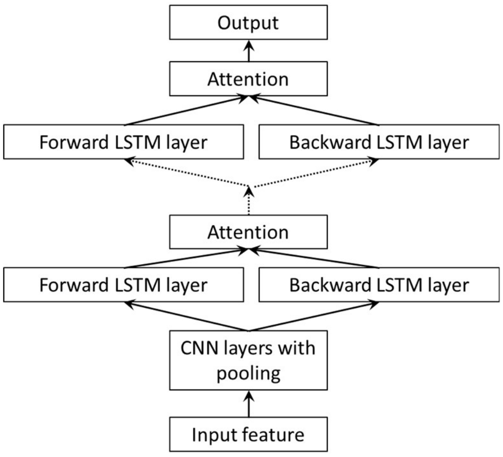

Figure 3: *Description of BLSTM layers with Attention mechanism.*

3.2.1. Input features

We employed a 43-dimension concatenated feature vector consisting of 40-dimension Mel-Frequency Cepstral Coefficient (MFCC) feature and 3-dimension F0 feature, including Probability of Voicing (POV) feature warped by normalized cross correlation function (NCCF), pitch feature and delta pitch feature. Spectral features were extracted through a 25ms window with a 10ms frame shift. All the feature sets were applied cepstral mean and variance normalization (CMVN).

3.2.2. Model setup

DNN was established as one baseline. The DNN had 6 hidden layers and each layer consisted of 1024 Rectified Linear Units (ReLUs). The network was trained using the mini-batch stochastic gradient descent (SGD) based back-propagation (BP) algorithm to minimize the cross-entropy (CE) loss. The input consisted of 10 preceding frames, current frame and 10 succeeding frames and each frame had 43-dimensional feature as mentioned in *3.2.1*. The batch size was set to 128 and an exponentially decaying schedule which began with an initial learning rate of 0.001 is adopted.

The LSTM and BLSTM baseline were established with 3 layers, in which each layer had a forward layer with 1024 cells, a projection layer with 256 units for dimensionality reduction. BLSTM also had a backward layer which had the same configuration with the forward layer except the opposite direction. The input was the super-vector consisting of one preceding frame, current frame and one succeeding frame and each frame was 43-dimentional MFCC feature vector as mentioned above. The LSTM and BLSTM was unfolded 20 steps in time. The network was trained using the SGD based back-propagation through time (BPTT) algorithm.

The input feature vector of proposed CNN-BLSTM with attention model was the same input as that of BLSTM baseline. The CNN was constructed with 3 convolutional layers and each layer had 64, 128 and 256 convolutional kernels. We used filter whose size was 3 and only did convolutional operation along frequency axis. The pooling layers following each convolutional layer employed size-of-2 max pooling unit. If necessary, zero padding was employed for some layers. Three BLSTM layers had the same configuration with BLSTM baseline. As for CNN-BLSTM with attention, attention layers concatenated the hidden states from both forward and backward layers at each time step as mentioned in Section. 2. The scoring layer in Eq. (7) had 1024 hidden units and we mapped the hidden states to the attention layer. The width of windows mentioned in 2.2 was explored.

4. Results and Discussions

The experimental results were shown in Table 2.

Table 2: *TER of different systems (ATT denotes Attention).*

System	**Overall**	**Five Tones**
DNN baseline	8.27%	11.28%
LSTM baseline	8.05%	11.09%
BLSTM baseline	7.30%	10.87%
CNN-BLSTM	7.03%	10.72%
ATT-BLSTM	6.24%	10.27%
CNN-BLSTM with ATT (window width: 10)	**5.35%**	**9.30%**

As shown in Table. 2, the TER of the five tones were higher than that of the overall since the silence and other unvoiced regions were relatively easy to be recognized. Therefore, we considered the TER of five tones as the main measure in order to demonstrate the performance of tone recognition later. Results in Table. 2 showed that BLSTM based models achieved better performance than DNN model. And the performance of BLSTM could go a step further by sending the output of convolutional layer to BLSTM layers. This kept consistency with [20]. In addition, introducing attention mechanism reduced the error rate significantly and finally our proposed CNN-BLSTM with attention mechanism achieved the best result with a relative reduction by 17.6% in TER of five tones from 11.28% by DNN to 9.30%. In detail, the results of three types (insertion, deletion and substitution) were shown in Fig. 4. Each error type was computed over all of samples including five tones and non-tone label. It was seen that comparing to the DNN baseline, our proposed model achieved a relative reduction by 85.2%, 41% and 23%, a relative reduction by 66.7%, 40.4% and 20.3% comparing to the LSTM baseline and a relative reduction by 50%, 38.5% and 17.7% comparing to the BLSTM baseline.

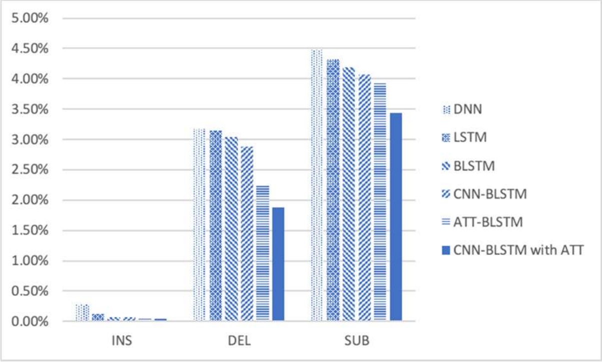

Figure 4: *Comparison of three error types between different systems.*

Table. 3, Table. 4, Table. 5 and Table. 6 showed the confusion matrix of four tones for DNN baseline, LSTM baseline, BLSTM baseline and our proposed model. By comparison in detail, we found that BLSTM based model significantly improved the detection accuracy especially for that of Tone 3. The effect of *co-articulation* from adjacent syllables made the F0 contour of Tone 3 deform a lot and sometimes the contour of Tone 3 was incomplete so that it might lead DNN mistakenly recognize as other tone patterns. Benefiting from the connection between hidden layer, LSTM had the stronger ability to model time-series so that it worked better than DNN. And, BLSTM could handle the effect from both preceding syllables and following syllables so it performed better than unidirectional LSTM. The accuracy of detecting some easily-confused, which was caused by some rules, tone pairs such as Tone 2-Tone 3 and Tone 3-Tone 4 was also improved by BLSTM based model. In addition, attention further improved the recognition performance and our proposed model achieved the best result.

Table 3: *Confusion matrix for DNN baseline.*

	Tone 1	**Tone 2**	**Tone 3**	**Tone 4**
Tone 1	92.0%	2.9%	0.8%	3.3%
Tone 2	3.0%	90.1%	4.7%	1.7%
Tone 3	1.2%	10.5%	82.7%	3.7%
Tone 4	2.4%	1.2%	2.8%	93.6%

Table 4: *Confusion matrix for LSTM baseline.*

	Tone 1	**Tone 2**	**Tone 3**	**Tone 4**
Tone 1	92.7%	2.9%	0.5%	3.4%
Tone 2	2.9%	90.2%	4.5%	2.0%
Tone 3	1.1%	9.9%	83.9%	2.6%
Tone 4	2.4%	1.3%	2.2%	94.0%

Table 5: *Confusion matrix for BLSTM baseline.*

	Tone 1	**Tone 2**	**Tone 3**	**Tone 4**
Tone 1	93.3%	2.6%	0.4%	3.5%
Tone 2	3.1%	90.7%	4.3%	1.8%
Tone 3	0.8%	8.5%	84.5%	2.4%
Tone 4	2.0%	1.0%	1.4%	95.6%

Table 6: *Confusion matrix for CNN-BLSTM with ATT.*

	Tone 1	**Tone 2**	**Tone 3**	**Tone 4**
Tone 1	94.0%	2.4%	0.3%	3.0%
Tone 2	3.0%	91.1%	3.9%	1.9%
Tone 3	0.8%	8.0%	85.8%	2.3%
Tone 4	2.2%	1.0%	1.4%	95.3%

We also explored the effect of different windows widths on the performance of tone recognition. Fig. 5 and Fig. 6 showed the line chart of results. We found that the window width of 10 was the best. Smaller window might lead the model to utilize lacking context information inside one tone or between tone pairs. However, it seemed that the larger window had a slight effect on the performance.

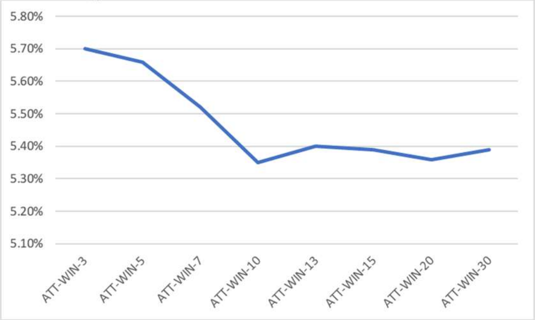

Figure 5: *Overall TER with different windows widths for attention.*

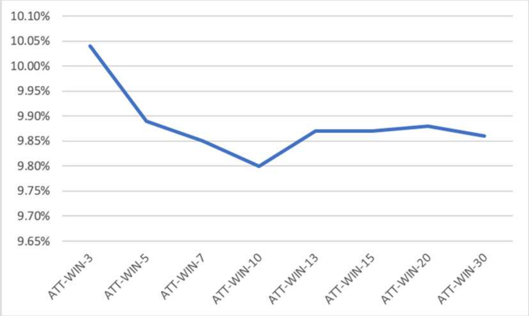

Figure 6: *Five Tone's TER with different windows widths for attention.*

5. Conclusion

In this paper, we explored a LSTM based architecture which utilized the strong capability of modeling time sequence in tone recognition task in which there are many variations affect the recognition performance. As a result, the CNN-BLSTM with attention mechanism achieved the five-tone error rate (TER) of 9.3% which is the best result comparing to other models.

6. Acknowledgements

This work is supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), Research Funds of State Language Commission (ZDI135-51), Wutong Innovation Platform of Beijing Language and Culture University (16PT05), the Fundamental Research Funds for the Central Universities (18YJ030006). The asterisked author is the corresponding author.

7. References

[1] J. Zhang, S. Nakamura and S. Hirose. (2005). Tone nucleus-based multi-level robust acoustic tonal modeling of sentential F0 variations for Chinese continuous speech tone recognition. Speech Communication. 46. 440-454, 2005.

[2] L. S. Lee, C. Y. Tseng, and C. J. Hsieh, “Improved tone concatenation rules in a formant-based Chinese text-to-speech system,” Speech and Audio Processing, vol.1, no.3, pp. 287-294, 1993.

[3] C. Shih, “The phonetics of the Chinese tonal system,” AT&T Bell labs technical memo, 1987.

[4] Y. Xu, “Contextual tonal variations in Mandarin.” J. Phonetics 25, 61–83, 1997.

[5] N. Umeda, “F0 declination is situation dependent,” Journal of Phonetics, Vol. 10, no.3, pp. 279-290, 1982.

[6] Y. Xu, “Effects of tone and focus on the formation and alignment of F0 contours,” Journal of Phonetics, vol. 27, no. 1, pp. 55-105, 1999.

[7] M. Chen, Z. Yang, and W. Liu, “Deep neural networks for Mandarin tone recognition,” International Joint Conference on Neural Networks, IEEE, pp. 1152-1158, 2014.

[8] Lin, Ju, et al. "Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks." Journal of Signal Processing Systems (2018): 1-11.

[9] C. Chen, B. Razvan, et al., “Tone classification in Mandarin Chinese using Convolutional Neural Networks,” INTERSPEECH, 2016.

[10] H. Sak, A. Senior, and F. Beaufays, “Long Short-Term Memory Recurrent Neural Network Architecture for Large Scale Acoustic Modeling,” INTERSPEECH,2014.

[11] A. Graves, N. Jaitly, A. R. Mohamed, “Hybrid speech recognition with Deep Bidirectional LSTM,” Automatic Speech Recognition and Understanding, IEEE, pp. 273-278, 2014.

[12] D. Bahdanau, K. Cho, Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” Computer Science, 2014.

[13] K. Xu, J. Ba, R. Kiros, et.al., “Attend and Tell: Neural Image Caption Generation with Visual Attention,” Computer Science, pp. 2048-2057, 2015.

[14] V. Mnih, N. Heess, A. Graves, et.al., “Recurrent models of visual attention” Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), Vol. 2, pp. 2204-2212, 2014.

[15] J. Chorowski, D. Bahdanau, D. Serdyuk, et.al., “Attention-Based Models for Speech Recognition,” Computer Science, 2015.

[16] J. Zhang, K. Hirose, “Tone nucleus modeling for Chinese lexical tone recognition,” Speech Communication, pp. 447-466, 2004.

[17] F. A. Gers, J Schmidhuber and F. Cummins, “Learning to Forget: Continual Prediction with LSTM,” Neural Computation, pp.2451-2471, 2000.

[18]] S. Gao, B. Xu, H. Zhang and T. Y. Huang, “Update of Progress of Sinohear: Adavanced mandarin LVCSR System At NLPR,” ICSLP, 2000.

[19] T. Ko, V. Peddinti, D. Povey, et.al., “Audio augmentation for speech recognition,” INTERSPEECH, pp.3586-3589, 2015.

[20] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio, “How to Construct Deep Recurrent Neural Networks,” ICLR, 2014.