Tagging Tone for Mandarin Pinyin Based on Sequence Labelling

Zhaopeng Qian¹*, and Kejing Xiao²

¹School of Biological Science & Medical Engineering, Beihang University, Beijing, 100191, China
²Information School, Renmin University of China, Beijing, 100872, China

Abstract. Generally, fundamental frequency (F0) is applied as the clue for tone recognition in normal speech. Tone recognition in whisper speech without F0 could be based on the temporal and spectral cues. However, the Mandarin Electro-Laryngeal (EL) speech with fixed F0 has no tone information. Therefore, the tone recognition of Mandarin EL speech is so difficult. And the researches about tone recognition for Mandarin EL speech is insufficient. In this paper, a new method labelling the tone for pinyin is proposed based on the context information to identify the tone of Mandarin speech without tone information. The experiment result shows that the precision, recall and the F value are all above 97% based on the test dataset. The amount of semantic information influences the performance of proposed method. If the amount of semantic information is little, the accuracy of tone labelling would be poor. The result shows that the proposed method has a good precision and robustness. The method can label the tone for pinyin without any tone information only based on the context information. The proposed method can label tones for tonal language, such as the Mandarin speech.

1 Introduction

Tone plays an import role in Mandarin speech, even influencing the understanding of semantic information [1]. In addition of four tones including Tone1 (T1), Tone2 (T2), Tone3 (T3) and Tone4 (T4), the Mandarin speech also has the fifth tone (T0), usually used to weaken the tone in whole sentence [2]. Tone identification is very useful for Speech Recognition [3], Speech synthesis [4] and the Cochlear Implant [5]. Tone identification is mainly based on the fundamental frequency (F0) or pitch [6, 7]. A large amount of researches about tone recognition has been done for healthy speech [8-11], however, not all kinds of the speech has whole F0 information, such as whisper speech [12], esophageal speech [13] and electro-laryngeal speech [14]. Therefore, tone identification for speech without F0 information is a very difficult task.

Some researches explain that besides of F0 contour, temporal and spectral parameters also contain the tone information of speech [15,16]. Therefore, the temporal and spectral cues can be also used to recognize tones from whisper speech. Chen X et al. [17] proposed an Ant-Colony-Clustering-Neural-Networks based method to recognize the Mandarin whisper speech tones. However, the Electro-Laryngeal (EL) speech differs from the whisper speech a lot. The EL speech is generated by laryngectomee using Electro-Larynx.
Especially the Electro-Larynx with fixed F0 is the most popular auxiliary sound device. And the EL speech has no tone information because of the fixed F0. It is very hard to pronounce the continuous Mandarin EL speech with tone variation by finger. Therefore, tone identification for the Mandarin EL speech based on the acoustic features cannot be achieved, due to the Mandarin EL speech has no tone information.

Advanced technology based on deep learning makes it possible that the Mandarin EL speech can be recognized as the pinyin sequence [18]. Therefore, tone identification for Mandarin EL speech has become possible according to the tone clues hiding in semantic information of context. Labelling tone for pinyin can be achieved based on machine learning according to context information. Certainly pinyin sequence contains the whole semantic information of the sentence and the tones information of the all syllables. The most popular sequence labelling tools are Conditional Random Field (CRF) [19], Hidden Markov Model (HMM) [20] and the Maximum Entropy Markov Model (MEMM) [21].

State-of-the-art sequence labelling model [22, 23] contains the Bi-direction Long Short Temporal Memory Unit (Bi-LSTM), Convolutional Neural Networks (CNN) and CRF. However, in this paper only the Word Embedding is used to represent the word vectors of pinyin enough, and the CNN need not be used. Therefore, the Bi-LSTM-CRF based end-to-end technology is used to label tone for pinyin.

The most important contribution of this paper is the work making tone identification for Mandarin EL speech possible using sequence labelling tools based on Bi-LSTM-CRF. Comparison of classification and sequence labelling is used to show the higher performance of the proposed method. The proposed method is a novel idea based on sequence labelling tools according to tone clues hiding in semantic information of sentence. This proposed method differs from the traditional method based on acoustic tone clues, and it identifies the tone for speech only based on context.

2 Tone identification based on sequence labelling

A novel of method based on sequence labelling is proposed to identify the tone for Mandarin EL speech. In addition, the proposed method prefers to use the contextual semantic information but not classify the syllables into different tones. In this paper, the pinyin sequence of Mandarin EL speech can be regarded as a group of linear sequence, such as Equation (1),

$$X = \{x_1, x_2, \ldots, x_n\}$$

where X is pinyin sequence for Mandarin speech, $x_i$ is the i th syllable unit of PinYin. The tone set corresponding to X can be represented as $\text{Tone} = \{T0, T1, T2, T3, T4\}$. Finally, the tones sequence corresponding to the input PinYin sequence can be labelled by the proposed method.

The proposed tone identification method includes Word Embedding, Bi-LSTM and CRF. And the framework of method is shown in Figure 1.
Figure 1. Framework of tone identification for pinyin based on Bi-LSTM-CRF.

Pinyin without tone is translated into the word vectors using Word Embedding. Vectors are mapped as the state probabilities by Bi-LSTM. The probabilities are decoded to the tone label using CRF. Note that the first T1 should be modified as T4 in the example “yi1 zhang1” should be “yi4 zhang1”.

2.1 Word embedding based on Word2Vec for Pinyin

Word Embedding is a group of technologies in the field of Natural Language Processing and Representation Learning [24]. Words, phrases or even sentences can be mapped into vectors. Mathematically, Word Embedding is used to map the word from one-dimension space to a continuous low-dimension space. One of the most popular Word Embedding technology is Word2Vec [25]. In this paper, pinyin is mapped into the vectors based on Word2Vec. Gensim [26] is used as the tool to train the Word2Vec model. And the pinyin vectors can also be obtained by Gensim tool.

2.2 Sequence labelling based on Bi-LSTM-CRF

Bi-LSTM is a recurrent neural networks framework with the Long-Short-Temporal-Memory (LSTM) unit including the forward and backward information of the whole sequence. The whole process of tone identification for pinyin based on sequence labelling by Bi-LSTM is shown in Figure 1.

CRF is an undirected graph model combining the characteristics of the maximum entropy model and the HMM. CRF performs well in the sequence labelling tasks, such as words segmentation, part-of-speech tagging and named entity recognition. All of the output for sequence are the conditional probabilities corresponding to the tone label, calculated by Equation (2),

$$ p(y|z; W, b) = \frac{\prod_{i=1}^{n} \psi_i(y'_{i-1}, y_i, z)}{\sum_{y' \in Y(z)} \prod_{i=1}^{n} \psi_i(y'_{i-1}, y'_{i}),} \text{ (2)} $$

where $\psi_i(y'_{i-1}, y_i, z) = \exp(W_{y'y}^T y'_{i-1} + b_{y'y})$, $W_{y'y}^T$ is the weight matrix, $b_{y'y}$ is the bias. The above parameters are corresponding to the pair of label $(y', y)$. The maximum conditional likelihood estimation, shown as Equation (3), is used to get the parameters for training CRF. The training set consists of $\{(z_i, y_i)\}$.

$$ L(W, b) = \sum_{i} \log p(y|z; W, b) \text{ (3)}$$

The label sequence $y^*$ can be decoded by CRF, where the maximum conditional probability can be calculated by Equation (5).

$$ y^* = \arg\max_{y \in Y(z)} p(z; W, b) \text{ (4)} $$

The training and decoding process are solved by Viterbi algorithm [27, 28].
2.3 Experiment setup

The “People's Daily” open database [29] is chosen as the materials, and the Chinese materials are processed into PinYin without tone and with tone separately by pypinyin tool [30]. More than 100 thousand sentences exist in the speech materials database, where 10% is used for testing and 90% is used for training. The test examples including 20 Chinese characters, 20 Chinese words, 20 Chinese proverbs (words with 4 Chinese characters), 20 Chinese short sentences and 20 Chinese long sentences are used to test the performance of the proposed method based on the sentences with different amount of semantic information.

The precision, recall rate and F value (F1-Measure) are used to measure the performance of the proposed method based on the testing dataset of “People’s Daily”. The evaluation result is calculated statistically many times under 95% confidence.

3 Results

The comparison experiment for the traditional method and proposed method is designed based on the testing dataset of “People’s Daily”. The comparison result including the precision, the recall rate and the F value is shown in Table 1.

| Methods         | Precision | Recall | F1-Measure |
|-----------------|-----------|--------|------------|
| DT              | 66.52%    | 66.70% | 66.61%     |
| SVM             | 65.34%    | 65.50% | 65.42%     |
| ANN             | 68.80%    | 69.20% | 69.00%     |
| Bi-LSTM-CRF     | 97.03%    | 97.01% | 97.02%     |

Note. DT means the method based on Decision Tree; SVM means the Support Vector Machine; ANN means the Artificial Neural Networks; Bi-LSTM-CNN-CRF is the method we proposed.

Results of Table 1 is calculated by testing data from “People’s Daily”. Obviously, the method based on sequence labelling can achieve a high tone labelling accuracy. The accuracy of every tone is listed in Table 2.

| Tone Type | Accuracy | STD    |
|-----------|----------|--------|
| T0        | 98.70%   | 0.0490 |
| T1        | 94.50%   | 0.4518 |
| T2        | 97.14%   | 0.4698 |
| T3        | 96.07%   | 0.4698 |
| T4        | 96.55%   | 0.4381 |
| Average   | 96.07%   | 0.0098 |

The accuracy of tone may be influenced by the contextual semantic information. The sentences with different amount of semantic information are used to test the performance of the proposed method. The result is shown in Table 3.

| Tone Type            | Accuracy | STD |
|----------------------|----------|-----|
| Chinese Characters   | 40.00%   | 0.50|
| Chinese Words        | 64.17%   | 0.40|
| Chinese Proverb      | 55.00%   | 0.24|
| Chinese Short Sentences | 90.74%  | 0.10|
| Chinese Long Sentences | 89.95%  | 0.08|
| Average of Tone Accuracy | 67.97%  | 0.36|
Obviously, the tone accuracy of the Chinese short sentences is the highest, and the following is the accuracy of the Chinese long sentences. No significant differences are between the two. This result illustrates that the performance of the proposed method is severely influenced by the amount of the semantic information.

4 Discussion and conclusion

Tone identification technology plays an important role not only in speech understanding, speech recognition and speech synthesis, but also in the pathological speech lacking tone variation. The researches for Mandarin EL speech recognition has been worked, however the result is pinyin without tone information [18]. The research in this paper can add tone for pinyin (Mandarin EL speech recognition result). If TTS is combined with the pinyin with tone, the intelligibility and naturalness of Mandarin EL speech would be improved effectively.

The precision, recall rate and F value of the proposed method all arrive above 97%, much higher than the traditional method based on classification tools. Please note that tone variation is very complicated in the continuous Mandarin speech. Therefore, the performance of proposed method is limited. Maybe only the speech materials enriched by more labelled pinyin and tone can improve the accuracy. Furthermore, the results of experiment based on different amount of the semantic information show that the accuracy of Chinese Characters is low, around 40%. Accuracy of Chinese Words and Chinese Proverbs arrives above 55%; accuracy of sentence can be above 90%. The results illustrate that the proposed method can identify the tone for pinyin sequence with high accuracy based on enough semantic information.

The proposed method can identify tone for pinyin sequence using contextual semantic information effectively. In addition, the proposed method can label tone for pinyin sequence. The proposed method can solve the problems that pathological speech lacks tone information.

This study was supported by the Open Project Program of National Engineering Laboratory for Agri-product Quality Traceability, (No. AQT-2018-YB4), and Beijing Natural Science Foundation (No. 4194079).

References

1. Speer S R, Shih C L, Slowiaczek M L. Prosodic structure in language understanding: evidence from tone sandhi in Mandarin. J. Language and Speech, 32(4): 337-354, (1989).
2. Howie J M, Howie J M. Acoustical studies of Mandarin vowels and tones. M. Cambridge University Press, (1976).
3. Lin C H, Lee L S, Ting P Y. A new framework for recognition of Mandarin syllables with tones using sub-syllabic units. C. 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 2: 227-230, (1993).
4. Chomphan S, Kobayashi T. Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis. J. Speech Communication, 50(5): 392-404, (2008).
5. Wei C G, Cao K, Zeng F G. Mandarin tone recognition in cochlear-implant subjects. J. Hearing research, 197(1-2): 87-95, (2004).
6. Lee T, Ching P C, Chan L W, et al. Tone recognition of isolated Cantonese syllables. J. IEEE Transactions on speech and audio processing, 3(3): 204-209, 1995.
7. Yang W J, Lee J C, Chang Y C, et al. Hidden Markov model for Mandarin lexical tone recognition. J. IEEE Transactions on Acoustics, Speech, and Signal Processing, 36(7): 988-992, (1988).
8. Chen S H, Wang Y R. Tone recognition of continuous Mandarin speech based on neural networks. J. IEEE Transactions on speech and audio processing, 3(2): 146-150, (1995).
9. Wang H, Ho T H, Yang R C, et al. Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data. J. IEEE transactions on Speech and Audio Processing, 5(2): 195-200, (1997).
10. Peng G, Wang W S Y. Tone recognition of continuous Cantonese speech based on support vector machines. J. Speech Communication, 45(1): 49-62, (2005).
11. Liu J, He X, Mo F, et al. Study on tone classification of Chinese continuous speech in speech recognition system. C. Sixth European Conference on Speech Communication and Technology. (1999).
12. Meyer-Eppler W. Realization of prosodic features in whispered speech. J. The journal of the acoustical Society of America, 29(1): 104-106, (1957).
13. Most T, Tobin Y, Mimran R C. Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. J. Journal of communication disorders, 33(2): 165-181, (2000).
14. Liu H, Wan M, Ng M L, et al. Tonal perceptions in normal laryngeal, esophageal, and electrolaryngeal speech of Mandarin. J. Folia Phoniatrica et Logopaedica, 58(5): 340-352, (2006).
15. Fu Q J, Zeng F G. Identification of temporal envelope cues in Chinese tone recognition. J. Asia Pacific Journal of Speech, Language and Hearing, 5(1): 45-57, (2000).
16. Kong Y Y, Zeng F G. Temporal and spectral cues in Mandarin tone recognition. J. The Journal of the Acoustical Society of America, 120(5): 2830-2840, (2006).
17. Chen X, Zhao H, Yu Y. Tone Recognition of Whispered Mandarin Using Ant Colony Clustering Neural Network. J. Journal of Applied Sciences, 26(5): 511-515, (2008).
18. Qian Z, Wang L, Zhang S, Liu C, Niu H. Mandarin Electrolaryngeal Speech Recognition Based on WaveNet-CTC. J. Journal of Speech, Language, and Hearing Research, 62(7): 2203-2212, (2019).
19. Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. J. (2001).
20. Thede S M, Harper M P. A second-order hidden Markov model for part-of-speech tagging. C. Proceedings of the 37th annual meeting of the Association for Computational Linguistics. (1999).
21. McCallum A, Freitag D, Pereira F C N. Maximum Entropy Markov Models for Information Extraction and Segmentation. C. ICML, 17(2000): 591-598, (2000).
22. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. J. arXiv preprint arXiv:1508.01991, (2015).
23. Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. J. arXiv preprint arXiv:1603.01354, (2016).
24. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. C. Advances in neural information processing systems. 3111-3119, (2013).
25. Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space. J. arXiv preprint arXiv:1301.3781, (2013).

26. Khosrovian K, Pfahl D, Garousi V. GENSIM 2.0: a customizable process simulation model for software process evaluation. C. International Conference on Software Process. Springer, Berlin, Heidelberg, 294-306, (2008).

27. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.. J. IEEE transactions on Information Theory, 13(2): 260-269, (1967).

28. Forney G D. The viterbi algorithm. J. Proceedings of the IEEE, 61(3): 268-278, (1973).

29. China Core Newspapers Full-text Database (CCND), People’s Daily (2014, 1.01-1.23), from http://gb.oversea.cnki.net/kns55/brief/result.aspx?dbPrefix=CCND

30. Pypinyin. (2014). Retrieved May 3, 2015, from http://pypinyin.readthedocs.org/en/latest/index.html#api

31. Humboldt W. On language: The diversity of human language-structure and its influence on the mental development of mankind. J. (1988).