ASR-based Input Method for Postal Address Recognition in Chinese Mandarin

Ling Feng Wei and Sun Maosong

School of Computer Science and Technology
Tsinghua University, China
andrew.l.wei@gmail.com, sms@tsinghua.edu.cn

Abstract. As the automatic speech recognition technology (ASR) has becoming more and more mature, especially with statistical language modeling built with web scale data, and with the utilization of Hidden Markov Model probabilistic framework, speech recognition has become applicable to many domains and usage scenarios. In particular, speech recognition can be applied to task such as Chinese postal address recognition. This paper presents the first attempt ever, in both academic and commercial settings, to create an ASR-based input method for postal address recognition in Chinese Mandarin. By customizing the statistical language model to such domain, and incorporating the knowledge from the structural information provided by geo-topology, our language model successfully captures the signals from geographical contextual information and self-correct possible mis-recognitions. Experiment results provide evident that our approach based on speech recognition achieves a faster and a more accuracy input method compare to traditional keyboard-based input.

Keywords: Automatic Speech Recognition, Postal Address Recognition, Statistical Language Modeling, Chinese Speech Recognition

1 Introduction

The market of postal services in China has increased significantly over past few decades. According to National Bureau of Statistics of the People's Republic of China, the total number of workforce behind this market has reached beyond 700,000 people outputting a total delivery of mail and packages of more than 61 billion pieces annually. In order to provide a more reliable postal service in such a large scale, postal tracking for mail and packages has become a standard service provided by many of the shipping and postal companies today. However, to enable such service, the shipping and postal companies often need to first hire a large amount of workers to manually type in the mailing address into their logistic information system before enabling the tracking service.

Currently, there is no special input method developed dedicated to postal address input in Chinese. According to the hand written address provided by the sender, the typist then enters the address by hand using PC based input

method such as Sogou Pinyin into the logistic system. Such input process is insufficient because keyboard based Pinyin input method is slow, and Pinyin to Hanzi conversion is not perfect. Furthermore, any error made during this data entry process will introduce additional costs to the service provider.

This paper is motivated to provide an alternative, and a more efficient way of entering Chinese postal address based on using automatic speech recognition technology (ASR). A new ASR-based input method, namely Voice Postal Input (VPI) system, built from this research, will provide a faster and more accuracy input method compare to traditional keyboard typing. Therefore, shipping service providers can lower the costs needed to hire large number of typist. Ultimately, this will translate to a cheaper shipping service for everyone in China. To our knowledge, this will be the first attempt, both in academic and commercial settings, to apply ASR technology to postal address recognition in Chinese Mandarin. Our experiment results have proven the feasibility of our approach, and have shown that high recognition accuracy can be achieved.

2 Background

Data entry process has always been human intensive and time consuming. Especially in Chinese language, typing in Chinese has been a difficult task due to the nature of the language[1]. Because of its facility to learn and use, Pinyin is by far the most popular keyboard input method in China [1]. However, Pinyin input suffers from several challenges. It is slow in terms of input speed compared to voice [2], and the Pinyin to Hanzi character conversion is error prone due to typographical errors during typing and the conversion itself is far from perfect [3]. There are 410 syllables in Chinese and they correspond to over 6000 common Chinese characters. This implies that on average, each syllable corresponds to about 17 characters. As the result, a Chinese typist must then select a choice from a multiple candidate list by typing an extra number key to identify the correct character from the list [4].

Furthermore, previous study has shown that a Chinese character' s Pinyin contains 4.2 Roman characters. Based on a skilled typist, an average keystroke takes 200 ms. Then the Pinyin typing time for a Chinese character would be about 840 ms. An extra time of 450 to 600 ms is also required to identify the correct choice of character from the pinyin candidate lists [4]. Finally, the total time spend to type a Chinese character with Pinyin input method would then fall into the range of 1290 ms to 1440 ms per character.

To improve the data entry process for Chinese postal address, a new input method by voice therefore is proposed. An ASR-based input method will avoid the typographical errors from user as well as the error-prone Pinyin to Hanzi conversion. Knowing that speech is the most effective and fastest form of human communication [6], ASR-based input method will introduce significant improvement to postal address entry process.

3 The Proposed Approach

The Voice Postal Input (VPI) system we created is based on client-server based software architecture where the client captures the audio data from end user, and the computational intensive speech recognition resides on the server. Our server end integrated with two of the most popular ASR products for Chinese Mandarin, namely iFLyTek Voice Cloud and Nuance Recognizer. Although iFlyTek Voice Cloud has been successfully integrated into many products in China, and has proven to have high recognition accuracy for open speech dictation [5], but it is lacking the flexibility for the developer to customize a language model for a domain specific task. In later section, we will show customizing a language model becomes very in postal address recognition. Nuance Recognizer, on the other hand, gives the developer the ability to customize a language model for domain specific task. For flexibility and future studies, our VPI system can easily be expanded to include any additional recognizer from its server end.

Taking advantages of the structural information of Chinese postal addresses, we also introduced a process named Geo-Topology Realignment (GTR) which takes the recognition results as an input, and makes corrections to possible misrecognitions through the knowledge learned from the Chinese geographical topology. GTR correlates the relationship between province, city, district and street information and form a geographical topology offline. When GTR receives the results from the recognizers during run time, it tries to match the recognized string against the knowledge from this topology to identify possible invalid address combinations and makes correction accordingly. Finally, the output of GTR will be sent back to the web client and will be presented as final result.

4 Experiment Setup

4.1 Evaluation Data Set

The evaluation data used to test the VPI system consists of 5170 recordings. Each audio recording contains exactly one postal address randomly picked from 150 thousand valid postal addresses collected. Each postal address is read out by both male and female voice in official Chinese Mandarin without carrying any local accent. The recording session was done in a control environment with minimal background noise, and each participant was asked to speak with their normal speaking rate. Due to the business requirement from postal companies, house numbers were to be ignored. This evaluation data set is used throughout this paper to provide consistent measurements and meaningful comparisons between different processes and features.

4.2 Evaluation Metrics

Two common methods were used to evaluate the performance of recognition accuracy. word accuracy, W_{Acc} which is derived from Levenstein distance, or Edit

Distance, is used to measure word level accuracy, and Sentence Accuracy, S_{Acc} , is used to measure how well the recognizer perform in terms of the whole address. To provide further insight to how well the recognition is in terms of different geographical level, the following metrics are also introduced each measuring the accuracy at province, city, district and street level respectively.

- P_{Acc} word accuracy at province level only
- C_{Acc} word accuracy at city level only
- D_{Acc} word accuracy at district level only
- St_{Acc} word accuracy at street level only

5 Experiment Results and Analysis

5.1 Using Open Speech Recognition

In first attempt, we first establish a baseline from using iFLyTek Voice Cloud. Using such an open-speech recognizer, our system suffers from confusing similar or identical sound characters. In extreme cases, out-of-context vocabularies, in this case, being the non-postal address related words, are found in the recognition result. Table below shows some of the errors produced by iFlyTek Voice Cloud.

Table 1. Recognition errors produced by iFlyTek Voice Cloud

ID	Ortho String	Recognized String
I-1	山东省青岛市胶南市旺富路	山东省青岛市胶南市王府路
I-2	天津市东丽区川铁路	天津市东丽区穿铁路
I-3	吉林省白城市洮北区洮河大路	吉林省白城市姚北区辽河大路
I-4	浙江省杭州市桐庐县定大线	浙江省杭州市桐庐县顶大仙
I-5	上海市闵行区墨江路	上海区民航路漠江路
I-6	天津市武清区丰收路	天是无情去丰收路
I-7	四川省成都市郫县禹庙下街	对方声称都市郫县与标赛杰
I-8	四川省成都市郫县汇川街三段	吃饭尚成东是背线绘春街三段

In cases like the ID I-1 , the recognizers were confused with acoustically similar or identical street names such as “旺富路” with “王府路” while both being valid street names. In I-2, “川铁路” with “穿铁路” were phonetically identical words. However, the term “穿铁路” would most likely not to be a name of a street because of the term “穿铁” does not make much sense in Chinese language. Case I-3 and I-4 suffers the similar misrecognition issues. I-7 and I-8 demonstrate more severe errors where terms like “对方声称都市” and “吃饭尚成东是背线绘春”, which makes very little sense in Chinese language, and are out of context of postal address, were found in the recognition result. We observed that iFlyTek Voice Cloud, being an open-speech recognizer, is trying

to predict a sequence of words which are most likely to generate the observed acoustic feature without specifically considering the postal address context. More specifically, a character like “川” and “穿”, which are phonetically identical, was misrecognized since “穿” is a more popular character in Chinese language. Since open speech recognizer such as iFlyTek Voice Cloud is designed to handle general dictation task, its language model would most likely be trained with common Chinese vocabularies. As a result, Chinese common terms such as “吃饭”, “对方” and “声称” might appear many times in its training corpus, and therefore most likely have a higher probability compare to terms or characters appears in Chinese postal address.

5.2 Customized SLM for Postal Address Recognition

We further investigate on recognition performance using Nuance Recognizer, and to explore the benefits of ngram customization. In a domain specific task such as postal address recognition, being able to customize the language model become useful because it gives a better approximation on priori probability of how words are connected in the context of Chinese postal address. The intuition behind ngram customization is to train a language model with words only appeared in postal addresses in China. As the result, terms related to postal address like the city “成都市” will have a higher probability compare to a similar sounding term “声称都市” which was previously produced as a misrecognition by iFlyTek Voice Cloud.

We developed two bigram models based on Chinese character or Chinese word as a feature. In the case of Chinese postal address, we take the complete province, city, district or street name such as the province “广东省”, or the district “朝阳区” as a Chinese word. These two language models were generated from 150 thousand valid postal addresses in China. Good-turing smoothing technique was applied to both language models. Table below shows the recognition results from our character bigram model, and word bigram model.

Table 2. Comparison of errors produced by character and word model

ID	Ortho String	Character Bigram Model	Word Bigram Model
N-1	山东省青岛市胶南市旺富路	山 东 省 青 岛 市 胶 南 市 王 府 路	山东省 青岛市 胶南市 旺富路
N-2	天津市东丽区川铁路	天 津 市 东 丽 区 川 街	天津市 东丽区 川铁路
N-3	吉林省白城市洮北区洮河大路	吉 林 省 白 城 市 窑 北 区 蛟 河 大 路	吉林省 白城市 洮北区 辽河大路
N-4	上海市闵行区墨江路	上 海 区 民 航 路 墨 江 路	上海市 闵行路 墨江路
N-5	四川省成都市郫县禹庙下街	四 川 省 成 都 市 郫 县 庙 下 街	四川省 成都市 郫县 禹庙下街
N-6	上海市闸北区公兴路	上 海 市 夏 北 区 共 兴 路	上海市 闸北区 公兴路

Both character and word bigram model eliminates the out-of-context mis-recognition. However, when model with character as feature, the recognizer still tends to produce errors with characters sound very similar to each other, and performs poorly compare to word bigram model. For instance, ID N1 of table x.x, the street name “王府路” sounds similar to “旺富路”, a character bi-gram model have seen both the term “王府”, and “府路” as well as “旺富” and “富路”. Using Baidu’s search result as a measuring baseline, the term “王府” occurred 77 million times in the search result, and the term “府路” appeared 8.3 million times. “旺富” appeared 330 thousand times, and “富路” appeared 85 million times. A character bigram model favors the more popular term of “王府路” over the less popular “旺富路”, which leads to a misrecognition. Similarly, “川街” can be found 16 million times from Baidu’s search result compared to the less popular “川铁” appearing 865 thousand times in the result. Using character as a feature lacks of the contextual information of the previous postal address term. A word bigram model overcome the errors produced by the character bigram. For example, knowing the city name being “胶南市” can help to increase the probability of street name “旺富路” over the street name of “王府路” because “王府路” never appeared in the city “胶南市.” Similar cases can be found for street name “民航路” and “闵行路”, and the district and street combination of “夏北区共兴路” and “闸北区公兴路”

5.3 Geo-Topology Realignment

A word bigram model significantly improve the recognition accuracy. However, due to the lack of geographical knowledge, ngram model still suffers from some of the mis-recognition errors listed below. We have identified the following five major types of mis-recognition errors from our bigram model.

Table 3. Recognition errors produced by word bigram model

ID	Ortho String	Recognized String
C-1	浙江省衡州市江山市丰新线	浙江省杭州市江山市丰新线
C-2	上海市闵行区天星路	上海市闵行区田兴路
C-3	重庆市渝中区罗汉寺街	重庆市渝中区罗汉寺
C-4	四川省成都市锦江区洗瓦堰路	四川省成都市锦江区洗瓦**路
C-5	香港特别行政区屯门区青麟路	香港潮南区青岭路

C1 - Invalid address combination

Valid address related words are correctly recognized, but does not form a single valid address. For example, both the city “衡州市” and “杭州市” belongs to the province “浙江省,” and they are phonetically similar to each other. However, the sub-district of “江山市” only belongs to the city

“衡州市.” Therefore, the city and sub-district combination of ” 杭州市江山市” does not form a valid address.

C2 - Similar sounding words
Although a word-based bigram significantly eliminates some of the confusion between similar sounding words, but it still produces mis-recognition such as the similar sounding street names of “天星路” and “田兴路.” Due to the popularity of the street name, a street “田兴路” might appear more often in the training set than “天星路.” Such empirical approach ultimately leads to mis-recognition.

C3 - Missing words
The end user interface built for this system has a push-to-talk button to allow a user to input by voice. Sometimes the user talks before pushing the button or release the button before finishing the voice input. This would cause the recognizer to miss words at the beginning or end of the sentence.

C4 - Encoding issue
The Nuance Recognizer integrated into our VPI system has a software level issue that it sometimes produces Chinese character with encoding issue.

C5 - Other errors
We observe a very small amount of errors were introduced from human interference such as noises created from recorder adjusting the microphone during recording, and low recording volume due to inappropriate microphone position. We choose not to focus to solve this type of error in this paper.

To tackle the misrecognition errors from our ngram model, a geo-topology was generated offline based on 198853 valid addresses in China. 33 province level nodes, 1320 city level nodes, 30972 district level nodes and 103446 street level nodes were created with appropriate links between each level of nodes in the geo-topology. This geo-topology forms a set of rules to validate the recongition result, and to correct possible misrecognition errors based on the relation between each node in the geo-topology. Two correction methods below are designed for the GTR process.

- Skip Match. If two words, w_x,w_y in the recognized string can match to two nodes, n1 and n2 in the geo-topology, and if these two nodes are not immediately connected, meaning one is not the parent of the other, then the GTR process will find the union set of n1’ s child nodes and n2’ s parent nodes, or vice versa. If such union set has only one unique node, we then replace or insert the name of this node with the string between w_x, and w_y. Figure xx below shows the logic of Skip match. In this example, which is taken from C1 of Table 3., the city “衡州市” was misrecognized as “杭州市”. These two cities have almost identical sounds, and therefore easily be misrecognized. Although both cities belong to “浙江省” province, but the sub district “江山市” only belongs to “衡州市”, but not “杭州市.” GTR takes advantage of the knowledge from geo-topology to automatically correct such misrecognition.
- Pinyin Match. If the union of two nodes in geo-topology, n1 and n2, has more than one node, we cannot directly apply Skip Therefore, a Pinyin

Fig. 1. GTR Skip Match Logic

Match is proposed. The union nodes of n1 and n2 will be converted into Pinyin, and then the node whose Pinyin has closest phonetic edit distance to the misrecognized string will replace the recognize string. The example below which comes from a real misrecognition string in our experiment is corrected by GTR. The street "旺富路" was misrecognized as the more popular street "王府路" The Pinyin representation of these two words are "Wang4Fu2Lu4" and "Wang2Fu3Lu4" respectively. The city node "胶南市" in the geo-topology contains more than one child whose value represents the street names belonging to this city. GTR performs a fuzzy match based on edit distance on the Pinyin of these street names and finds the street "旺富路" being the closest match. GTR then replace the correct term "旺富路" with the misrecognized "王府路" to "旺富路". The figure blow is the graphical representation of the Pinyin match logic of GTR.

5.4 Overall Recognition Accuracy

We consolidate our experiment results and provide a summary on recognition accuracy using different techniques below.

When using iFlyTek Voice Cloud, experiment iFlyTek, our system performs poorly with a word level accuracy of 87.79%. In experiment iFlyTek+GTR, GTR is applied to the results from iFlyTek Voice Cloud, there is about 2% absolute improvement is observed. When GTR is applied to the results produced by Nuance Recognizer with a customized word bigram (Experiment Word.Bigram+GTR), there is approximately 4% absolute improvement compare to the original bigram results reaching 96.54%.

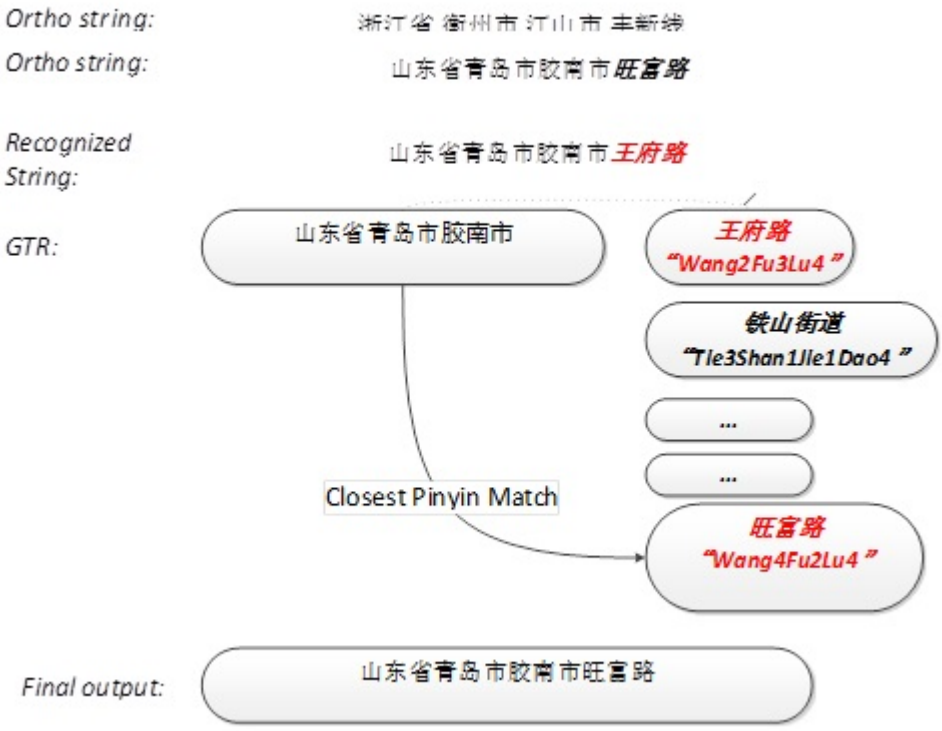

Fig. 2. GTR Fuzzy Pinyin Match Logic

Table 4. Recognition errors produced by iFlyTek Voice Cloud

Experiment Name	S_{Acc}	W_{Acc}	P_{Acc}	C_{Acc}	D_{Acc}	St_{Acc}
iFlyTek	41.86%	87.79%	98.16%	97.80%	89.39%	74.50%
iFlyTek+GTR	66.01%	89.68%	96.01%	98.42%	91.16%	83.94%
Character.Bigram	28.28%	78.44%	94.70%	88.96%	76.31%	72.50%
Word.Bigram	68.45%	92.42%	98.96%	96.82%	92.23%	89.51%
Word.Bigram+GTR	83.27%	96.54%	99.74%	99.23%	96.59%	92.77%
iFlyTek+Word.Bigram + GTR	85.07%	96.93%	99.66%	99.53%	97.12%	93.57%

In iFlyTek+word.bigram+GTR experiment, The best accuracy performance is observed when GTR is applied to the results produced by both iFlyTek and word bigram model achieving 96.93% word accuracy rate, and with the sentence accuracy is at 85.07%. This means for every 100 addresses inputted using VPI system, only about 15 addresses will contain errors. There is less than 1% that the misrecognition error will occur at province or city level, and about 2% chance that the error will occur in district level. Street level accuracy is slightly lowered to 93.57%.

5.5 Speed Performance

As described earlier in this section, our VPI system consists of a web client to collect audio data from end user, and then the voice data will feed into two recognizers, the result of the recognition will be passed to a post process of GTR. Separate measurements have been taken for each of the individual sub-task, including:

t_1 time it takes to read an address
t_2 transmission time for voice data from client to server
t_3 processing time required by iFlyTek Voice Cloud
t_4 processing time required by Nuance Recognizer
t_5 processing time required by GTR
t_6 transmission time of results from server to client

The total time to input an address through VPI system is then,

$$T = t_1 + t_2 + t_3 + t_4 + t_5 + t_6 \tag{1}$$

A complete breakdown of VPI system processing time is shown in table below.

Table 5. Recognition errors produced by iFlyTek Voice Cloud

Process ID	Process Name	Avg. Time (ms)
t_1	Voice Recording	2676
t_2	Transmission from Client to Server	5
t_3	iFlyTek Recognition	1592
t_4	Nuance Recogntion	3543
t_5	GTR	569
t_6	Transmission from Server to Client	5
t_7	Total Process Time	8380

The average input speed using VPI system is about 762 ms per character derives from the average process time of 8390 ms for an 11 character Chinese postal address. VPI system performance can further be enhanced. Current version of runs the speech recognition in serial process, meaning it will run iFlyTek

first and wait for the recognition complete before running Nuance Recognizer. This workflow can be further improved by running both iFlyTek and Nuance recognition in parallel with a multithreaded module. It is estimated that 1951 ms of processing time can be saved. In addition, audio streaming allows further time saving. Instead of spending 2682 ms to record an input, audio streaming allows VPI system to start capturing voice data and start the recognition process right away. With a buffer size of 1 second, audio streaming will send the recorded data to server every second. This will also significantly cut down the processing time of VPI system. By entering postal address through voice using VPI system, the data entry person is estimated to be 33.15% - 63% faster than by using traditional keyboard with Pinyin input based on the estimation of 1129 ms per character typing speed for an average user [4].

6 Conclusion

We focuses on implementing an ASR-based input system for Chinese postal addresses. We aim to examine the feasibility and usability of such system for the commercial settings. With a proper language model trained specifically with address specific data, and integrating geo-topology knowledge, the VPI system is able to achieve very high recognition accuracy. On the other hand, using VPI system, the input rate per character is estimated to be 762 ms per character, which is about 33% faster compare to keyboard based Pinyin input method. To our knowledge, our ASR-based input method is the first attempt ever to applied speech recognition accuracy in the domain of postal address recognition. The implication of this work is that using speech as an input for Chinese postal address is indeed feasible. Shipping and postal companies in China can use VPI system to improve their data entry process to reduce operation costs. Ultimately, people in China can also be benefitted with a cheaper shipping service due to such reduction.

7 Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61170196, 61133012 and National Program on Key Basic Research Project (973 Program) under Grant 2014CB340501.

References

1. Chen, Zheng, and Kai-Fu Lee. "A new statistical approach to Chinese Pinyin input." In Proceedings of the 38th annual meeting on association for computational linguistics, pp. 241-247. Association for Computational Linguistics, 2000.
2. Hartley, James, Eric Sotto, and James Pennebaker. "Speaking versus typing: a casestudy of the effects of using voicerecognition software on academic correspondence." British Journal of Educational Technology 34, no. 1 (2003): 5-16.

3. Chen, Zheng, Jian Han, and Kai-Fu Lee. "Language input architecture for converting one text form to another text form with tolerance to spelling, typographical, and conversion errors." U.S. Patent 6,848,080, issued January 25, 2005.
4. Wang, Jingtao, Shumin Zhai, and Hui Su. "Chinese input with keyboard and eye-tracking: an anatomical study." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 349-356. ACM, 2001.
5. "iFlyTek Voice Cloud," iFlyTek, 2 May 2014. [Online]. Available: http://open.voicecloud.cn/. [Accessed 2 May 2014].
6. Erden, Mustafa, and Levent M. Arslan. "Automatic Detection of Anger in Human-Human Call Center Dialogs." In INTERSPEECH, pp. 81-84. 2011.