LINGUISTIC AND ACOUSTIC ANALYSIS OF CHINESE PERSON NAMES

Wen-Jie CAO, Bo XU, Juha Iso-Sipilä*
National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing
{caowj, xubo}@nlpr.ia.ac.cn
*Nokia China R&D Center, Beijing
juha.iso-sipila@nokia.com

ABSTRACT
In this paper, we give results on our recent study on Chinese person names. The analysis is based on a corpus of 1 million names. The results include the syllable lengths and surname composition of the names in the corpus, full name and given name statistic results and analysis, tonal pattern analysis of Chinese full and given names, and name confusion analysis when given number of names are extracted from the corpus. Since pronunciation is involved in the tonal pattern of the names and confusion analysis, we give our strategy of dealing with the multi-pronunciation characters in Chinese person names. A model is brought forward to estimate the upper-limit to the recognition accuracy when N words are randomly extracted from the corpus, and has been used in our name analysis.

1. INTRODUCTION
Many fields, for example, text parsing field, speech recognition field, involve proper name processes. Chinese person name identification/recognition is one important application of these processes. To gain more challenging performances of systems, researchers need to utilize more information and characteristics of Chinese person names. So, a comprehensive analysis is needed, which may properly reflect current state.

A Chinese person name is formed by surname and given name. Most surnames are monosyllabic, and few are disyllabic, which represent one surname, or combination of two monosyllabic surnames. For surnames, since many surnames were rooted from the place people lived, or the official positions of people’s ancestors, the distribution of surnames is different in different places. For given names, almost all given names are monosyllabic or disyllabic. People’s selection of names (name length and characters used in given names) is influenced by many factors, including gender, dialect, time, psychology, geology, belief, etc. [3]. Compared to those of most western languages, Chinese name is more flexible, and it is impossible to include all the different Chinese names in a relatively small vocabulary. Some researchers have found rules in surnames and character selections in given names. Yet monographs on the analysis of Chinese names, especially those based on large and representative name corpus, are rare and precious.

We use a 1 million Chinese person name corpus (1,000,400 entities, size: 7.45MB) to do the statistic work. The entities were mainly extracted from many newspapers, and some were inputted manually. This corpus has an attribute of stochastic, and is statistically significant. We analyzed about 992,500 name entities of the corpus. In the other 8,000 names, 6,000 entities are transliterated names; other 2,000 entities are considered as “potentially” Chinese names, for we couldn’t find from the information we could access that the surnames used in these entities are really surnames. Considering that there might be input errors, we left this part untreated. Since the frequency of each surname in these 2,000 names is quite low, this may not affect the significance of the statistic result. In the rest of the paper, “the corpus” refers to the analyzed 992,500 names.

In this paper, section 2 gives some general information and the composition of the corpus and the analysis of surnames and names. A comparison with previous statistic data is also given in the paper. A comparison is also made with previous data. Section 3 provides our process of multi-pronunciation characters and tonal pattern analysis. Section 4 deals with the confusion analysis of the names. In the end, we give our conclusions and future plan.

2. CHARACTER STATISTICS AND ANALYSIS OF THE CHINESE PERSON NAME CORPUS
4,253 different characters are involved in the whole corpus (1,000,400). 4,064 characters are used in the 992,500 Chinese names. In these 4,064 characters, 956 characters are involved in surnames, and 3,994 characters are found in the given names of these entities. Characters used in given names can be commendatory or neutral. They can also be derogatory, which especially in villages is believed to help children with such names to grow up more easily. ISCSLP will cover, but not limited to, the following topic areas.

In the name corpus, most names are tri-syllabic, whose number is about 891,750 (89.850%). The second is disyllabic, and its number is about 100,400 (10.116%). The least frequent names are tetrasyllabic (the surnames are two characters), which contain 340 names with the proportion of 0.034%. The ratio of disyllabic names to tri-syllabic names is 1: 8.882. Traditionally, people tended to select disyllabic given names, and most full names are trisyllabic. For a time (1970’s-1990’s), there was a fashion to select monosyllabic given names for babies in many places. This trend originated from some cities, and then swept some of the countryside. But people found that monosyllabic given names are more prone to be same with that of other people’s and can express less information than disyllabic given names do. From the late 1980’s, the trend gradually faded and 2-syllable given names become more common.
2.1 Analysis of Surnames

In the extraction of surnames, we use a 570-surname set [2] to do a first-step extraction. Then by using the statistic information of the corpus, we found out other possible surnames. For each surname that we were not sure, we use the abundant information of the Internet and modern Chinese dictionary to confirm it and find out its correct pronunciation.

From the corpus, we extracted 974 surnames, in which 951 are monosyllabic and 23 are disyllabic. These surnames can cover most Chinese names (99.8%, estimated from the corpus, if the 2,000 “potentially” Chinese names are calculated).

2.2 Analysis of Names

In this part, we analyzed given names and full names.

2.2.1 Given names

In the given name analysis, we analyzed monosyllabic and disyllabic given names. About 100,450 given names are monosyllabic. Character set used in monosyllabic given names contains 2,341 characters. About 891,920 given names are disyllabic. 3,075 characters are used in the 1st syllables of disyllabic given names; and 3,406 characters are used in the 2nd syllables of disyllabic given names. Totally, 3,994 characters are used in disyllabic given names.

From the corpus, we extracted 974 surnames, in which 951 are monosyllabic and 23 are disyllabic. These surnames can cover most Chinese names (99.8%, estimated from the corpus, if the 2,000 “potentially” Chinese names are calculated).

2.2.2 Full Names

We analyzed the full names. If characters used in two names are the same, we say they are repeated. The number of unique names is the number of different names in the name corpus. The following is a contrastive table of our statistic data with that of the sample and that of a statistic of 1.6 million names in Motorola China Research Center. Superscript 1 means data of Motorola China (similarly in Table 5).

Table 3 accumulative probabilities of characters of given names

Table 4. A comparison of full names

From Table 4, we can see that in most cases, the proportions of repeated names are increased with the increase of the number of person names. Yet case is different between our data and that of Motorola China when given names are disyllabic.

We have also found from the statistics, that if we disregard gender factor, which is among the factors that influence the selection of names, the most frequent names of ours are in line with that of the sample.

3. PRONUNCIATION AND TONAL PATTERN ANALYSIS OF THE NAMES

Since the analysis of the pronunciations and tones of the names is related to the pronunciations and tones of characters, we firstly took some measures to deal with the multi-pronunciation characters. For a character that has one pronunciation and multi-tones, we consider it to be a multi-pronunciation character.

3.1 Pre-processing: Treatment of Multi-pronunciation Characters

There are over 1,000 multi-pronunciation characters embodied in the newest "Modern Chinese Dictionary". 396 multi-pronunciation characters appeared in the corpus, where 387 characters appeared in given names, and 138 characters in...
surnames. Compared with that of text, treatment of multi-
pronunciation characters in names has its own characteristics:
- When used as surnames, most multi-pronunciation characters have only one pronunciation
- Many multi-pronunciation characters have only one pronunciation in names, e.g.: 华 hua2, 荣 hong2
- Some Chinese names have repeated characters, such as 吴 wu2 吴 wu2. The tone of the 2nd character is a neutral tone
- Some characters have specific pronunciations in specific position in a name, for example: "wei2" in the last syllable of names is often pronounced as "wei2".

Regarding these characteristics, we deal with the multi-
pronunciation characters based on lexicon and rules. We made a
lexicon of properly notated multi-pronunciation characters, and we use rules for repeated characters in given names and specific pronunciations in specific positions. About 315,600 names in the
corpus have multi-pronunciation characters. After the processing,
we extracted 305,900 names. Still we could hardly find the most
proper pronunciations of the remained 9,700 names. We didn’t
use this portion in our analysis.

### 3.2 Pronunciations of Full Names

We compared the number of different pronunciations between the
names of ours and that of Motorola China. Then we get Table 5.

<table>
<thead>
<tr>
<th>Name length (syllables)</th>
<th>Total number</th>
<th>AP1</th>
<th>AP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>99,888</td>
<td>4.36</td>
<td>6.06</td>
</tr>
<tr>
<td>2</td>
<td>137,137</td>
<td>6.42</td>
<td>9.12</td>
</tr>
<tr>
<td>3</td>
<td>882,420</td>
<td>1.86</td>
<td>2.21</td>
</tr>
<tr>
<td>3</td>
<td>1,497,923</td>
<td>1.51</td>
<td>2.17</td>
</tr>
</tbody>
</table>

Table 5. A comparison of pronunciations of full names

In Table 5, the second column is the total number of people. Column AP1 is the average number of people that share the same pronunciation with tones. AP2 is the average number of people that share the same pronunciation without tone. We can see that
tone information decreases the repeated rates for a large portion.

### 3.3 Tonal Pattern of the Names

Tone information is important for tonal language speech recognition. Even for whispered speech, 64.0% tone recognition
can be achieved [7]. In this paper, we analyzed the tonal pattern of
Chinese person names.

For a tri-syllabic Chinese person name, if the pronunciation of a
name is “cao2 wen2 jie2”, its tonal pattern is “2 2 2”. We analyzed the
tonal pattern of full names and given names respectively. Because the number of 4-syllable names is too small, we didn’t calculate names with 4 syllables.

### 4. NAME CONFUSION ANALYSIS

In this part, we bring forward a statistical model to estimate the
confusion probability between each two names of the randomly generated name list, and check the name confusion status of the corpus when randomly extracting N names from the corpus, where N equals to 20, 50, 100, 200, and 500. Then, based on these estimated confusion probabilities, we estimated the upper-limit to the recognition accuracy of name lists of size N.

#### 4.1 Estimation of Recognition Rate

In the estimation, firstly, we use a large test corpus, which
records the recognition results of a LVCSR in syllables, to train
syllable confusion matrix. Then, we use the generated confusion matrix to estimate the confusability between two names.

Although Chinese has only 409 syllables, we didn’t use the original test corpus directly. This is because the frequencies of the syllables in the corpus are quite different, and the lowest syllable frequency may be too low to estimate the confusion probability between this and other syllables accurately (sparse data problem). To resolve this problem, we estimated two phoneme confusion probability matrices, confusion matrix of 24 consonants, and confusion matrix of 37 simple or compound vowels, from the corpus. Supposing that a consonant cannot be confused with simple or compound vowels, and vice versa, we re-estimated syllable confusion matrix with these two phoneme confusion matrices. Supposing two syllables s1 and s2, where s1=ci1vi1, s2=ci2vi2 (ci and vi are consonant and vowels of si, i=1,2), the probability of confusing s1 with s2 is:

\[
P_{s1s2} = P_{ci1ci2} P_{vi1vi2}
\]

(EQ 1)

On computing confusion probability between two words, we assume a model that somewhat like a dynamically built finite state network. Each syllable of the shorter transcription (a
sequence of syllables of a word) acts as the state of the network, and syllables of the longer transcription acts as the arcs of the network. Probability of each arc equals to the confusion probability between the syllable of the arc and the syllable of the state at which the arc points. We use dynamic programming (DP) matching. The boundary condition is that the first and the last syllables of the two words should be matched respectively.

As figure 3 illustrates, transcriptions of two names are $T_1=S_1S_2$ and $T_2=S_1' S_2' S_3'$, where $S_1=\text{"zhang"}$, $S_2=\text{"li"}$, $S_1'=\text{"chang"}$, $S_2'=\text{\"lim\"}$, $S_3'=\text{\"qiu\"}$. We want to find the probability of confusing $T_2$ with $T_1$. Let syllables of $T_1$, that is \"zhang\", \"F\", be the states of the network, and the sequence $T_2$, that is \"chang lin qiu\", be the output sequence.

![Figure 3](image)

Then the state shift path could be the sequence labeled by the solid lines, or the sequence labeled by the dashed lines. Probability of each path is the product of the probabilities of the arcs in the path. The biggest probability will be the confusion probability between $T_2$ and $T_1$.

For a randomly generated name list of size $N$, we can estimate the recognition accuracy (ERA) could be expressed as:

$$ ERA_{ij} = P_i / \sum_{j=1}^{N} P_j $$

(EQ 2)

### 4.2 Result

The extraction and computation was repeated 2,000 times for each $N$. Besides, we have checked in each extraction the number of name pairs, triplets, and quadruplets, and so on, in which the names have the same pronunciation (tone omitted). Limited by the space, we only give the average value. Here are the results.

<table>
<thead>
<tr>
<th>Extraction Number</th>
<th>Average ERA (%)</th>
<th>Average number of repeated transcriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.999866</td>
<td>0.003</td>
</tr>
<tr>
<td>50</td>
<td>0.998458</td>
<td>0.0425</td>
</tr>
<tr>
<td>100</td>
<td>0.998972</td>
<td>0.1435</td>
</tr>
<tr>
<td>200</td>
<td>0.996899</td>
<td>0.6065</td>
</tr>
<tr>
<td>500</td>
<td>0.991430</td>
<td>3.819</td>
</tr>
</tbody>
</table>

Table 6. Average ERCR and average number of repeated transcriptions for randomly generated name list sized N

From Table 6, we can see the average ERA hasn’t decreased greatly with the increase of the extracted names. But repeated names in the extracted list increase greatly. Table 7 shows that in a name list, the recognition accuracy of a name is mainly influenced by the number of names with repeated transcription. That means, for a person name speech recognizer, with the increase of the lexicon, the whole performance of the recognizer may not decrease greatly, but for some names in the lexicon, the probabilities of being recognized as other names may be twined, tripled, or more. In practice, the risk may be quite more than the theoretical estimation.

### 5. CONCLUSION AND FUTURE WORK

A general report for our analysis of a 1 million person name corpus, as well as its comparison with previous research, is presented in this paper. Through our statistics and analysis of character usage, pronunciation and tonal pattern condition of the Chinese person names, we hope the results will be helpful to corresponding researches, for example, voice tag voice dialing Chinese person names, we hope the results will be helpful to corresponding researches, for example, voice tag voice dialing systems. We also give a measure, which uses the recognition results of a LVCSR, to estimate the confusability of two names, and list the results in the paper. For the future, we may modify the word confusion analysis model to make the “upper-limited” recognition accuracy closer to practical situation. And we will perfect the model and put it into more use.

### 6. ACKNOWLEDGEMENT

The research work described in this paper is supported by the National Nature Science Foundation of China under grant number 9835003 and 60175012. It is also supported by the National Key Basic Research Program of China under grant number G1998030504.

### 7. REFERENCES


