ABSTRACT

This paper presents our recent efforts in developing a speaker independent LVCSR engine for Mandarin Chinese using our multilingual database GlobalPhone. We describe a two pass approach, in which the recognition first generates phoneme hypotheses and second transform these into Chinese character hypotheses. We show how this approach can reduce complexity and increase flexibility. We evaluate and compare different systems including different base units for speech recognition as phoneme units versus syllables. Furthermore we analyze the influence of tonal information. Our currently best system shows very promising results achieving 15.0% character error rate.

1 INTRODUCTION

With the distribution of speech technology products all over the world, the fast and efficient portability to new target languages became a practical concern. Our Janus Recognition Toolkit (JRTk) is language independent and we have already shown that the underlying recognition methods and techniques can be applied to several languages [1]. To setup a recognizer in a new language the acoustic models, the pronunciation dictionary and the language model have to be trained or adapted. In this paper we describe our work in bootstrapping a Chinese LVCSR system from a multilingual recognizer engine. Since the input of Chinese characters to the computer is a very time consuming process LVCSR systems for Chinese languages are of very special interest. The benefits of Pinyin based speech recognition are the following:

- Smaller search space: The mapping between Pinyin syllables and Chinese characters is very ambiguous and using Pinyin decrease the dictionary
- No Out-of-Vocabulary (OOV) rate: There are only about 1300 different syllables including the tonal information and no segmentation into words
- Easy adaptation: No adaptation from existing tools to the Chinese character set is necessary, we can use the accustomed segmented word model like in other languages
- Modularity: Separate examination of Pinyin to Character conversion is applicable for text-to-speech
- Error tracking: Pinyin transcription is closer related to audio than Chinese character transcription and therefore makes error tracking easier.

2 WORD SEGMENTATION AND PINYIN CONVERSION

In Mandarin Chinese every character is spoken in a mono syllabic manner. The over 10,000 different characters can be expressed by about 1344 Pinyin syllables which consist of a combination of 408 base syllables and 5 tones. Some of the 2040 possible combinations do not appear in common Chinese speech
and we can limit the set of Pinyin syllables to 1344. By putting all 1344 Pinyin syllables in the vocabulary every spoken utterance can be expressed in terms of segmented Pinyin. After the reverse character conversion the Segmentation does not exist anymore ..., so that the Out-of-Vocabulary (OOV) problem is eliminated in the recognition process. This results in a compact and efficient recognition engine.

2.1 The Pinyin conversion

We have used two separate steps to achieve this conversion. In the first step we segment the Chinese words and in the second step we translate the segmented Chinese words into the Pinyin representation.

![Figure 1: The Pinyin converter](image)

The final Pinyin converter can handle large numbers, year dates, percentage in the right manner and tries to give translations for English abbreviations into likewise pronounced Chinese characters.

2.2 The Chinese characters conversion

After the recognition process we have to convert the Pinyin hypothesis back to Chinese characters to make the output readable and performance results comparable. This is the inverse conversion of the process described in Figure 1. However, this process is much more ambiguous and very large context information in necessary. Nevertheless, in this task data sparseness is no longer a problem because we can label a big corpus of six years of Peoples Daily newspaper with the Pinyin converter. We have invented a method to automatically learn a minimized set of translation rules. During learning, the tool reports the performance rate, the size of the rule set, and translation errors with their full context. This information helps to control the rule learning process.

![Figure 2: The Chinese characters conversion](image)

With about half a million rules we achieve an error rate of 3.2 %. As an approximate over-all error rate we can add this error rate to the recognition error rate. In the worst case recognition errors can influence the context for the conversion and it is possible that we can get an
larger error rate than the sum of the two single error rates. But often the recognition and the conversion reports the same error so that one error is counted twice. The error difference between the Pinyin hypothesis and the Chinese character output of the best recognizer is less than 2.6%.

3 THE GLOBALPHONE DATABASE
All experiments have been carried out in the framework of the GlobalPhone project. The aim of this project is the development of a multilingual recognition engine. For this purpose a large speech database has been collected which currently consists of 13 languages, namely Arabic, Mandarin and Wu Chinese, Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. For each language about 100 native speakers were asked to read 20 minutes of newspaper articles. Their speech was recorded in office quality, with a close-speaking microphone. The GlobalPhone corpus is fully transcribed including spontaneous effects like false starts and hesitations. Further details of the GlobalPhone project are given in [1]. The Mandarin of the multilingual database was collected in three places in north, middle and south mainland China. The different places ensure a wide spreaded coloring of Mandarin dialect. We tried to get an uniform distribution of ages (between 18 and 65) and education levels. Our database consists of 10214 utterances with a totally length of 28.6 hours of speech spoken by 132 speakers of both gender, 112 speakers are used for training, 10 speakers each for the development and evaluation test set.

To train the language model we used 82.5 Mio words from Peoples Daily and Xinhua newspaper. The trigram perplexity of the language model is 207. The dictionary has a size of 17000 words including each syllable, which results in an OOV-rate of 0%.

4 THE RECOGNITION ENGINE
We liked to integrate the resulting Chinese recognition system into a multilingual speech recognizer framework. To do this in an easy way we decided to use similar preprocessing and acoustic modeling for all languages. During the preprocessing the dimensionality of the feature set is reduced to the first 24 LDA parameters of the 13 mel cepstral coefficients, power, zero-crossing and their first and second derivatives calculated from 16 kHz sampled input speech.

The system is a fully continuous 3-state HMM with emission probability modeled by a mixture of 16 Gaussians with diagonal variances.

4.1 Bootstrapping
For bootstrapping we generated a mapping from a multilingual system including English, German, Spanish, and Japanese phonemes [11] and wrote labels for the training data. In the next step we initialized the Gaussian codebooks with k-means and trained along the previously created labels and again wrote labels with the resulting system. We repeated this procedure several times until this context independent system reached its performance maximum. For a context independent system the polyphonic tree of all occurring quintphones (containing cross-word models with up to one phoneme lookahead to adjacent words) has been clustered down to 3000 codebooks by using linguistic motivated questions about the phonetic context.

4.2 Syllable vs. phoneme units for speech recognition
We developed a tool to generate parametric controlled phoneme sets and corresponding dictionaries. Using this tool we compared two promising mappings from phonetic information to acoustic models. In the first case we have mapped for every Pinyin syllable the beginning consonant, the middle vocal construct and the ending consonant on its own acoustic model. For the middle vocal construct we distinguish between five different tonal information.

<table>
<thead>
<tr>
<th># Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginnings</td>
</tr>
<tr>
<td>Middle vocals with tone</td>
</tr>
<tr>
<td>Middle diphthongs with tone</td>
</tr>
<tr>
<td>Middle triphthongs with tone</td>
</tr>
<tr>
<td>Endings</td>
</tr>
<tr>
<td>Σ</td>
</tr>
</tbody>
</table>

Table 2: Composition of the first phoneme set

The inter syllable coarticulation in Chinese is only reflected in minor degree compared with Indo-European languages like English. Associated with the fact that there are only about 1300 frequently used syllables including the tonal information, we decided to use an acoustic model for the whole syllable in the second case. The first context dependent system outperforms the corresponding phoneme based system by 29.1% to 30.8% error rate on the word based Pinyin hypothesis. While building a context independent system some severe run-time and memory consumption problems arose, caused by the huge amount of more than 1000 acoustic models. This forced us to break the further development of the syllable based context independent system until we have changed the JRTk. The expectation for a performance increase is not as large as for the phoneme based system, because the inter-syllable coarticulation is much smaller than the intra-syllable coarticulation for Mandarin Chinese.

4.3 Incorporating tonal information in the feature vector
Additionally, two different acoustic modellings of the tonal information were compared. Besides the implicit modeling of the tonal information through training of mostly 5 different phonemes for a vocal construct, we performed an approach by explicit detecting pitch information and adding 18 generated pitch characteristics to the feature vector before performing LDA ensuring that at least 6 pitch parameters are left in the resulting feature vector.

4.4 VTLN

In Vocal Tract Length Normalization (VTLN) a linear or nonlinear frequency transformation compensates for different vocal tract lengths [10]. Finding good estimates for the speaker specific warp parameters is a critical issue. For VTLN, we keep the dimension constant and warp the training samples of each speaker such that the Linear Discriminant is optimized. Although that criterion depends on all training samples of all speakers it can iteratively provide speaker specific warp factors. By training with speaker specific warp factors and estimating good warp factors for testing we can decrease the error about more than 1%.

4.5 Results

The table below shows the progress of the Mandarin system. PWE is word error rate based on the Pinyin hypothesis, while CWE is word error rate based on the Chinese character hypothesis. The word recognition error rates are reported for better comparison to other languages results. However, to compare our system to other Chinese systems, the commonly used character based error rates are presented in the last column (CCE) which is 15.0% for our currently best system.

<table>
<thead>
<tr>
<th>Systems</th>
<th>PWE</th>
<th>CWE</th>
<th>CCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>First bootstrapped version</td>
<td>50.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Data correction</td>
<td>43.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pinyin tool improvements</td>
<td>34.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Full training set and large LM: 112 train speaker + 82.5 LM</td>
<td>30.8%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Context dependent system</td>
<td>24.1%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Speaker normalization</td>
<td>22.9%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Including explicit pitch</td>
<td>21.8%</td>
<td>24.3%</td>
<td>16.1%</td>
</tr>
<tr>
<td>Currently best system</td>
<td>20.7%</td>
<td>23.3%</td>
<td>15.0%</td>
</tr>
<tr>
<td>Best syllable based system</td>
<td>27.7%</td>
<td>31.2%</td>
<td>21.3%</td>
</tr>
</tbody>
</table>

Table 3: System performance

5 CONCLUSION

Our experience shows that a nonnative developer can port a system to a new language within less than 6 months. Our currently best system shows 15.0% error rate on character output. Furthermore, it turns out that JRTk easy can be adapted to be used in several different languages. This is an important result concerning the integration into the framework of a multilingual recognizer. While building the Chinese recognizer we have learnt some methods to automate important steps building a new recognizer from the scratch and porting our Janus Recognition Toolkit (JRTk) to the windows platform.

6 ACKNOWLEDGMENTS

The authors wish to thank all members of the Interactive Systems Laboratories especially Tianshi Wei, Jing Wang and Jiaxing Weng from the GlobalPhone team for collecting and validating the Chinese database.

REFERENCES

[6] Gao, Yuqing; Hon, Hsiao-Wuren; Lin, Zhiwei; Loudon, Gareth; Yagunanthia, S.; Yuan, Baosheng: Tangerine: A Large Vocabulary Mandarin Dictation System, ICASSP 95, Detroit (Volume 1, Page 77)
[7] Lyu, Ren-Yuan; u.a.: Golden Mandarin (III) – A User-Adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary, ICASSP 95, Detroit (Volume 1, Page 57)
[8] Ma, Bin; Huang, Tairei; Xu, Bo; Zhang, Xijun; Qu, Fei: Context-Dependent Acoustic Models for Chinese Speech Recognition, ICASSP 96, Atlanta (Volume 1, Page 455)