MANDARIN LARGE VOCABULARY SPEECH RECOGNITION USING THE GLOBALPHONE DATABASE

J. Reichert, T. Schultz, and A. Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
{juergen,tanja,waibel}@ira.uka.de

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 Pinyin 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. In terms of portability three aspects distinguish the Chinese language from other languages:

- The Chinese ideographic characters do not reflect the pronunciation of a word
- Modern Chinese written text lacks the segmentation into words
- In spoken Mandarin the tonal information is necessary to distinguish meanings

Chinese ideographic characters do not allow to generate automatically the pronunciation for an unknown Chinese character. Chinese text are written in strings of Chinese characters without any delimiter between adjacent words. Thus Chinese languages lacks the natural segmentation into words which can be used as basis units for speech recognition purposes as known from Indo-European languages like English. Units can be found by segmenting text into single characters, by using prosodic information [3] or by defining semantic meaningful strings. Conventional Chinese systems analyze information about syllabic structure and tone separately and combine the information by synchronization in later steps [3]. Recent systems integrate the tonal structure [2], [5]. The number of phonemes used for acoustic modeling varies between 33 [3] and 100 [5] phonemes. Differences in the definition of phonemes exists since tonal information is assigned to different parts of phonemes in one syllable. Some researchers define initial and final parts of syllables [4] and handle the intra-syllable structure of a model. Most of the current systems use HMM-based acoustic modeling [2][6][7][8][9]. Whereas most studies operate on Chinese characters we decided to deal with the *Pinyin* transcription. Pinyin are word segmented phoneme transcriptions as used in the people's republic of China. The benefits of Pinyin based speech recognition are the following:

- Complexity: The mapping between Pinyin syllables and Chinese characters is very ambiguous. Using Pinyin results in smaller search space with decreased dictionary size.
- No Out-of-Vocabulary (OOV) rate: Only about 1300 different Pinyin syllables are required for full coverage.
- Compatibility: 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 recognition 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 Pinyin syllables which consist of a combination of 408 base syllables and 5 tones. Since some of the 2040 possible combinations do

not appear in common Chinese speech, 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 transform the segmented Chinese words into the Pinyin representation.

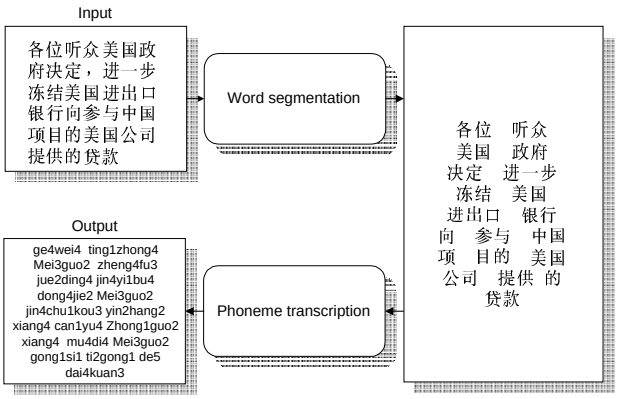

Figure 1: The Pinyin converter

In the first approach we only searched for the word with the maximal length in a lexicon to split a character string later on we added probabilities to support the splitting decision. For the phoneme transcription we start with a simple character mapping to the most likely Pinyin syllable. But the resulting error rate was unacceptable because both steps are ambiguous and context information is necessary to resolve the equivocations. In the segmentation process there exists no safe decision rule when to split two Chinese characters because nearly every part of a Chinese word with more than two characters has its own meaning. Furthermore the ending character can be combined with the beginning character of the following word resulting in correct words.
E.g.: 进出口 = Import/Export, 出口 = exit, 进 = to enter, 出 = out , 口 = mouth,
The phoneme transcription is not only a simple mapping because about 13 % of the Chinese characters has more than one pronunciation.
E.g.: 乐 = le4 (joy) or yue4 (music)
In many cases additional pragmatic knowledge is required to disambiguate different meanings. Due to the facts, that only few data of high quality are public available, we integrated different data sources like word lists, statistical information and manually Pinyin labeled Chinese text. We implemented for each step a 3 stage back-off process to exploit most optimal the existing data. The length of the resulting word units varies between 1 and 10 syllables with an average length of about 2 syllables per word, which is similar to word units in Indo-European languages. This length is a good trade off between a useful string length for acoustic disambiguation and language modeling context as well as a limited size of resulting vocabulary units. Since the Pinyin reflects the pronunciation of the spoken character we can use the Pinyin conversion tool for creating a pronunciation dictionary by simply adding some pronunciation rules for exceptional cases.
This converter gives only 1.5% Pinyin error rate compared to hand edited texts on Peoples Daily corpus. Word segmentation faults, mostly resulting from incorrect splitting proper names, are no serious problem for preparing training data and pronunciation dictionary.

System	Segmentation	Pinyin conversion
Brute-Force	--	~ 20 %
+Lexicon max length	~ 6 %	~ 20 %
+probabilities	8 %	6 %
Pinyin converter 1	4.9 %	2.1 %
Final converter	3.8 %	1.5 %

Table 1: Performance of word segmentation and Pinyin conversion

In order to get good transcription for training we add rules to handle large numbers, year dates, percentage to the final converter. We also introduce mapping rules for English abbreviations and acronyms into likewise pronounced Chinese characters.

2.2 The Chinese characters conversion

For speech-to-text purposes like dictation applications we have to convert the Pinyin hypothesis back to Chinese characters to make the output readable and performance results comparable. This is the reverse 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 (provided by LDC 1997) with the Pinyin converter. We have invented a method to automatically learn a minimized set of translation rules. To control and improve the rule learning process we supervise the tools feedback which consists of performance rate, the size of the rule set, and translation errors with full context.

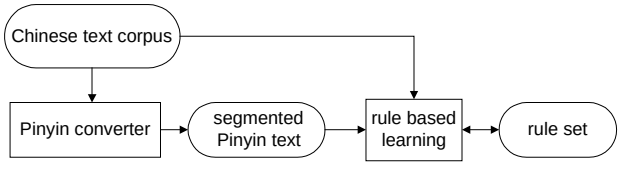

Figure 2: The Chinese characters conversion

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 part of the multilingual database was collected in three places in north, middle and south mainland china. The different places ensure a widespread 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

Our goal is 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, zerocrossing 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 dependent 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 vowel construct we distinguish between five different tonal information.

	# Phonemes
Beginnings	21
Middle vowel with tone	35
Middle diphthongs with tone	63
Middle triphthongs with tone	19
Endings	3
$\sum$	141

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 acoustic models based on 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 Tonal information

Additionally, two different acoustic modeling 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 results of other languages. 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.

Systems	PWE	CWE	CCE
First bootstrapped version	50.0 %	-	-
Data correction	43.0 %	-	-
Pinyin tool improvements	34.0 %	-	-
Full training set and large LM: 112 train speaker + 82.5 LM	30.8 %	-	-
Context dependent system	24.1 %	-	-
Speaker normalization	22.9 %	-	-
Including explicit pitch	21.8 %	24.3 %	16.1 %
Currently best system	20.7 %	23.3 %	15.0 %
Best syllable based system	27,7 %	31.2 %	21.3 %

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 implemented methods to automate important steps building a new recognizer from scratch. Furthermore we ported 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, and Kjell Schubert for implementing the pitch tracker.

REFERENCES

[1] *Schultz et al:* Language Independent and Language Adaptive LVCSR. ICSLP 1998.

[2] *Chen, C.J., Gopinath R.A., Monkowski M.D., Picheny M.A., and Shen K.:* New Methods in Continuous Mandarin Speech Recognition, Eurospeech 97 Rhodes 1543-1546

[3] *Lyu, R-Y. et al:* Golden Mandarin (III) - A User-adaptive Prosodic-Segment-Based Mandarin Dictation Machine for Chinese Language with Very Large Vocabulary. ICASSP 95, vol1, pp57.

[4] *Ma, B., Huang, T., Xu, B., Zhang, X., and Qu, F.:* Context-dependent Acoustic Models for Chinese Speech Recognition, ICASSP 96, Atlanta, pp 455-458.

[5] *Zhan, P., Wegmann, Steven, Lowe, Steve:* Dragon Systems 1997 Mandarin Broadcast news System, Darpa Workshop, Lansdowne, 1998.

[6] *Gao, Yuqing; Hon, Hsiao-Wuren; Lin, Zhiwei; Loudon, Gareth; Yaganantha, 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, Taiyi; Xu, Bo; Zhang, Xijun; Qu, Fei*: Context-Dependent Acoustic Models for Chinese Speech Recognition, ICASSP 96, Atlanta (Volume 1, Page 455)

[9] *Ho, Tai-Hsuan; Yang, Kae-Cherng; Huang, Kuo-Hsun; Lee, Lin-Shan* : Improved search strategy for large vocabulary continuous mandarin speech recognition, ICASSP 98, Seattle (Volume 2, Page 825, Paper number 2381)

[10] *Martin Westphal, Tanja Schultz, Alex Waibel:* Linear Discriminant - a new method for speaker normalisation, Proceedings of the ICSLP 98, Sydney, Australia, 1998.

[11] *Tanja Schultz and Alex Waibel:* Fast Bootstrapping of LVCSR Systems with multilingual Phoneme Sets, Proceedings of the Eurospeech'97 Vol. 1 pp 371-373, Rhodes, Greece, September 1997.