ADAPTING WESTERN LANGUAGE RECOGNIZER FOR CHINESE RECOGNITION

Hong XU, Frédéric BEAUGENDRE and Hugo VAN HAMME
Lernout & Hauspie Speech Products nv.
Sint-Krispijnstraat 7, 8900 Ieper
E-mail: {hong.xu, frederic.beaugendre, hugo.vanhamme}@lhs.be

ABSTRACT

In this paper, we will present a speaker-independent Mandarin Chinese recognition system. The system is designed to adapt an existing Western language recognizer in order to support Chinese recognition. We will only focus on the stage of acoustic recognition. Since Chinese has its special characteristic (a tonal language) comparing with Western languages, we will first develop a system special for tone recognition using pitch information. Then we will present two methods for the integration of tone information into the existing system.

1. INTRODUCTION

In Mandarin Chinese, there exist about 100,000 commonly used words. Each word is composed of one or several characters. There are more than 10,000 commonly used characters, and all of them are monosyllable. However, there are only about 1320 phonologically different syllables called *tonal syllables*, which are also known as pinyin. Since the introduction of computers to the Chinese community, one particular problem has been the difficulty to use a conventional keyboard as input for the more than 10,000 Chinese characters. Thus, many methods have been proposed for Chinese keyboard input, like for example Chajei, Pinyin, Five Strokes, etc.. However, none of them is as convenient as for alphabetic languages such as Western languages. Therefore, speech recognition of Chinese has become a very important challenge to overcome this problem.

A particular difference between Chinese and Western languages is the fact that a specific tone is associated to each syllable, and that tones convey lexical meaning. Although there are dialectal differences in the number of tone types, in the case of standard Mandarin Chinese, the number of tones is 5: 4 lexical tones and a fifth neutral tone. When the tones are disregarded, only 410 syllables can be phonetically distinguished. Those are called the *base syllables*. Furthermore, each base syllable consists of an INITIAL and a FINAL, similar to the consonant and vowel in English. There are in total 22 INITIALs and 40 FINALs.

In the last decades, many researchers have developed different methods for Chinese speech recognition. A survey about Chinese recognition can be found [3]. Conventionally, Chinese Recognition can be regarded as two-stage process. In the first stage, the acoustic recognition is performed to select syllable candidates. Tone recognition and base syllable recognition are performed separately. The tone recognition is based on tone models using mainly pitch information. The base syllable recognition uses models for the 408 base syllables (or INITIAL-FINAL units) based on vocal tract parameters. The results of both recognition parts are then combined to form the tonal syllable solution. A lexicon is then used to construct a word lattice and finally a language model is applied to find the final sentences in Chinese characters.

The aim of this paper is to present a system of automatic recognition of the Chinese language, based on an adaptation of the current speaker-independent recognizer for Western languages. In the next section, we first introduce our baseline system for Chinese recognition using the existing recognizer. Then, the tone recognition process is described. In section 3, we present two different methods for integrating tone and syllable recognition. Experiments on different tasks for speaker-independent recognition have been conducted and results will also be shown in the paper. Finally, a comparison with other Chinese recognition systems is given in section 4.

2. BASELINE SYSTEMS

In the baseline system for Chinese recognition, we use the existing recognizer for Western languages with special language developments for Chinese. In the first part of this section (1.1), tone information is disregarded since the treatment for tones does not exist in the Western language recognizer. The second part of the section describes the baseline system for tone recognition (1.2).

2.1. Syllable Recognizer Disregarding Tones

In literature, the INITIAL-FINAL structure is widely used as a standard speech unit for Chinese speech recognition. A number of experiments have been run in order to optimize the choice of units to use. The phoneme set which gives the best trade-off between recognition performance and size of the models use Initial Context Dependent and Final Context Independent units. For this structure, called “ICD”, there are 40 Context Independent (CI) FINALs and 103 Right Context Dependant (RCD) INITIALs. Those units are modeled using discrete density HMM, with 3 states for INITIALs and 4 states for FINALs (469 states altogether). The resulting models are called "ICD-TI" models since tone information is not considered.

It has been shown in [7] that feature vectors based on spectral parameters also have discriminating capability on phonemes with different tones. Therefore, in this paper, we also try to design FINALs as tone-dependent (named ICD-TD models). The ICD-TD models can already be regarded as models designed for tonal syllables. That is, each FINAL would have 5 versions with 5 different tones. For example, FINAL "a" will become "a1", "a2", "a3", "a4" and "a5". The numbers following the FINAL indicate the 5 different tones. If all FINALs with tones are considered, there will be 103*3 + 40*5*4 = 1109 states for such models. However, some FINALs with certain tones don't exist, thus 980 states remain only.

The baseline system is a multiple feature stream, discrete-density HMM recognizer. 80 speakers' speech data was used for clustering (in order to build the codebooks) and training of the models. In the training set, each speaker has recorded about 2/3 of the 410 base syllables, 200 isolated words, 100 sentences and about 100 digit-related utterances.

2.1.1. Experiments and Results

As our purpose is to design a speaker-independent recognizer, we use 10 speakers (5 females and 5 males, different from the speakers of the training set) for testing.

Four different experimental tasks have been considered:

- Isolated Syllables;
- Isolated Words;
- Connected Digits;
- Sentences

For the "Isolated Syllables" task , two different syntaxes were generated. The first one (s410) contains the 410 base syllables (each base syllable attached with one of its possible tones). A second one (s1300) contains 1330 tonal syllables. For isolated words, different vocabularies were also considered. One set includes about 110 command words (w100) for which about 16% of the words are not in the training set. A second one (w120) contains about 120 trained words. By combining these two sets, and adding the 1000 most frequent Chinese words, a larger lexicon (w1200) was also considered for our experiments. The "Connected Digits" (cdd) task contains isolated digits, connected digits of short length (4-5 digits), telephone numbers and credit card numbers. The "Sentences" task contains about 200 sentences for each speaker. A finite state graph is used as syntax for this task, no statistical language models being used at this stage. Models are trained in a general task-independent way. The average recognition results for the 10 speakers are shown in table 1 in word error rates percentage.

	s410	s1300	W100	w1200	cdd	Sent.
ICD-TI	36.30	NA	3.05	10.10	11.40	3.47
ICD-TD	31.83	54.43	1.94	6.44	8.97	3.25

Table 1: Word error rate (in %) for the baseline system disregarding tones

2.2. Tone Recognition

A particular difference between Chinese and Western languages is the fact that a specific tone is associated with each syllable, and that tones convey lexical meaning. This makes the recognition of tones a very special task for a Chinese recognizer.

A robust measurement of pitch and voiced/unvoiced decision was thus implemented and applied for tone recognition. Both Hidden Markov Models (HMMs) and neural networks are almost equally successful for this task [3]. In order to stay as close as possible to the standard architecture of the recognizer, we used Discrete Density HMM. Moreover, Discrete Density Models have shown to be more successful for this task than Continuous Density Models ([4], page 168).

The feature vector used for tone recognition consists of:

- Normalized log-pitch P
- Delta log-pitch δP
- Normalized log-energy E
- Delta energy δlog-E

A single codebook is used for tone recognition. A number of experiments, which will not be described here, have been carried out in order to optimize the number of states per tone (4), the vector of features selected (see above), and the size of the codebook.

Context dependent (CD) tones have to be considered in order to handle the co-articulation effects. Even if 175 contexts should be considered to cover all the possible tone combinations, it has been shown that we can reduce this list to 23 contexts only, if the characteristics of the tone behavior are carefully considered [6]. In particular, the tone level at the beginning and the end if each tone should be especially examined. The same list of 23 tones is used here, augmented with four "isolated tone" models (lexical tone surrounded by silence) and a special CD tone for the particular case described in 3.1.2. Therefore, the total number of CD tone models is 28.

The databases used for training of those tone models and testing of tone recognition rates are identical to those used for the *base syllable* experiments (see above)

Tone recognition results on the test sets described in 2.1.1 are given in table 2.

	s1300	W100	w1200	cdd	Sent.
CDTM	14.8	18.5	18.2	28.6	38.3

Table 2: Word error rate (in %) for tones recognition only

3. INTEGRATION OF TONE INFORMATION INTO THE BASELINE

In this section, we propose two techniques for taking the tone information into account. The first method uses the HMMs described in the previous section. In the second method, the tone features are added as a new stream already during the training phase.

While the spectral co-articulation between syllables is limited, the tone co-articulation is prominent. When INITIALs and FINALs take the role of a phoneme, these effects cannot be modeled using conventional triphones. Therefore, we opt for a 2-pass search, where the first pass ignores tone co-articulation. In the ICD context, we can then use an available N-best algorithm that takes no co-articulations between words or syllables. Backtracing can only happen at the word level to save memory. The second pass rescores the N-best list using forced alignment and can, at least in principle, model complex dependencies. The architecture is outlined in Figure 1.

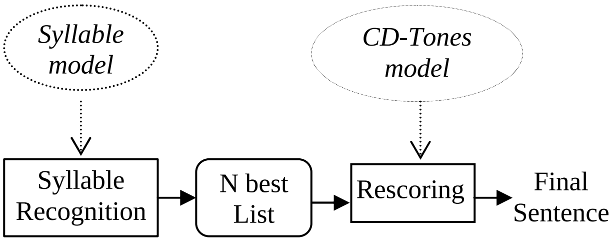

Figure 1. Two pass recognition for Chinese

However, since the tone models relate to voiced frames and FINALs only, the rescoring is not trivial. Rather than using the segmentation into INITIALs and FINALs as derived by the 1st pass (which would violate our constraint on the backtracing capabilities of the 1st pass search), we adopt other methods.

3.1. Tone Co-articulation Modeling

In the tone model, three digits are used to present the tone models with different contexts. For example, “122” states for “tone 2” preceded by tone 1 and followed by tone 2. For the training of CD tone model, we use a transcription of what has been actually pronounced (pinyin transcription) instead of a transcription at the Chinese character level. As a particular case in Chinese, the tone sequence “33” may be pronounced as “23”, “32” or “22” depending on the preceding and following contexts. For example:

我 买 书
wo3 mai3 shu1 => wo2 mai3 shu1

我 想 买 本 书
wo3 xiang3 mai3 ben3 shu1 =>
wo2 xiang3 mai2 ben3 shu1

我 很想 买 那 本 书
wo3 hen3 xiang3 mai3 na4 ben3 shu1 =>
wo3 hen2 xiang2 mai3 na4 ben3 shu1

Therefore, in the training set, “33” contexts are not any more encountered. But during recognition, a tone 3 followed by tone 3 may occur. Therefore, we made multiple transcriptions for the context dependent tones containing “33” for the second pass rescoring. For example, “133” is modeled as a multiple transcription of “123”, “132” and “122”, while “331” is modeled as “231”.

3.2. Integrating Separately Trained Tone Models

In the first method, the first pass ignores the tone feature stream. In the second pass, we augment the emission models with a weighed version of the separately trained CD-tone models of section 2. These have the advantage of being well-trained since they are pooled over all FINALs.

An extra VQ label is assigned to the unvoiced frames. INITIALs receive a uniform distribution on their tone codebook. By properly weighting the emission probabilities of the tone codebook, the 2nd pass can now be completed.

However, the experiments showed that the integration of tones and syllables is not trivial: different weights on tones and syllables would conduct different results. For example, in table 3, the result of the second pass (indicated by *) for ICD-TI model is obtained by using the integrated model where the weight on syllable and tone is 1:1, it is worse than that in the first pass. Therefore, more techniques are used to balance the weights between tones and syllables and between INITIALs and FINALs since tones are only with FINALs. The final results of this two-pass integration are shown in the following table (except the one with *).

		s410	s1300	w100	cdd
ICD-TI	Pass1	36.26		3.05	11.40
	Pass2	37.97*		2.59	9.91
ICD-TD	Pass1	31.83	54.43	1.94	8.97
	Pass2	29.73	49.10	1.66	6.48

Table 3: Word error rate (in %) for two-pass recognition

3.2. Direct Integration

Instead of training the models of the tones and syllables independently and integrating them later in the recognition process, we also propose to use Pitch (described in section 2.2) as a feature stream, similar as the other features. This method is also proposed in [1].

The emission models are trained with an extra codebook for pitch. The first pass useds context-dependent tone assumptios on the FINALs. This is fully in-line with the ICD-TD modeling context. With the addition of the pitch, we call this model ICD-PT. Refer to the row Pass 1 in table 4 for recognition performance.

		s410	s1300	w100	cdd
ICD-PT	Pass1	25.04	36.34	1.02	4.48
	Pass2	22.06	33.31	0.83	3.72

Table 4: Word eror rate (in %) for Direct Integration

Cross-word and cross-syllable co-articulation effects are handled by the second pass. The N-best hypothesis obtained from the 1st pass are re-scored using a forced alignment using tone-context-dependent final models. These are easily trained, since the tone context is known for each final in a training utterance.

In this integration method, we need not to tune the parameters between the tones and syllables. This is implicitly done in the training section by estimating the parameters for emission models. From table 4, we can see that considering the context dependent tones always improve the recognition.

Finally, by including gender modeling with automatic gender selection, we obtain the final performance listed in table 5 using the ICD-PT paradigm.

	s410	s1300	w100	w120	w1200	cdd	Sent.
Pass1	20.12	30.76	0.83	0.10	2.50	3.93	2.46
Pass2	18.87	28.95	0.65	0.10	NA	3.93	NA

Table 5: Word error rate (in %) for Chinese Recognition(final system)

4. COMPARISONS AND CONCLUSIONS

In this paper, we have presented some methods for adapting a western language recognizer to a Chinese recognition task. We have proposed two different methods to integrate the tone and syllable recognition. Substantial improvements can be achieved by integrating tone information, the best results being obtained using context dependent tones.

System	Basic Unit	HMM	Base Syl.	Tone	Tonal Syl.	Speaker
[4]	Syl.	Cont.	91.67	97.80		1 M.
[7]	Sub-syl.	Cont.	96.38	94.85	92.0	2 M.
[8]	Sub-syl.	Cont.	88.70	91.60	81.4	2M+2F
[2]	Syl.	Cont.	47.25	92.32	43.62	Indep.
[9]	Sub-syl.	Cont.			91.3	Indep.
L&H	Sub-syl.	Disc.	81.17	90.9	71.05	Indep.

Table 6: Comparisons of recognition (in accuracy %) of different systems

Table 6 compares the results of our final system (marked as L&H) and other existing systems [2,5,7,8,9]. M represents Males and F represents Females. All the systems use continuous HMM for modeling except ours. The first three systems [4,7,8] are speaker-dependent continuous speech recognizers. In [2], the system is designed for a speaker-independent large-vocabulary task. It uses an isolated-character input mode. In [9], the system is a speaker-independent dictation machine using isolated-word input and a vocabulary of 1000 words. The result shown in the table 6 is the average over 5 males. Our system gives an averaged 2.5% word error rate on 10 speakers for a 1000 word recognition task.

We can conclude that, by extending the techniques for an existing discrete-density HMM western language recognizer, we can achieve a system for the Chinese language that is comparable to continuous-density HMM recognizers.

5. ACKNOWLEDGEMENT

This work is partially supported by the IWT (Flemish Government of Belgium).

6. REFERENCES

[1] C-J Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny and K. Shen, *"New Methods in Continuous Mandarin Speech Recognition"*, EUROSPEECH, 1997.

[2] Tang-Hui Chiang and Yi-Chung Lin and Keh-Yih Su, "On Jointly Learning the Parameters in a character-synchronous Integrated spech and language model", *IEEE transactions on Speech and Audio Processing,* Vol 4, No. 3, pp. 167-189, 1996.

[3] Stephen W.K. Fu, C. H. Lee, Orville L. Clubb, "A Survey on Chinese Speech Recognition", *Communications of OLIPS* Vol. 6, pp. 1-17, 1996.

[4] L-S Lee, C-Y Tseng, H-Y Gu, F-H Lin, S-L Tu, S-H Hsieh and C-H Chen, "Golden Mandarin (I) - a real-time mandarin speech dictation machine for Chinese language with very large vocabulary", *IEEE transactions on*

Speech and Audio Processing, Vol. 1, pp. 158-176, 1993.

[5] Lin-Shan Lee, Chiu-Yu Tseng, Keh-Jiann Chen, I-Jung Hung, Ming-Yu Lee and Lee-Feng Chien, *"Golden Mandarin (II) - an improved single-chip real-time mandarin dictation machine for chinese language with very large vocabulary"*, Proc. ICASSP, Vol. 2, pp. 227--230, Mineapolis, 1993.

[6] Lin-Shan Lee, “Voice dictation of mandarin chinese”, *IEEE Signal Processing Magazine*, Vol. 14, pp. 63-101, July 1997.

[7] "Chih-Heng Lin and Chih-Heng Wu and Pei-Yih Ting and Hsin-Min Wang "Frameworks for recognition of Mandarin syllables with tones using sub-syllabic units", *Speech Communication*, Vol. 18, pp. 178-190, 1996.

[8] Hsin-Min Wang, Tai-Hsuan Ho, Rung-Chiung Yang, Jia-Lin Shen, Bo-Ren Bai, Jenn-Chau Hong, Wei-Peng Chen, Tong-Lo Yu and Lin-Shan Lee, "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data", *IEEE transactions on Speech and Audio Processing*, Vol. 5, No. 2, pp 195-200, 1997.

[9] T-H Chiang, C-M Pengwu, S-C Chien and C-H Chang,, *"CCLMD’96: Towards a Speaker-Independent Large-Vocabulary Mandarin Dictation System"*, ICASSP, Vol 1., pp. 1799-1802, Detroit, 1997.