ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

CHINESE-ENGLISH MIXED-LINGUAL KEYWORD SPOTTING

Shan-Ruei You, Shih-Chieh Chien, Chih-Hsing Hsu, Ke-Shiu Chen, Jia-Jang Tu, Jeng Shien Lin, and Sen-Chia Chang

Computer & Communications Research Laboratories
Industrial Technology Research Institute, Hsinchu
{shangray, saga, hsjus, koche, santu, jslin, chang}@itri.org.tw

ABSTRACT

Base on our former experience in the "ITRI 104 Auto Attendant System"[1] of using keyword spotting for Mandarin speech recognition, a Chinese-English mixed-lingual keyword spotting for catering the speaking style of Taiwanese is present. Detailed descriptions and discussions for developing the mixed-lingual auto-attendant system are included in this paper, especially for solving different scoring scales in the decoding phase and the re-scoring phase for these two languages. In the decoding phase, we propose a bias-compensation method to make up the score-gap in the likelihood calculation of using Chinese and English acoustic models. To select the most probable result from the recognized hypotheses, the method for normalizing the combination scores of using different scoring mechanisms in the re-scoring phase is also presented.

1. INTRODUCTION

In ITRI, we had built a keyword spotting based auto attendant system for Chinese-name query since July 2000, and it processes about 1500 calls everyday till now. About six thousand Chinese names are served in this system and the recognition rate is more than 92%. However, Chinese mixed with English is also commonly used in Taiwan, such as person names, car ID, personal ID, etc. A system with the capability of recognizing Chinese-English mixed-lingual speech is more suitable for the speaking style of Taiwanese. Base on our former system, several problems have to be solved for building the mixed-lingual keyword spotting system. The first one is the accent problem. If we use native English recognizer to recognize Taiwanese-accented English, the recognition result will be very poor. This problem can be easily solved through collecting the Taiwanese-accented English and developing a recognizer for Taiwanese-accented English speech. Second, the acoustic models for Chinese and English could be trained in different training conditions, such as recording condition and modeling resolution, and resulted in different scoring scales for these two languages. To compensate the score-gap problem, two methods, a fixed bias-term and a bias-weight incorporating with the duration information, are presented. Third, after decoding phase, we combine different information to select the most probable result for output. The Chinese part combines acoustic score, verification score, and tone score to generate the combination score. However, there are only acoustic score and verification score for English part. The dynamic range of the combination scores through using different scoring mechanism will be very different, and that makes system difficult to decide the final result for output. Therefore, in the re-scoring phase, a mapping table is developed to solve the normalization problem.
The architecture of this Chinese-English mixed-lingual keyword spotting is shown in figure 1.

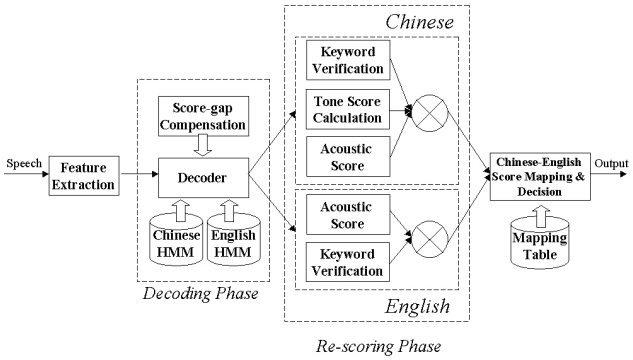

Figure 1 Mixed-Lingual keyword spotting

2. MIXED-LINGUAL KEYWORD SPOTTING

2.1. The former system - Chinese keyword spotting

Language dependent HMM-based acoustic models are used in our former system. We employed 138 sub-syllable models, including 100 3-state right-context-dependent INITIAL models and 38 5-state context-independent FINAL models, for recognizing Chinese speech. This part

of HMM was trained by using the MAT2000[2] database. For keyword spotting, two common-used filler-words are chosen to absorb the garbage words in front of and after the keywords, and we employed 163 pre-fillers and 29 post-fillers (Figure 2-1) for this purpose.

For selecting the most reliable result, N candidate-keywords with the acoustic scores, verification scores, and tone scores are provided to our re-scoring mechanism to select the most possible keyword for the final output.

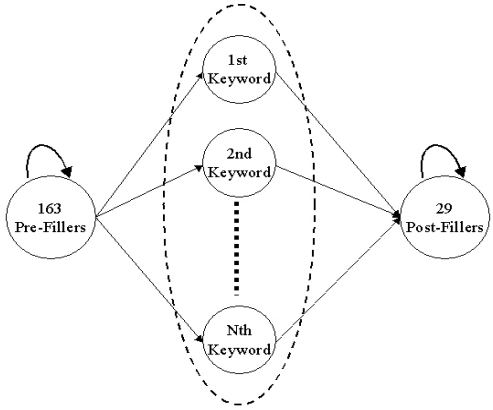

Figure 2-1 The Chinese keyword spotting

2.2. Models for Taiwanese-accented English speech

Language dependent HMM-based acoustic models are also used for English speech recognition. Three sets of acoustic models with different context-dependencies [3], including 40 context-independent phone models, 134 right-context-dependent phone models, and 595 left-right-context-dependent phone models, are combined to improve the recognition performance on Taiwanese-accented English speech.

2.3. Mixed-lingual keyword spotting

By directly using the Chinese and English HMMs for our preliminary experiment on the mixed-lingual speech recognition, we found that the Chinese keywords are recognized as English keywords easily, and the likelihood score of applying Chinese models is lower than the one applying English models in average. That means there is a score-gap between the two languages. The score-gap might come from different recording conditions, different training sets, and the modeling resolutions. It also could be came from the fact that the Taiwanese-accented English with highly pronunciation variations. Same experience on the score-gap problem can also be found in [4].

2.3.1. Score bias and duration model

Two methods are present to overcome the score-gap problem. The first one is using a fixed bias-term, b_{L_i}, on the log-likelihood calculation for each frame to compensate the score-gap, i.e.

$$S_t = \log P_{t,L_i} + b_{L_i} \quad (1)$$

The duration information and a bias-weight, w_{L_i}, are applied instead of fixed bias-term in the second compensation method, i.e.

$$S_t = \log P_{t,L_i} + w_{L_i} \log P_{t,L_i}^{DUR} \quad (2)$$

The phone-dependent (or sub-syllable-dependent) duration models with bi-asymmetric Gaussians (Figure 2-2) are referenced to calculate the term of P_{t,L_i}^{DUR}. The duration models were trained by using the MAT2000 database. The experiment shows that using the duration information and the bias-weight to compensate the score-gap is more suitable than that using the fixed bias-term.

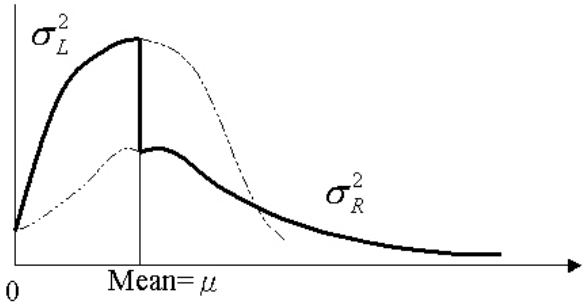

Figure 2-2 Bi-asymmetric Gaussians

2.3.2. Score mapping

As mentioned earlier, a re-scoring mechanism is applied to improve the system reliability, and the acoustic score, verification score, and tone score are referenced for the final decision of Chinese-keyword output. However, no tone score can be applied for the English part. That results in different dynamic range on the combination scores for these two languages. It's hard to select a proper answer using the combination score with different scoring mechanism. Therefore we need a normalization process to solve such scaling problem.

In the system view, the setting of false rejection rate is an important factor for system operation. Base on the relationship between false rejection rate and combination score, a score-mapping function can be found easily, and the combination score of one language can be projected to another language through using the same false rejection rate. Figure 2-3 shows the relation curves of false rejection rate and combination score for Chinese and English. The dash-line shows one of the examples that projects the combination score of Chinese (or English) to

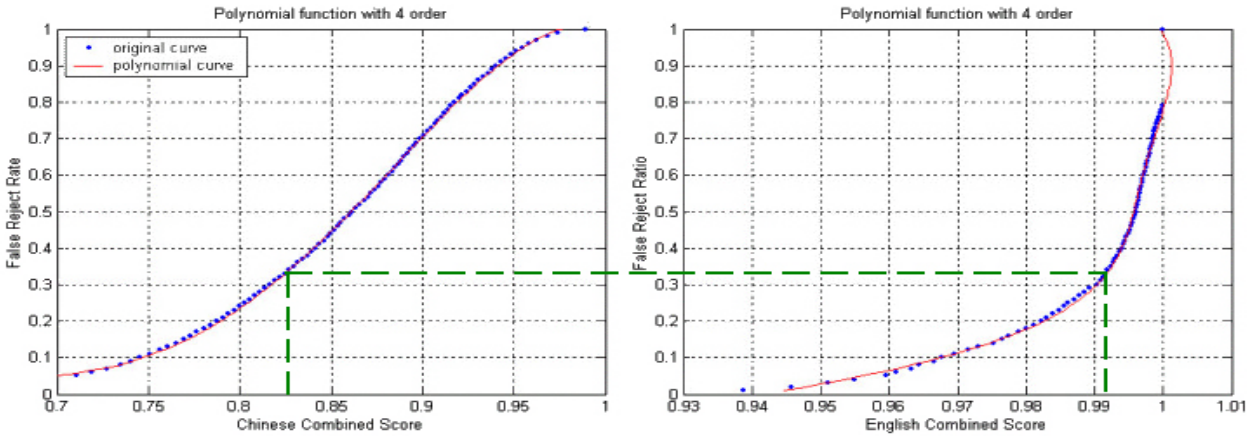

Figure 2-3 The relation curve of false rejection rate and combination score

the combination score of English (or Chinese) with the same false rejection rate of 0.33.

2.4. Threshold setting for system operation

Considering the friendly issue, two thresholds are set for system operation. First threshold is rather strict, which keeps the false rejection rate of top1 at 30%. The second one is slightly lax, which keeps the false rejection rate of top3 at 3%. The system will make a call-transfer to the top1 person directly if the combination score is larger than the first threshold. It will reject the answer directly and prompt the user to speak again if the combination score is smaller than the second threshold. And it offers the possible candidates to user to make a selection if the combination scores are between the first and the second thresholds.

Database	Annotation	Utt.	
TDB-0	Alphabets, isolated words, and sentences in English	male	5931
		female	5826
		total	11757
TDB-1	114 English person names	male	176
		female	37
		total	213
TDB-2	885 English isolated words	male	508
		female	172
		total	680
CDB-1	Chinese person names	total	936
CEDB-1	TDB-1 + CDB1	total	1149

Table 3-1 The Taiwanese-accented corpus

3. CORPUS

We collected the Taiwanese-accented corpus for training and testing. Detailed descriptions of the corpus are listed in Table 3-1. TDB-0 is the training database for training the Taiwanese-accented English models., and TDB-1 and TDB-2 are the testing database of Taiwanese-accented English. The corpus of TDB-0 includes English alphabets, English isolated words and English sentences. The corpus of TDB-1 includes 114 English names. The corpus of TDB-2 includes 885 isolated words that are extracted from dictionary published by ministry of education. The corpus of CDB-1 is a Chinese database collected by auto attendant system [1] in ITRI, and its vocabularies are all Chinese person names. The corpus of CEDB-1 is a Chinese-English database, and it is the combination of CDB-1 and TDB-1.

4. EXPERIMENTAL RESULTS

4.1. Taiwanese-accented English speech recognition

The features used in the following experiments consist of 12 MFCCs, 12 delta MFCCs, the delta log-energy, and the delta-delta log-energy. We used TDB-0 to train the Taiwanese-accented English HMMs. To improve the recognition rate, the decision tree [3] is used to cluster the RCD (right context dependent) and the LRCD (left and right context dependent) models based on 40 context independent phonemes. Table 4-1 shows the experimental results of using different combinations of HMMs. The recognition performance is improved when more contextual information are considered.

Model	TDB-1	TDB-2
CI40	73.24%	72.35%
RCD134	79.34%	75.29%
LRCD595	82.63%	69.41%
LRCD595+RCD134+CI40	86.85%	78.53%

Table 4-1 Text dependent model testing

4.2. Score compensation for mixed-lingual keyword spotting

We use a fixed bias-term (equation (1)) to compensate the score-gap between English and Chinese keyword recognition. The vocabularies of keywords contain 114 English names, 1000 Chinese names and 37 mixed lingual (Chinese and English) names. And the results (row-3, 4, 5 of table 4-2） show the improvements on the Chinese speech with different fixed bias-terms, however, the negative results for the other targets, English and Mixed speeches, are also obtained.

We also use duration information into the likelihood calculation (equation (2)) and turn the bias-term into the bias-weight. The table 4-3 shows results of using different bias-weights for Chinese and English. It suggests that using the duration information and the bias-weight to compensate the score-gap is more suitable than that using the fixed bias-term.

CE_BIAS	Fixed bias-term		
	TDB-1	CDB-1	CEDB-1
0.00	91.08%	85.15%	90.61%
1.00	88.26%	91.77%	93.43%
1.50	86.85%	92.52%	91.08%
2.00	83.57%	93.27%	89.20%

Table 4-2 The keyword-spotting results of using the fixed bias-term

CH DW	EN DW	CE Duration Model		
		TDB-1	CDB-1	CEDB-1
1.00	1.00	92.02%	89.10%	93.43%
1.00	1.50	91.08%	90.38%	93.43%
1.00	2.00	91.08%	91.99%	93.43%
1.00	3.00	89.67%	93.38%	92.96%

Table 4-3 The keyword-spotting results of using duration information and bias-weights

4.3. Score mapping of combination score

Database		CEDB-1
Top1		96.71%
Top2		97.18%
Top3		97.18%
High Thsld	FR	61.65%
	SUB	0.94%
Low Thsld	FR	6.28%
	SUB	1.88%

Table4-4 The performance of mixed-lingual keyword-spotting

We use a mapping table of combination score for English and Chinese keyword spotting. CEDB-1 is used for this experiment. The vocabularies in database are the same as Section 4.2. The result, Table4-4, shows that the system performance and friendly issue can be further improved and solved by the mechanism of threshold setting. High threshold keeps the false rejection rate at 61.65% so the system can make a call-transfer for 38.35% correct queries directly. Low threshold keeps the false rejection rate at 6.28% so we would only reject 6.28% correct queries wrongly. The recognition rate of Top1 rises to 96.71%. The mechanism also maintains stable recognition rate.

5. CONCLUSION

In this paper, we presented a Chinese-English mixed-lingual keyword spotting for expanding the usability of our auto-attendant system. To solve different scoring scales for these two languages, a bias-compensation method was proposed to make up the score-gap and a mapping table was developed to normalize the combination scores of using different scoring mechanisms for Chinese and English.

Because the pronunciational variation is a very serious problem in Taiwanese-accented English, we are planing to collect more speech data for Taiwanese-accented English speech recognition in the future. Besides, we will integrate other speech modules, such as speech pre-classifier, speaker classifier, and speaker-adaptation to improve our auto-attendant system.

6. REFERENCES

[1] W.-C. Shieh, S.-C. Chien, J.-S. Hsu, and S.-C. Chang, “Improvement on ITRI 104 Auto-Attendant System,” CCL Technichal Journal, No. 96.

[2] H.-C. Wang, “MAT – A Project to Collect Mandarin Speech Data Through Networks in Taiwan,” Computational Linguistics and Chinese Language Processing, Vol. 2, No. 1, pp. 73-89, Feb. 1997.

[3] P.-Y. Liang, J.-L. Shen, L.-S. Lee, “Decision Tree Clustering for Acoustic Modeling in Speaker-Independent Mandarin Telephone Speech Recognition,” ICSLP 98.

[4] B. Ma, C.-T. Guan, H.-Z. Li, and C.-H. Lee, "Multilingual Speech Recognition with Language Identification," ICSLP 2002.