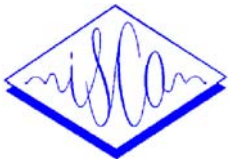

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DEVELOPMENT OF THE PHILIPS 1999 TAIWAN MANDARIN BENCHMARK SYSTEM

Chiwei Che, Nick Wang, Max Huang, Hank Huang and Frank Seide
Philips Innovation Center, Taipei
24 FB, Sec. 1, Chung Hsiao West Road. Taipei, Taiwan
Email: cche@prlt.research.philips.com

ABSTRACT

This paper describes the Philips Large Vocabulary Continuous Mandarin speech recognition system for the 1999 Taiwan benchmark. The basic system architecture is based on the Philips LVCSR technology developed for Western languages. However, several modifications are made in order to better suitted processing Chinese spoken languages. In the paper, we present some experimental results on the two tasks we participated in this benchmark: digit and continuous syllable. For the development set, we were able to obtain a digit/syllable error rate of 2.9%/23.9%. At the final evaluation, our system achieves the lowest error rate of 3.1%/24.3% among all participating sites.

Keywords: Mandarin, benchmark, speech recognitionour

1. INTRODUCTION

Speech recognition benchmark activity on Mandarin has only been started recently. These include the 1997 HUB4NE task by NIST at United State, the 1998 863 national project evaluation by China government at Mainland China, and most recently, the 1999 ASR benchmark by ROCLING[1] at Taiwan. It clearly shows the increasingly importance and emerging need of the Mandarin recognition technology. It is also largely dued to the fact that speech provides the most convinent man-computer interface for Chinese input.

In this paper, we describe the development process of the Philips Large Vocabulary Continuous Mandarin recognition system for the 1999 Taiwan ASR benchmark. Although Philips has gained many experiences in participating Western languages speech reognition benchmark, this is an initial effort for Asian language. We believed that through the benchmark pratice, investigation and understanding of the common tasks will bring speech recognition research a step furthur towards Mandarin.

[1] Computational Linguistic society of Taiwan

The Taiwan ASR benchmark consisits of the following tasks: continuous syllable, continuous digit and isolated/noisy digit where the Philips took part in the first two. Material used in all tasks are speech recorded via local Taiwan telephone network. The corpus, MAT-160A and NUM-100A [3] are provided to sites with no training data. However, it is not limited to use other self acquired acoustic training matreials. A development set data, including 500 utterances, is given for sites to develop their system. The final evaluation uses another set of 1000 utterances for final system evaluation.

Organization of this paper are given as follows. In section 2, we describe the baseline system set up. Recognition experiments conducted for digit and syllable string tasks are illustrated in section 3. In section 4, the summary of result and future direction is presented. Final section gives the conclusion.

2. BASELINE SYSTEM

The baseline system uses continuous density HMM in modelling the sub-syllablic unit in Mandarin. Front-end feature consists of Mel Frequency Cepstral Coeefficients (MFCC) and its time derivitives. The current benchmark system does not have any treatment on tone.

The available acoustic training corpus is given in the following table:

Corpus	# Speakers	Short desc.
MAT-160A	160	Recorded via telephone. Isolated syllable, command word and continuous speech
NUM-100A	100	Recorded via microphone. Continuous digit with length ranges from 1 to 7.
MAT-800	800	Same as MAT-160A

Table 1. Language resources used for the benchmark

3. EXPERIMENTAL RESULTS

3.1 The Continuous Digit Task

For the continuous digit task, whole digit HMM models are used instead of sub-syllabic modelling in our normal setup. Some preliminary experimental results suggested that the use of whole digit model produce superior recognition performance. NUM-100A is used to train the gender independent acoustic model. Recognition result on the 500 utterances development set (Dev set) is 18.1% digit error rate. The performance is far from what we expected on the telephone digit string task. Mismatched training and testing data might cause such degradation. The system parameters are also nearly optimized.

3.1.1 Incoporating Finite State Network (FSN) for length constraints

We made a strong assumption that number of digit in one utterances cannot exceed 7. The idea is then to guide the recognizer search with such a constraint. The error rate drops from 18.3% to 9.7%. An order of magnitude improvement is achieved.

3.1.2 Corpus Effect and System Optimization

Realizing the difference between the initial training corpus and Dev set, we then use MAT-160A plus MAT-800 to train acoustic models based on telephone data. A 6.5% error rate is then obtained. Thus, the data driven solution provides a relatively 30% gain on top of FSN. After we replace the training corpus, some standard system optimization procedures are made. First, we train a gender dependent acoustic model, that gives another 26% improvement and the error rate is now 4.8%. Finally, we optimize the number of state for each digit HMM models and the word penalty during the search. It results in the error rate of 3.2% without using any adaptation.

3.1.3 Speaker Adaptation

Unsupervised speaker adaptation is realized in two steps. First, in feature domain, we apply the Vocal Tract Normalization (VTN) procedure. A maximum likelihood approach is used to select the frequency-warping factor for VTN. Secondly, the Maximum Likelihood Linear Regression (MLLR) is used after the preliminary pass of decoding. A global matrix is used for sentence based adaptation. Iterative application of MLLR is applied, however, no further improvement is found. Combining the VTN and MLLR, the system achieves an error rate of 2.9%.

3.2 The Continuous Syllable Task

Compared to the continuous digit task, setup used for continuous syllable is much more complicated. First, we use a preme/core-final structure to model the sub-syllabic units of Mandarin. It is mostly based on our previously reported study on Mandarin phonetic modeling [8]. Recognition experiments of continuous syllable are divided into results using or without word/phrase based bigram language model (LM). The word/phrase text segmentation uses the longest match approach. Lexicon uses the top 40K entries from the Academy Sinica lexicon. MAT-160A and MAT-800 are used to train the right context dependent preme and context independent core-final models. The baseline system produces a 35.6% syllable error rate without using any LM.

3.2.1 Corpus and Language Model Effect

The following figure illustrated the syllable error rate in terms of amounts of training data and the application of language model.

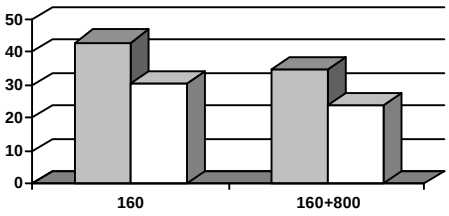

Figure 1. Corpus and language model effect on syllable error rate

The horizontal axis represents the syllable error rate and the vertical axis shows the number of speakers data contained in the training corpus. Grey/white bar gives the performance with/without LM. Not surpassingly, adding more training data decrease the error rate. However, one interesting observation is the large error reduction by using LM. Ideally, we would imagine that the intention of measuring the syllable accuracy is purely on the acoustic level. Obviously, the Dev set contains very rich syntactic information. Finally, we decide to submit both results (with/without LM) to the evaluation and designated the language model as the preferred one.

3.2.1 Optimization and Adaptation

Similar optimization techniques described in previous section on continuous digit are applied to the syllable task. The recognition performances are measured with no language model. Gender dependent modeling gives 33.9% error rate and ap-

plication of speaker adaptation can further reduce it to 33.3%.

3. RESULT SUMMARY AND FUTURE DIRECTION

The following table gives a summary of improvement on the continuous digit task:

Techniques	Error rate (%)
Baseline	18.1
+FSN	9.7
+MAT160+800	6.5
+GD	4.8
+HMM Length Est.	3.8
+Word Penalty Opt.	3.2
+VTN, MLLR	2.9

Table 2. Summary of improvement by different techniques

As shown in table 2, use of FSN and matched training data give most of the gain. However, the application of some up to date technology, such as VTN, MLLR, also reduces the error rate to a large extent. For the continuous syllable task, it uses already an optimized system, but similar improvement by standard optimization and adaptation techniques is observed. The final official released results for both tasks are given in the following table.

	Digit Task	Syllable Task
Dev Set	2.9	23.8
Eval Set	3.1	24.2

Table 3. Recognition performance (error rate in %) of development data and evaluation data

From the above table, we observed quite similar recognition performance for both the Dev and Eval set data.

4. CONCLUSION AND FUTURE WORK

Our main focus on this benchmark evaluation is to achieve the best performance on Mandarin recognition. It has been paid off by applying up to date LVCSR techniques, although originally developed for Western languages. Our system gives state of the art performance and achieved the lowest digit and syllable error rate among all participants. In the future, we will continue on the investigation of tone which is not included in this benchmark, also the duration modeling which is effective in digit task. At last, we would also like to work more on the accent adaptation for a more robust system.

5. ACKNOWLEDGEMENT

Acoustic training materials are partly provided by Computational Lingustic Socisity of Taiwan, Republic Of China (ROCLING). The authors want to thank Prof. Wang Hsiao Chuan for assistance during the evaluation.

6. REFERENCES

[1] Legetter, C. J. and Woodland, P. C. (1996), Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Comput. Speech Lang., vol. 9, pp. 171-186.*

[2] Chiou, R. L. and Wang, H. C. (1998), A Preliminary Test of MAT-160 Speech Database in Connected Syllables Recognition. *Proceeding of the 1998 International Symposium on Chinese Spoken Language Processing*. pp. 89-92.

[3] Wang, H. C. (1997), MAT – A Project to Collect Mandarin Speech Data Through Telephone Networks in Taiwan. *Computational linguistic and Chinese language Processing. Vol. 2, no. 1, pp. 73-90.*

[4] Chen, C. J., Gopinath, M. D., Monkowski, M. D., Picheny, M. C. and Shen, K. (1997), New Methods in Continuous Mandarin Speech Recognition. *Proc. EuroSpeech*, pp. 1543-1546.

[5] Aubert, X. et al. (1994), Large Vocabulary Continuous Speech Recognition of Wall Street Journal data. *ICASSP*, vol. II, pp. 129-132.

[6] Dugast, Ch et al. (1995), The Philips Large Vocabulary Recognition System for American English, French and German. *Proc. Euro-Speech*, pp. 197-200.

[7] Wang, H. M. et al. (1995), Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary but Limited Training Data. *Proc. ICASSP*, pp. 61-64.

[8] Seide, F. and Wang, J. C. (1998), Phonetic Modelling in the Philips Chinese Continuous Speech Recognition System. *Proc. of the 1998 Interment national Symposium on Chinese Spoken Language Processing*. pp. 54-59.

[9] Beyerlein, P. et al. (1997), Modelling and Decoding of crossword Context Dependent Phones in The Philips Large Vocabulary Continuous Speech Recognition System. *Proc. EuroSpeech*, pp. 1163-1166.

[10] Beulen, K. and Ney, H. (1998), Automatic Question Generation for Decision Tree Based Tying. *Proc. ICASSP*, Vol. 2, pp. 805-808.

[11] Hong, XU., Frederic, B., and Hugo, VH. (1998), Adapting Western Language Recognizer for Chinese Recognition. *Proc. of the 1998 International Symposium on Chinese Spoken Language Processing*. pp. 60-63.