IMPROVEMENTS OF THE PHILIPS 2000 TAIWAN MANDARIN BENCHMARK SYSTEM

Yuan-Fu Liao1, Nick Wang1,2, Max Huang1, Hank Huang1 and Frank Seide1

1Philips Research East-Asia, Taipei
2Graduate Institute of Communication Engineering, Taiwan University, Taiwan
24FB, 66, Sec. 1, Chung Hsiao W. Rd., Taipei, China
E-mail: {leo.liao, nick.wang, max.huang, frank.seide, hank.huang}@philips.com

ABSTRACT

In this paper, we present the Philips large vocabulary continuous Mandarin speech recognition system developed for the 2000 Taiwan Speech Input Technology Assessment. We systematically integrated key Mandarin components with up-to-date Western-language techniques to build up a state-of-the-art Mandarin speech recognition system. These technologies include robust pitch extraction/tone modeling, context-dependent preme/core-final units, Chinese phrase/syllable trigram language model, linear discriminant analysis (LDA), cross-syllable modeling/decoding, speaker clustering and maximum likelihood linear regression (MLLR) adaptation. Among them, the major breakthroughs were our robust pitch extraction/tone modeling technology and the treatment of coarticulation across syllable boundaries. For the development set, we dramatically reduced last year’s best error rates by relative 44.8%~67.8% on all three categories we participated. Moreover, for the evaluation set, we achieved the lowest unit error rates on all three categories.

1. INTRODUCTION

Speech recognition benchmark activities on Mandarin have just been started recently. These include the 1997 HUB4NE task by National Institute of Standards and Technology (NIST) in United States, the 1998 863 National Hi-Tech Project evaluation in China, and most recently, the 1999 and 2000 Speech Input Technology Assessments by the Association for Computational Linguistics and Chinese Language Processing (ROCLING) in Taiwan [1]. This clearly shows the increasing importance of the Mandarin recognition technology. In this paper, we report the system we used in the recent 2000 Taiwan benchmark.

In the 1999 Taiwan benchmark, an initial effort was made to apply Philips’ large vocabulary continuous speech recognition (LVCSR) techniques to Mandarin. The major work was to use the preme/core-final units to model the syllable structure of Mandarin since Mandarin is a monosyllabic language [2]. For this second time that a benchmark was held, more specific characteristics of Mandarin were considered to improve the recognition accuracy. First, Mandarin is a tonal language. The five lexical tones constitute a phonemic element important for Mandarin speech recognition. The same syllable with different embedded tone will have a totally different meaning. Therefore, pitch features and tone modeling are necessary. Second, Chinese language uses ideographic characters. Every Chinese character can be used as a single-character word and a phrase is made of several these kind of single-character words. Since each character is pronounce as exactly one syllable, pronunciation effect across these Mandarin words (or we should say syllables) is much stronger. The careful treatment of coarticulation across word (syllable) boundaries is necessary for Mandarin speech recognition. Therefore, in this year, we tried to improve our system by systematically integrating these key components that we had developed for Mandarin as well as many other necessary speech recognition techniques that had been successfully deployed in Western-language systems.

The Taiwan benchmark was motivated and came with a 3-year project to collect Mandarin speech data across Taiwan (MAT) of over 7,000 Taiwanese individuals via the telephone network [1]. The first benchmark was held in 1999 and had built a healthy infrastructure for Mandarin speech recognition research. In this second time the benchmark was held, the assessment consisted of the following tasks: (1) continuous-syllable and (2) digit recognition. These tasks were further divided into four contests. They were

- (1-A) continuous-syllable recognition with limited training corpus
- (1-B) continuous-syllable recognition with unlimited training corpus
- (2-A) isolated-digit recognition with limited training corpus
- (2-B) continuous-digit recognition with unlimited training corpus

Philips took part in the contest 1-A, 1-B and 2-B. For those contests with limited corpus, three official corpora, MAT-160, MAT-400 and NUM-100A were provided and only allowed for system training. A development set including 500 continuous-syllable and 500 continuous-digit utterances was given the same as 1999. This set was selected from the first two years MAT recordings. But the NUM-100A was a different microphone recording. The 2000 evaluation set used another 1,500 continuous-syllable, 1,200 isolated-digit and 1,500 continuous-digit utterances selected from the last year MAT recordings. The difference between the first two years and the last year MAT recordings was that the later had about 30 additional newspaper items.

Organization of this paper is as follows: In Section 2, we describe the training corpora and the baseline system. In Section 3, we describe the newly added technologies that led to a significant overall improvement. The summary of results is given in Section 4. Conclusions are in the final Section.
2. BASELINE SYSTEM AND THE EFFECT OF LANGUAGE RESOURCES

The training corpora we used this year included MAT-160+MAT-400 (from now on we refer to it as MAT-560) DB3-5 for contest 1-A, MAT-2000 DB3-5 for contest 1-B and MAT-2000 DB2 for contest 2-B. All corpora came from the first two years of MAT recordings. Among them, MAT-560 was the official training corpus. Unlike those non-validated corpora MAT-160 and MAT-800 used last year [2], the new MAT-2000 was carefully validated [3]. We therefore used the MAT-2000 to correct the transcriptions of the MAT-560 since MAT-560 was a subset of MAT-2000. Some statistics of those corpora are listed in Table 1.

<table>
<thead>
<tr>
<th>Contest</th>
<th>Corpus</th>
<th>Spks</th>
<th>Utrs</th>
<th>Hrs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-A</td>
<td>MAT-560 DB3-5</td>
<td>560</td>
<td>28,668</td>
<td>16.3</td>
</tr>
<tr>
<td>1-B</td>
<td>MAT-2000 DB3-5</td>
<td>2,232</td>
<td>113,475</td>
<td>62.5</td>
</tr>
<tr>
<td>2-B</td>
<td>MAT-2000 DB2</td>
<td>2,232</td>
<td>10,933</td>
<td>8.5</td>
</tr>
</tbody>
</table>

Table 1. Language resources used in the 2000 benchmark.

The baseline system just followed our 1999 benchmark setting [2]. Continuous density hidden Markov models (HMM) were used. A 28-dimensions feature vector included Mel Frequency Cepstral Coefficients (MFCC), its first and second order time derivatives was used. The syllable recognizers used within-syllable right-context-dependent (RCD) premes and context-independent (CI) core-finals [2,4]. The numbers of states were automatically estimated from training corpus. The bottom-up state tying were used to reduce system parameters. For contest 1-A, gender independent (GI) models were used. For contest 1-B, gender dependent (GD) models were used. The digit recognizer used GD and whole word digit models.

By using larger training corpora, we increased the maximum number of mixture splits from five to six for contest 1-A and to seven for 1-B to increase system resolution. But we left out the VTN, MLLR, finite state network (FSN) and language models [2] in order to examine the effect of different language resource. Recognition results on the development set are listed in Table 2. It is worth noting that our 2000 baseline results were already better than 1999’s baseline and comparable with 1999’s best results [2].

Table 2. Recognition results of the 2000 baseline system, 1999’s baseline system and 1999’s best results. Here no language model or adaptation was applied for all baseline systems. But the VTN and MLLR adaptation were used for the 1999’s best results (unit error rate in %).

<table>
<thead>
<tr>
<th>Contest</th>
<th>2000 baseline</th>
<th>1999 baseline</th>
<th>1999 best</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-A</td>
<td>34.0%</td>
<td>42.5%</td>
<td>-</td>
</tr>
<tr>
<td>1-B</td>
<td>30.7%</td>
<td>33.9%</td>
<td>33.3%</td>
</tr>
<tr>
<td>2-B</td>
<td>3.2%</td>
<td>3.2%</td>
<td>2.9%</td>
</tr>
</tbody>
</table>

3. IMPROVEMENTS

For the benchmark 2000, we systematically integrated several key Mandarin components and many other necessary speech recognition techniques that had been successfully deployed in Western-language systems. They included robust pitch extraction and tone modeling, LDA, cross-syllable modeling and decoding. Chinese phrase/syllable trigram language model, unsupervised adaptation based on speaker clustering and MLLR adaptation.

Among them, the major breakthroughs were our robust pitch extraction and tone modeling technology [5-6] and the treatment of coarticulation across syllable boundaries i.e., cross-syllable modeling and decoding [7]. In the following, we describe all techniques for system building. It has to be noted that only pitch features were added into our digit recognizer.

3.1 Robust pitch features extraction and tone modeling

Since Mandarin is a tonal language, robust pitch extraction [5-6] was used to generate five pitch features including pitch, its first and second time derivatives, degree-of-voicing and its first order derivatives. Some parameters of our pitch extraction algorithm were adjusted to adapt to telephone environment. Especially, the dynamic programming search interval were adjusted to include the whole sentence. The smooth range of pitch contour was also extended. These 5 pitch features were merged with the 28 MFCCs to become a 33-dimension feature vector and used in both digit and syllable recognition. By using the pitch features, the core-final models were expanded according to their different embedded tones. Since the number of models became much larger, the bottom-up state tying were applying to both RCD premes and tone-dependent (TD) core-finals. It reduced the number of system parameters to let each parameter have enough training material.

The results are shown in Table 4. About 10% relative error rate reduction were achieved for contest 1-A and 1-B. It was interesting that pitch features could also dramatically help the recognition of digit string, 43.7% relative gain was achieved.

One possible reason was that pitch could help the segmentation of digits, because the pitch contour is considered in the vowel part. The other reason was that a sequence of two connected single-vowel digits, “1-1” (yi1-yi1), “2-2” (er4-er4) and “5-5” (wu3-wu3), were often only recognized as a single digit, since there were almost no clear syllable-boundary cues from the view point of MFCC features. Here the numbers after the phonetic transcription were the tone marks. But tone 4 had a high-and-low, and tone 3 had a middle-low-high pitch contour pattern, the boundary cues were more clearly supported by the pitch features.

3.2 Linear discriminate analysis

LDA was used to explore the temporal information of current and neighbor frames and to diagonalize the speech features to fit the HMM assumptions. The vectors of the current, and its left and right frames (t-1, t, t+1) were compacted into 50-dimension vectors using the mixture-level LDA. The relative gains from
LDA were 7.9% and 5.6% for contest 1-A and 1-B.

### 3.3 Cross-syllable modeling and decoding

To take care of the strong coarticulation effect across the boundaries of the Mandarin syllables, the tonal core-finals were further expanded according to their succeeding preemes. This expansion resulted in two problems. One was that the number of possible combination was very large. Therefore, there were many unseen units and the training material were definitely not enough. The other was that the decoder had to keep the decision open and follow all possible hypotheses until the next syllable was known. To solve the first problem, a decision-tree state-tying algorithm was applied [8]. Unlike those question sets used in Western language, the question set used here had to include additional tone context questions. As for the second problem, the Philips Research one-pass cross-word decoder that considered both language model history and cross-word context was used [7]. The performance of the cross-syllable modeling is shown in Table 4. About 10% relative gain was achieved for both contest 1-A and 1-B.

### 3.4 Chinese phrase/syllable trigram language model

By using the new one pass cross-word decoder [7], we now could directly apply the trigram language models in one pass instead of the lattice based two-pass trigram rescoring. A 400K text corpus and a 56K phrase lexicon were used to estimate the phrase trigram models for contest 1-B. But for contest 1-A, the limited training corpus one, only the MAT-560 text could be used. Since the training material was quite limited in MAT-560, only a syllable trigram was estimated for contest 1-A. In all language models, we carefully removed any texts that overlapped with the texts of the development set to avoid any bias. The results of applying the language model are shown in Table 4. For contest 1-A, the syllable trigram helped a little, 5.6% relative. But for contest 1-B, the using of the phrase trigram had a relative 60.4% gain.

### 3.5 Unsupervised adaptation based on speaker clustering and MLLR

Unlike our last year’s benchmark system, that performed the acoustic adaptation using only one sentence or the whole test data, this year we incorporated a speaker-clustering algorithm in the front-end of the MLLR adaptation. The clustering algorithm is based on the second order of the Kullback-Leibler (KL) distance and uses a bottom-up-tying approach to merge those utterances with similar speaker characteristics. This algorithm worked well in the HUB4 evaluation [9].

The unsupervised adaptation was implemented in three steps. The first step was to apply the speaker-clustering algorithm to cluster the input utterances. This was to make sure there was enough adaptation material for the succeeding MLLR operation. The second step was to recognize those clustered utterances using current reference model in order to generate a transcription. The third step was to apply MLLR adaptation using those transcriptions to update the reference model for each cluster. The last two steps could be repeated for several iterations.

The adaptation results after three iterations are shown in Table 4. Although, only a global MLLR matrix was used in the adaptation procedure, much larger gain compared to our last year’s results [2], 12.3%~15.7% vs. 2.0%, for contest 1-A and 1-B had been achieved.

### 4. SUMMARY AND DISCUSSION

In Table 4, we summarize the overall improvements on the development set for the three contests. Compare to the best results of the 1999 Taiwan benchmark, 23.3% and 2.9% for contest 1-B and 2-B [2], significant gains have been achieved. They were relative 44.8% and 67.8% error rate reduction for contest 1-B and 2-B. Evaluation results measured by the benchmarking committee are shown in Table 5. In all three categories we participated in, we achieved the lowest unit error rates. Finally, Figure 1 shows the development and improvement of all used techniques for the contest 1-B. As shown in the figure the improvements were consistent in all development directions.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Contest 1-A</th>
<th>Contest 1-B</th>
<th>Contest 2-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>34.0%</td>
<td>30.7%</td>
<td>3.2%</td>
</tr>
<tr>
<td>+ pitch &amp; tone modeling</td>
<td>30.5%</td>
<td>26.6%</td>
<td>1.8%</td>
</tr>
<tr>
<td>+ LDA</td>
<td>28.1%</td>
<td>25.1%</td>
<td>-</td>
</tr>
<tr>
<td>+ cross-syllable modeling</td>
<td>25.0%</td>
<td>22.5%</td>
<td>-</td>
</tr>
<tr>
<td>+ syllable or phrase trigram</td>
<td>23.6%</td>
<td>8.9%</td>
<td>-</td>
</tr>
<tr>
<td>+ speaker clustering &amp; MLLR</td>
<td>20.7%</td>
<td>7.5%</td>
<td>-</td>
</tr>
<tr>
<td>+ FSN</td>
<td>-</td>
<td>-</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

Table 4. Summary of improvement by different techniques on the development set (unit error rate in %).

<table>
<thead>
<tr>
<th>Contest</th>
<th>Philips</th>
<th>Competitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-A</td>
<td>33.3%</td>
<td>44.0%</td>
</tr>
<tr>
<td>1-B</td>
<td>12.2%</td>
<td>18.3%</td>
</tr>
<tr>
<td>2-B</td>
<td>3.5%</td>
<td>5.4%</td>
</tr>
</tbody>
</table>

Table 5. Evaluation results of Philips and its competitor (unit error rate in %).

It is worth noting that there were large performance gaps between the results of the development and evaluation set. This may not come from system over-fitting but from the mismatch of the development and evaluation set. The reasons are the following:

1. According to the official evaluation report, all participants suffered serious performance degradation. The error rates of our competitor even increased from 13.7% to 44.0% and from 8.0% to 18.3% for contest 1-A and 1-B.
2. The texts of the training corpora, development set and the evaluation set were different. The first two contained only well designed phonetic balance sentences. But the later had additional newspaper items.

3. After the benchmark, some rough analysis was done, we found that the development set was quite clean and homogeneous. But, we felt that the content of the evaluation set may be more close to real-life situation. For example, it contained recordings from children and old men. Some utterances were spoken very fast or very slow. Some utterances had strong channel noise. In addition, there are some totally empty utterances, some had only half a syllable, one even contained very long music after the speech segment.

5. CONCLUSIONS

In this paper, we successfully integrated key Mandarin components with up-to-date Western-language techniques to build up a state-of-the-art Mandarin speech recognition system. Our system achieved the lowest unit error rate on all three categories we participated in the 2000 Taiwan benchmark. We hope that the combination of those techniques marks a new milestone for Mandarin speech recognition. We also hope that sharing of our experience will be useful for future Mandarin speech recognition research.

ACKNOWLEDGEMENTS

The authors would like to thank Xavier Aubert, Matthew Harris, Reinhold Haeb-Umbach, and Carsten Meyer, their Philips Research Labotary Aachen colleagues, for their valuable supports. They also want to thank ROCLING and Prof. Hsiao-Chuan Wang for leading the collection of the MAT database and hosting the 2000 Taiwan Benchmark.

REFERENCES