ADVANCES IN MANDARIN BROADCAST SPEECH RECOGNITION

Mei-Yuh Hwang¹, Xin Lei¹, Jing Zheng², Ozgur Cetin³, Wen Wang², Gang Peng¹, Andreas Stolcke²,³

¹Univ. of Washington, Dept. of Electrical Engineering, Seattle, WA 98195 USA
²SRI International, Menlo Park, CA, 94025 USA
³International Computer Science Institute, Berkeley, CA, 94704 USA

{mhwang,leixin,gpeng}@ee.washington.edu, {zj,wwang,stolcke}@speech.sri.com,
ocetin@icsi.berkeley.edu

ABSTRACT

We describe our continuing efforts to improve the UW-SRI-ICSI Mandarin broadcast speech recognizer. We augmented acoustic and language model training data and obtained first results using the broadcast conversation genre. The front end was enhanced using features estimated by multilayer perceptrons, and we incorporated discriminative training techniques such as MPE and fMPE. The decoding architecture was also improved to take advantage of cross-adaptation and system combination. The net effect was a 24%–38% relative reduction in character error rates (CER) on a variety of test sets.

Index Terms— Mandarin speech recognition, character error rate, MLP features, discriminative training.

1. INTRODUCTION

In previous work [1], we built a competitive speech recognizer for Mandarin broadcast news. This paper describes our recent efforts to improve along a number of dimensions, from front end to acoustic modeling to language models (LMs). Section 2 summarizes the data used in training, development, and testing. The description of our development work starts with lexicons and language models in Section 3. Section 4 describes the acoustic modeling work, including a novel combination of discriminative MLP features, fMPE transforms [2], and MPE [3] model training. Section 5 describes our decoding structure and how cross-adaptation and confusion network combination (CNC) [4] are used in the overall system. Section 6 presents experimental results. Section 7 concludes and lists the challenges ahead.

2. DATA

2.1. Acoustic data

Our acoustic data are all from various LDC Mandarin corpora. Table 1 lists the sources of data used to train our acoustic models (AMs). The total amount of data is 465 hours of speech: 313 hours of broadcast news (BN) and 152 hours of broadcast conversations (BC). BC is a genre defined to include broadcasts characterized by spontaneous conversations, such as talk and call-in shows.

The sources of these recordings are indicated in the table. We used a flexible alignment algorithm to filter out segments with bad transcripts [5] in the TDT4 corpus, and kept about 89 hours of data. GALE (Global Autonomous Language Exploitation) is a new DARPA-sponsored program, aiming at translation into English. We used all the GALE Year 1 Q1 and Q2 acoustic data that come with human quick transcriptions. During development, we focused on the broadcast news genre. Our development data included the 2004 development set for the EARS program, dev04, and the evaluation set for that year, eval04. After we finished tuning our system based on dev04 and eval04 we then applied the system to the GALE extended dryrun evaluation set (ext06) and the BC development set defined by Cambridge University (dev05bc). These datasets also contain some foreign (non-Mandarin) speech. More information about the test sets used is in Table 2. Naturally, these sets are excluded from either AM or LM training.

2.2. Language model data

Table 3 lists all the textual data used in LM training and development. As in previous systems, Hub4 and all TDT2, TDT3, and TDT4 data were used for LM training. The TDT4 text corpus (LDC2003T16) contains more programs and more text sources (Xinhua and Zaobao) than the speech corpus and was used in its entirety for LM training. Additionally, the Multiple Translation Chinese Corpus parts 1, 2, and 3, and the Chinese News Translation...
The Gigaword corpus (LDC2005T06) were added to train the first LM in Table 3, labeled “TDT”.

The LDC GALE text data include all the transcriptions of the GALE acoustic data listed in Table 1, plus the GALE web transcription (closed-caption-like). These data are more similar to speech test data, as they correspond to real speech rather than written articles exclusively. They were used to train the second LM.

The Gigaword corpus contains articles from three newswire and newspaper sources: Central News Agency from Taiwan, Xinhua newspaper from China, and Zaobao from Singapore. They were thus subdivided into three sets to train three other LMs. For purposes of LM training, we considered all newswire text as being part of the BN genre.

National Taiwan University (NTU) downloaded news articles and conversation transcriptions from CCTV, PHOENIX, and VOA web sites (dated before February 2006) to cover some of the sources missing from the LDC GALE data. These data do not necessarily correspond to speech. Yet they are more like GALE data than the Gigaword corpus, as they are from the same broadcast sources, rather than from newswire articles. They were used to train a sixth LM.

Finally, Cambridge University (CU) downloaded newswire text from a variety of Chinese newspaper sources and BN transcripts from CNR, BBC, and RFA. They were used to train the seventh LM. The total number of words in all the training texts was 946 million.

To combine multiple LMs, we designated the GALE 2006 BN development set (dev06) as our LM tuning set. Multiple LMs were interpolated to form a single LM, so as to maximize the likelihood on dev06. Dev06 is a superset of dev04 and eval04. It also contains the 2003 NIST Rich Transcription BN evaluation set and some new data from the GALE Year 1 BN audio transcript release. Section 3 will present more details on the construction of our language models.

Note that in choosing the LM training data, all text data from the same months as contained in dev06 were excluded, to avoid memorization of the test data. However, because we had little BC data, we only excluded data within a one-week time window centered on dates covered by dev05bc.

3. LANGUAGE MODELS

To do Chinese word segmentation, we started from the 49K-word unigram LM trained in [1], and then added a few thousand Chinese words from the LDC lexicon and another few thousand names automatically identified in the TDT4 corpus. The added new words were given a constant unigram probability. We then used this expanded unigram to perform maximum likelihood segmentation on all text data. The most frequent 60K words in the training text were then chosen as our decoding vocabulary. This vocabulary included 1760 English words. As described in Section 2, and summarized in Table 3, seven N-gram LMs were independently trained and then interpolated to maximize the likelihood on dev06. Each individual LM was trained with modified Kneser-Ney smoothing [6] using the SRILM toolkit [7]. All partial words, noises, laughter, out-of-vocabulary (OOV) words, and non-Mandarin words that were not English were mapped to a special “garbage” word whose N-grams were also statistically trained. The garbage word was designed to absorb non-speech noises and OOVs.

Five LMs were used in decoding. One, a bigram, was trained and highly pruned for fast decoding in the first recognition pass. Two versions of trigrams were trained. A highly pruned version was again used for fast search and a full version for trigram lattice expansion and N-best generation. Two 5-gram LMs were trained for N-best rescoring. The first one, $\text{LM}_2^5$, was an interpolation of three 5-grams:

$$\text{LM}_2^5 = 0.94 \text{LM}_{full}^5 + 0.059 \text{LM}_{class}^5 + 0.001 \text{LM}_{pos}^5$$

The first one of these was the regular full 5-gram. The second LM component was a class-based 5-gram, based on 1000 word classes that were automatically generated using the SRILM toolkit (each word in the vocabulary was assigned a single word class). The third LM component was another class-based 5-gram, in which the classes were the 33 part-of-speech labels with which the training text had been automatically tagged. Although experiments showed that the two class-based N-grams did not result in significant improvements, we decided to keep them in the system to possibly improve robustness on unseen test sets.

The second 5-gram LM for N-best rescoring, $\text{LM}_2^5$, used context-based Jelinek-Mercer smoothing [8, 6] on the union of all training data counts. This LM did yield a small error rate reduction, albeit at the expense of considerable load time.

4. ACOUSTIC MODELS

4.1. Front End

Two front ends were used in our system, for the purpose of cross-adaptation and system combination. The first one computed 13-order MFCC and spline smoothed pitch [9], plus their first and second order differences, resulting in a feature vector of dimension 42.

The second front end used multilayer perceptrons to compute a feature related to phoneme posteriors [10], to be concatenated with the MFCC and pitch features. The final feature dimension from this front end was 74. (The features used in the second front end form a superset of those in the first front end.)

A clustering algorithm based on the mixture weights of an MFCC-based Gaussian mixture model was used to group all utterances within the same broadcast show into acoustically homogeneous clusters. Based on these pseudo-speaker clusters, vocal tract length normalization and component-wise feature mean and variance normalization were applied.

4.2. Pronunciations

Word segmentation was performed on the transcriptions of the acoustic training data using the 60K lexicon. We developed a Chinese character pronunciation dictionary with multiple pronunciations per character. This single-character dictionary contains around

<table>
<thead>
<tr>
<th>Data</th>
<th>BN</th>
<th>BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) TDT</td>
<td>17.7M</td>
<td></td>
</tr>
<tr>
<td>(2) GALE</td>
<td>3M</td>
<td>2.7M</td>
</tr>
<tr>
<td>(3) Giga-cna</td>
<td>451.4M</td>
<td></td>
</tr>
<tr>
<td>(4) Giga-xin</td>
<td>260.9M</td>
<td></td>
</tr>
<tr>
<td>(5) Giga-zbn</td>
<td>15.8M</td>
<td></td>
</tr>
<tr>
<td>(6) NTU-Web</td>
<td>95.5M</td>
<td>2.1M</td>
</tr>
<tr>
<td>(7) CU-web</td>
<td>96.8M</td>
<td></td>
</tr>
<tr>
<td>dev06</td>
<td>34.1K</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. LM training and development text data, in number of words.
8000 entries, including almost all the possible Chinese characters, with the first pronunciation being the most common pronunciation for a given character. To obtain the pronunciation of a new word (in the 60K lexicon, but not in the 49K lexicon built in [1]), we simply concatenated the most common pronunciations for each character in the word. This allowed us to quickly build the lexicon. The phone set was the BBN-based main-vowel phone set, which included 70 tonal phones plus one silence and one noise phone. The pronunciation of the garbage word was a sequence of at least two noise phones. For the GALE training data, we did not spend time transcribing English words with the Mandarin phone set. Instead, we simply set their pronunciations as a sequence of noise phones. The length of the sequence was made proportional to the length of the spelling. Once the pronunciation dictionary is constructed in this fashion, we could start training acoustic models.

For decoding test data, we adopted a different methodology to obtain pronunciations for English words. As mentioned earlier, there were more than 1700 English words in the decoding lexicon. To be able to recognize these English words, we used simple heuristic rules to create the pronunciations of English words with our Mandarin phone set.

4.3. Training criteria

The AMs used in the final system were all gender-independent, MPFE trained [11] with fMPE feature transforms. For the MFCC-feature front end, there were 3000 decision-tree clustered states with 128 Gaussians per state, denoted as 3000*128. Crossword triphones were used in the MFCC system with feature-space speaker adaptive training (SAT), via single-class constrained MLLR [12].

For the MLP-feature front end, we did not have enough time to train an equally complex system as with the MFCC-feature system. Instead, we trained a 3000*64 within-word triphone model without speaker adaptive training. For a detailed account of our investigation into combining MLP features, fMPE transforms, and MPFE training, please refer to [13].

5. DECODING

Our decoding structure consisted of two iterations of cross-adaptation, as illustrated in Figure 1. In the first iteration, first-pass decoding is performed using the within-word MLP-feature AM with a pruned trigram. The top hypothesis is used to cross adapt the word-SAT MFCC-feature AM. Next, we use the adapted models to re-decode the test data and generate lattices with a pruned bigram, followed by lattice expansion with the full trigram LM. The top hypothesis from the trigram lattice is then used for the second iteration of cross-adaptation, as shown in Figure 1. Finally, we generated 1000-best lists from the trigram lattices in the final stages. The two N-best lists were rescored, respectively, by two 5-gram LMs: LM^1 and LM^2, and then decomposed into character-level N-best lists. The 5-gram scores were then combined with acoustic scores and word insertion penalties to compute posterior probabilities at the character-level via confusion networks. The character string with highest posterior was generated as the final result.

6. EXPERIMENTAL RESULTS

To understand the contribution of LM training data and lexicon, we used a subset of dev06 to compare the perplexities of different LMs. This subset of dev06 did not contain OOV words in either 49K or 60K lexicons and did not contain any garbage words, in order to avoid any confounding effects due to the special nature of the garbage word. This subset had about 29.4K words. Table 4 shows the word perplexity of a few LMs. One way to achieve comparability of the two lexicons occurring in the table is to compare the perplexity or likelihood at the character level.

Table 4. Word perplexity of different LMs on a subset of dev06, which did not contain OOV or the garbage word.

Table: | LM                  | perplexity | log likelihood |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>49K full 4-gram</td>
<td>243.8</td>
<td>-75332</td>
</tr>
<tr>
<td>60K pruned 2-gram</td>
<td>359.5</td>
<td>-79499</td>
</tr>
<tr>
<td>60K pruned 3-gram</td>
<td>228.7</td>
<td>-73390</td>
</tr>
<tr>
<td>60K full 3-gram</td>
<td>193.0</td>
<td>-71094</td>
</tr>
<tr>
<td>60K full 5-gram LM^2</td>
<td>77.9</td>
<td>-58830</td>
</tr>
</tbody>
</table>

Before decoding, input broadcast shows were segmented into short utterances and subjected to pseudo-speaker clustering within a show, consistent with how the training had been processed. Speaker-based vocal track length normalization and utterance-based cepstral-mean and cepstral-variance normalization were then executed.

The first row in Table 5 corresponds to the best system described in [1] on dev04 and eval04. That system used 97 hours of acoustic data and 420M words of LM training data. The second row of Table 5 demonstrates the effect of adding more acoustic data. To do a fair comparison, the same adaptation hypotheses used in Table 2 of [1] were used here for adaptation on the new crossword+SAT MPFE trained model, and the same 4-gram word lattices were used to search for the best word sequence after adaptation. The small improvement could be blamed to the tight search space constrained by the lattices generated by the 49K lexicon.

The studies in [13] showed 8%-15% relative CER reduction using each of the three discriminative techniques (MLP features, MPFE models, and fMPE transforms) alone, and 23% relative...