

PORTING DECIPHER FROM ENGLISH TO MANDARIN

M. Hwang, X. Lei, T. Ng, M. Ostendorf

University of Washington, Seattle, WA
{mhwang,leixin,tng,mo}@ee.washington.edu

A. Stolcke, W. Wang, J. Zheng, V. Gadde

SRI International, Menlo Park, CA
{stolcke,wwang,zj,rao}@speech.sri.com

ABSTRACT

This paper describes our efforts in porting the SRI Decipher English system into Mandarin for transcribing telephone conversations. This includes all aspects of the system: the pronunciation phone set and lexicon, word segmentation, pitch features, discriminatively trained acoustic models with parameter sharing determined by decision trees, and web-data augmented language models.

1. INTRODUCTION

Our work on Mandarin conversational telephone speech (CTS) recognition is based on SRI Decipher system [1] and the SRI language modeling (LM) Toolkit [2]. This is the first complete attempt in porting Decipher to Mandarin. This paper will describe all the work involved.

Mandarin is the official spoken language for modern Chinese. There are many dialects in China, such as Cantonese, Taiwanese, etc. Almost every populated region has its own dialect. Unfortunately these dialects are acoustically so different that they can be treated as completely different languages. However, the writing is the same irrespective of which dialect is spoken. The Qing empire unified Chinese writing more than 2000 years ago until the 20th century when the China government simplified the traditional writing, while some other parts of Chinese community in the world still use the complicated yet more artistic traditional writing. Fortunately it is easy to learn to read the simplified writing after one is familiar with the traditional writing, and vice versa. Therefore, Chinese writing remains to be unified to a major extent. In this paper, 'Chinese' refers to the written form of the language, while 'Mandarin' refers to the spoken form. We focus on Mandarin dialect and simplified writing only. Furthermore, since GB coding is the most popular coding for simplified characters, we deal with Chinese texts in GB coding only, although the work done here should be able to be extended easily to other codings and traditional writing.

We will first summarize the decoding architecture for the 20-times real time (20xRT) system in Section 2, followed in Section 3 by a description of the training and test data involved. Then the work starts in Section 4 defining the

pronunciation set and segmenting Chinese texts into 'word' units. Section 5 describes the front end feature extraction subsystem and all the acoustic models (AMs) involved. Section 6 then turns to the development of language models, including collecting and filtering extra texts from internet. We then present the recognition character error rate (CER) on two test sets in Section 7. Finally, in Section 8, we summarize our efforts and discuss issues for future.

2. DECODING ARCHITECTURE

Figure 1 describes the Mandarin 20xRT system architecture, which is almost identical to the SRI Decipher English system. In a nutshell, the system output is a combination of two different front ends (MFCC and PLP), which involves a few iterations of unsupervised cross adaptation (i.e., MFCC output is used to adapt PLP acoustic models and vice versa).

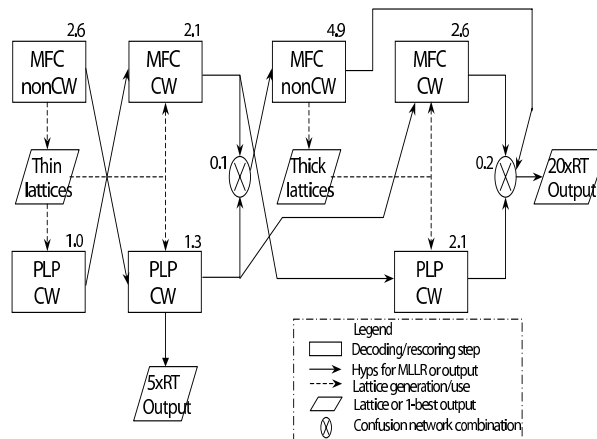


Fig. 1. SRI Decipher 20x real-time decoding architecture.

The MFCC within-word (or non-cross-word, nonCW) triphone acoustic model is used to generate word lattices using a bigram language model. The word lattices are then expanded into more grammar states with trigram scores by a trigram LM. Finally, three N-best lists are generated from the trigram lattices using three different adapted acoustic

models: MFCC nonCW triphones, MFCC CW triphones, and PLP CW triphones. The N-best word lists are then combined to generate a character-based confusion network for Rover, to obtain the final recognition result. For more details of the 20xRT system, please refer to [1]. The main differences of our Mandarin system from the SRI English system include: added pitch features in the front end, no duration modeling, no alternative pronunciations, no SuperARV language modeling, no Gaussian short lists for speeding up the decoding, and neither LDA/HLDA nor voicing features nor ICSI features were used.

3. TRAINING AND DEVELOPMENT DATA

The acoustic training data consisted of two parts: CH&CF and Train04. CH&CF is the Mandarin CallHome and Call-Friend corpora. According to the supplied manual segmentation, there are 45.9 hours of data in CH&CF. These were phone calls from the U.S. (usually one speaker) to mainland China (often more than one speaker) without any specific topic. Since families and friends between two continents tried to convey as much information about their lives as possible within the constrained time period, many speakers talked fast and many conversations involved abundant English words, such as 'yeah', 'okay', 'e-mail', 'Thanksgiving', etc.

Train04 was collected by Hong Kong University of Science and Technology (HKUST) in 2004 and consists of 251 phone calls. These were phone calls within China and Hong Kong by mostly college students, limited to 40 topics such as professional sports on TV, life partners, movies, computer games, etc. There are no multiple speakers on any conversational side. According to the manual segmentation, this consists of 57.7 hours of data. CH&CF and Train04 were used here for both acoustic model training and language model training.

HKUST also collected a 2-hour development set (Dev04, 24 phone calls) and a 1-hour evaluation set (Eval04, 12 phone calls), according to the same protocol as Train04. Due to all of the above differences between CH&CF and Train2004, Dev04 (and Eval04) are a lot easier than CallHome and CallFriend data. As the Train04 protocol will be continued in future data collection efforts, we focused on HKUST data only and reported results and system construction based on Dev04. The final system was then evaluated on Eval04.

4. WORD SEGMENTATION AND PRONUNCIATION PHONE SET

The transcriptions of Train04 are not word segmented. Therefore we used New Mexico State University's word segmenter (NMSU) [3] to segment both Train04 and CH&CF. The

word units then determined the training of both within-word and cross-word triphone acoustic models.

For phonetic pronunciation, we started from BBN's 2003 Mandarin pronunciation dictionary, which was based on the LDC Mandarin pronunciation lexicon. BBN dictionary used 83 tonal phones, in addition to 6 nonspeech phones to model silence and other nonspeech events. We obtained some CER improvement by using a few simple rules to merge rare phones [4]. The resulting phone set had 62 speech phones, plus one silence phone, one for laughter, and one for all other nonspeech events.

5. ACOUSTIC MODELING

5.1. Front end

First of all, all acoustic conversation sides of CH&CF, Train04, and Dev04 were segmented into utterances based on the LDC released segmentation information. For evaluation, the acoustic segmentation was not provided and therefore an automatic segmentation was run using gender-independent Gaussian mixture models (GMMs). Two GMM models were used, each with 100 Gaussians of 39-dimension MFCC cepstra and deltas: a foreground model for speech and a background model for silence. 0.5 seconds of silence were kept at the beginning and at the ending of each utterance segment. Both training and decoding were done at the utterance level. Our first Mandarin auto segmentation GMM model trained on CH&CF and Train04 did not prove to be better than the English model in recognition accuracy. Thus we used the English acoustic segmentation GMM throughout this work.

Similar to the English system, two front end systems were trained: MFCC and PLP, each using 12-order cepstra plus log energy. ESPPS pitch was passed to a lognormal tied mixture model [5] to alleviate pitch halving and doubling problems, then a smoothing algorithm similar to [6] was applied. First-order and second-order differences of all 14 dimensions were computed, yielding a 42 dimensional feature vector per 10ms.

A speaker-based VTL warping factor between 0.88 through 1.12 stepped by 0.04 was selected, by computing the maximum likelihood of a 128-mixture GMM of 12 dimension MFCC cepstra. The warping factor was then applied to the MFCC/PLP frequency bands before cepstra were computed. No word hypothesis was needed. Training the Mandarin VTLN GMM model using both CH&CF and Train04 did not yield better accuracy on one of our earlier systems than the one trained on CH&CF only. Therefore, the CH&CF VTLN GMM model was used. Next cepstral and pitch features were mean and variance-normalized for each speaker. Initial attempts to use HLDA to transform cepstral features with triple delta (13×4) back to 39 dimension did not yield

any gain and thus were not incorporated into the final system.

5.2. Acoustic Model Training

Though gender information is included in Train04, it is not reliable. Furthermore, the size of the training data was small compared with the English system. Therefore, gender independent models with VTLN were used for all the acoustic models. As explained in Section 2, nonCW triphone models were needed to generate word lattices. Both the MFCC and PLP front-end models followed the training procedure illustrated in Figure 2.

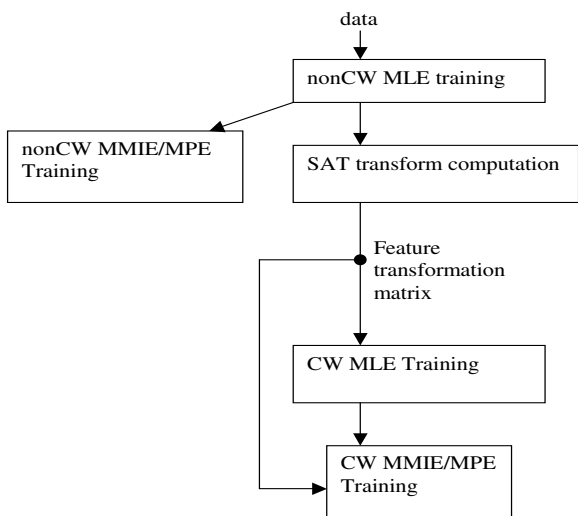


Fig. 2. Flowchart of acoustic model training.

One feature transform per speaker was estimated via 1-class speaker adaptive training (SAT) [7, 8] using the nonCW MLE model. The feature transform was then applied during CW training and testing. The final acoustic models were trained by alternating MMIE [9] and MPE [10, 11] training, which was seeded by maximum likelihood estimated (MLE) models.

All triphone HMMs were 3-state left-to-right Bakis topology and used decision-tree based state clustering. 3500 shared states were empirically determined, with at most 32 Gaussians per state. In building the decision tree for state clustering, we hand crafted 66 *categorical* questions, such as TONE1 which included all the vowel phones with the first tone, in addition to 65 *individual* phone questions. All of these questions could be asked of the left-context phone, the right-context phone, or the central phone. The same triphone in different word positions (beginning, middle, ending phone of a word) were treated identical. Triphones which represent the same toneless base phones were allowed to be clustered together; so were the different HMM state loca-

tions. For example, the first state of an EE1 triphone was allowed to be clustered with the last state of another EE3 triphone.

6. LANGUAGE MODELING

Among the 251 phone calls in Train04, we randomly reserved 25 phone calls as a heldout set to optimize LM perplexity. The rest of the Train04 transcriptions are together denoted as subTrain04 here.

As the lexicon used by NMSU is in an unknown binary format, we could not change the lexicon for word segmentation. After texts were segmented into words by NMSU, we selected the most frequent 11k words in CH&CF + subTrain04 data as our initial decoding lexicon. We then further did a longest-match segmentation on those words, in the training text, that were in NMSU lexicon but not in our initial 11k-word lexicon into a sequence of shorter words that were either in the 11k lexicon or became single-character words. Finally we added all single-character words in the final segmented text into the decoding lexicon, if they were not in initially. Therefore, the only out-of-vocabulary words in the training text were infrequent English words, which were mapped to <UNK> during LM training. The final decoding lexicon had about 11.5k unique words.

6.1. Baseline LM

Since the two training corpora were quite different, we first trained two different trigrams based on CH&CF and subTrain04 and then linearly interpolated them. The interpolation weight was optimized to minimize the word perplexity of the heldout set. As expected, the baseline LM, LM_0 , was heavily weighted on the subTrain04 data, as indicated in Table 1.

At the time of testing our LMs, we had not finalized our AMs yet. Therefore, we used an early set of AMs which were MMIE trained CW models. The decoding architecture was less than 5xRT, as indicated by the '2.6 MFC nonCW' and the '1.3 PLP CW' boxes on the left side of Figure 1. Row 1 of Table 1 indicates the word perplexity on Dev04, whose word segmentation was done by NMSU. It also shows the recognition CER on the acoustically auto-segmented Dev04 set, using the speech/silence GMM models as described in Section 5.1.

6.2. Conversational Web Data

To take advantage over the enormous amount of data on the internet, we selected the top 8800 4-grams from Train04 as queries to the Google search engine. We searched for the exact match to one or more of these n-grams within the text of web pages in GB encoding only. The web pages returned indeed mostly consisted of conversational style phrases such

	LM Weight				PPL	CER
	Sub04	CC	Web _c	Web _t		
LM ₀	0.87	0.13	-	-	269.3	38.8%
LM ₁	0.65	0.05	0.30	-	202.2	36.4%
LM ₂	0.64	0.04	0.16	0.16	192.6	36.1%
LM ₃	0.66	0.05	-	0.29	193.5	36.1%

Table 1. Word perplexity and CER on Dev04. Word segmentation of Dev04 was done by NMSU as all other text data. Dev04 speech was auto segmented here. Sub04 is subTrain04 text. CC is CH&CF text.

as ‘让你觉得不爽’ (made you out of sorts), ‘你也够呛’ (you have had enough), etc.

Some returned web pages were removed, such as duplicated pages and pages with corrupted or illegal GB codes. Text normalization, such as removing html markers, converting Arabic digits into the spoken form, etc., were applied. Sentence boundary was dictated by punctuation. Word segmentation was again achieved by the NMSU segmenter. Finally, pages with high perplexity computed by the baseline LM were filtered such that 60% of the total number of words of the entire retrieved documents were kept.

The conversational web data were then used to train a web trigram, Web_c. Then three way LM interpolation was applied to optimize the word perplexity on the heldout set. As shown in the row of LM₁ of Table 1, Web_c received a significant weight, indicating that our retrieval and filtering algorithms were effective in selecting conversation-like texts. CH&CF received an even smaller weight than in LM₀, re-confirming that CH&CF was quite different from the new collection by HKUST. LM₁ reduced Dev04 word perplexity by 25% relatively, which translated into 6% relative reduction in CER. We shared our filtered web data with both BBN and Cambridge University, who both confirmed the significant contribution of the web data.

6.3. Topic-based Web Data

Since Train04 was limited to 40 topics only, we tried to retrieve 40 topic-dependent collections from the internet, based on 3-word key phrases for each topic. The word sequence $w_1w_2w_3$ is considered as a key phrase for topic t if

$$\frac{C(w_1w_2w_3|t)/\alpha_t}{\sum_{j=1}^{40} C(w_1w_2w_3|j)/\alpha_j} > \theta$$

where $C(w_1w_2w_3|t)$ is the trigram count and α_t is the total number of word tokens, among subTrain04 whose topic is t . θ an empirical constant (we used 0.3 here). Using this criterion helps obtain topic-discriminant key phrases. For example, the above formula yields a small value for phrases common to many topics, such as ‘对对对’ (yes,yes,yes).

For a topic-related phrase, the formula yields a value closer to 1.

Some of the 40 specified topics were rare or not represented in subTrain04 at all. For those topics, we manually designed key phrases based on the brief descriptions that were provided to the subjects as part of the data collection protocol. After the key phrases were defined for all 40 topics, we then queried Google for 40 collections of web pages.

The sizes of the four training corpora are indicated in Table 2. As we have found in English, the web data is easily more than 100 times the in-domain data.

Source	Number of Words
CH&CF	479K
subTrain04	398K
Conversational web data	100M
Topic-based web data	244M

Table 2. Sizes of data sources for LM training.

These 40 collections then went through the same text normalization and data filtering. Finally one trigram, Web_t, was trained using these topic-dependent web data. Then a four-way LM interpolation was applied to obtain the optimized LM for the heldout set. The third row of Table 1 showed the performance of the 4-way interpolated LM on Dev04. There was marginal improvement over LM₁. Since the interpolation weight of the topic based web data was the same as that of the conversational web data, we verified the topic web data alone was as valuable as the conversational web data by doing another three-way interpolation as the last row of Table 1 shows. Its marginal benefit over the conversational web data could be explained by the larger amount of data and by the fact that these data are correlated to the topics in the HKUST collection. LM₂ was used in the final evaluation system.

For further LM investigation using web data, including LM adaptation, please refer to [12].

7. EVALUATION

Table 3 shows the CER of adding a new AM component incrementally. To evaluate effectively all the different components, we ran one-pass decoding, all constrained by the same thick trigram lattices generated by an adapted MFCC nonCW MPE model (see the ‘thick lattices’ box in Figure 1). The trigram scores in the thick lattices were computed by LM₂. Unlike earlier experiments, manually segmented Dev04 was used as the testbed in order to optimize Rover parameters. The time stamps provided by LDC was followed without any silence trimming. Table 3 shows the CERs of both the speaker independent (SI) model and the

unsupervised speaker adapted (SA) model, with the adaptation hypothesis from the other front-end system as indicated in Figure 1.

Acoustic Model	SI	SA
(1) PLP nonCW MLE, no pitch	41.5%	36.4%
(2) PLP nonCW MLE	40.4%	35.5%
(3) PLP CW MLE	39.5%	34.5%
(4) PLP CW SAT MLE	36.8%	34.0%
(5) PLP CW SAT MPE	35.3%	32.9%
(6) MFCC CW SAT MPE	36.2%	33.4%
(7) MFCC nonCW MPE	40.0%	33.6%
Rover (5)+(6)+(7)	–	31.7%

Table 3. CERs on manually segmented Dev04 set using 1-pass decoder, all constrained by the same thick lattices which contained LM₂ trigram scores.

From the table, we observed consistent improvements with pitch features, unsupervised adaptation, CW modeling, SAT feature transformation, and MPE training. Comparing rows (3) and (4), the SAT feature transform became less powerful, though still worthwhile, after adaptation. CW SAT on MFCC front end had hardly any impact on the CER of the top 1 word hypothesis after adaptation, as shown in rows (6) and (7). However, it still offered valuable different N-best hypotheses to be used by Rover. To prepare system combination, 2-pass decoding was run to generate a word-based N-best list using each of the three MPE AMs. The three N-best lists were finally combined to construct a character-based confusion network. System combination gave us 1.2% absolute gain over the best single system.

This final system architecture was then used in our 2004 evaluation on Eval04 data set and achieved 29.7% CER [13]. All three systems in the evaluation were considered to have no significant difference in CER. The CER difference between our system and others' was smaller on Eval04 than that on Dev04, probably because we did not fine tune our system on the Dev04 set.

8. DISCUSSION AND FUTURE WORK

We have ported the SRI Decipher English CTS system into a successful Mandarin CTS system. Our CER on the evaluation clearly indicated that we now have a state-of-the-art Mandarin CTS recognizer. Almost all of the technologies were able to maintain their merits as seen in the English system. The only notable exception is that we have not yet incorporated LDA/HLDA/PCA (principal component analysis) successfully.

System comparison after evaluation indicates two major deficiencies. First of all, as noted by other systems,

the silence of HKUST data at sentence boundaries provided by the LDC manual segmentation is extraordinarily long, which degrades the estimation of the speaker-dependent mean and variance of cepstra and pitch (possibly VTLN, too), as proportionally less speech data is used in the computation. Some degree of silence trimming needs to be done. The same issue needs to be addressed in the acoustic auto segmentation of the test data. 0.5 seconds of silence at utterance beginning and ending seem to be a little bit too long.

Secondly, our word perplexity of 3-gram LMs is significantly higher than those of other systems. Although the word perplexities are not directly comparable because of differences in vocabulary, it may indicate that our word segmentation is inferior. An improved version of the word segmentation algorithm is needed, along with a better lexicon for word segmentation and a new lexicon for decoding.

Additionally, we have not investigated different variants of class-based n-gram LMs. The topic-based LM adaptation in [12] also needs further exploration since no significant improvement was obtained. We also need to train the VTLN GMM model and acoustic auto-segmentation GMM model using Train04 acoustic data. Finally, the model size of the CW AMs may need to be increased compared with nonCW AMs, in order to model more co-articulation effects. We will investigate all of the above aspects in the near future.

Acknowledgments

The authors would like to thank BBN for sharing their 2003 phonetic pronunciation lexicon. This work was supported by the Defense Advanced Research Projects Agency grant MDA 972-02-C0038. The opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government. Portions of this paper are taken from [12].

9. REFERENCES

- [1] A. Stolcke et al., "Development of the SRI/ICSI/UW fall 2004 conversational telephone speech-to-text system," in *DARPA EARS Workshop*, 2004.
- [2] <http://www.speech.sri.com/projects/srilm/>.
- [3] W. Jin, "Chinese segmentation and its disambiguation," in *New Mexico State University, Technical Report MCCS-92-227*, 1992.
- [4] M. Hwang et al., "Progress on Mandarin conversational telephone speech recognition," in *International Symposium on Chinese Spoken Language Processing*, 2004.
- [5] M. Sonmez et al., "A lognormal model of pitch for prosody-based speaker recognition," in *Proc. Eur.*

Conf. Speech Communication Technology, 1997, pp. 1391–1394.

- [6] C.J. Chen et al., “New methods in continuous Mandarin speech recognition,” in *Proc. Eur. Conf. Speech Communication Technology*, 1997, vol. 3, pp. 1543–1546.
- [7] M. Gales, “Maximum likelihood linear transformations for HMM-based speech recognition,” in *Computer Speech and Language*, 1998, vol. 12.
- [8] T. Anastasakos et al., “A compact model for speaker adaptive training,” in *Proc. International Conference on Spoken Language Processing*, 1996.
- [9] P. Woodland et al., “Large scale discriminative training for speech recognition,” in *IEEE ASR*, 2000, pp. 7–16.
- [10] D. Povey et al., “MMI-MAP and MPE-MAP for acoustic model adaptation,” in *Proc. Eur. Conf. Speech Communication Technology*, 2003.
- [11] J. Zheng et al., “Explorations in discriminative training,” in *DARPA EARS RT04 workshop*, 2004.
- [12] T. Ng et al., “Web-data augmented language models for Mandarin conversational speech recognition,” submitted to ICASSP 2005.
- [13] http://www.nist.gov/speech/tests/rt/rt2004/fall/rt04f-evaldiscdoc/rt04f_official_results_20041022/stt_man_cts.htm.