

PROGRESS ON MANDARIN CONVERSATIONAL TELEPHONE SPEECH RECOGNITION

*Mei-Yuh Hwang, Xin Lei, Tim Ng, Ivan Bulyko, Mari Ostendorf^d
Andreas Stolcke, Wen Wang, Jing Zheng, Venkata Ramana Rao Gadde, Martin Graciarena²
Man-Hung Siu³, Yan Huang⁴*

mhwang@ee.washington.edu

1. University of Washington 2. SRI International

3. Hong Kong University of Science and Technology 4. International Computer Science Institute

ABSTRACT

Over the past decade, there has been good progress on English conversational telephone speech (CTS) recognition, built on the Switchboard and Fisher corpora. In this paper, we present our efforts on extending language-independent technologies into Mandarin CTS, as well as addressing language-dependent issues such as tone. We will show the impact of each of the following factors: (a) simplified Mandarin phone set, (b) pitch features, (c) auto-retrieved web texts for augmenting n-gram training, (d) speaker adaptive training, (e) maximum mutual information estimation, (f) decision-tree-based parameter sharing, (g) cross-word co-articulation modeling, and (h) combining MFCC and PLP decoding outputs using confusion networks. We have reduced the Chinese character error rate (CER) of the BBN-2003 development test set from 53.8% to 46.8% after (a)+(b)+(c)+(f)+(g) are combined. Further reduction in CER is anticipated after integrating all improvements.

1. INTRODUCTION

In recent years, much effort has been put into Mandarin Chinese large vocabulary continuous automatic speech recognition (ASR). On one hand, it is natural to want to expand ASR to one of the world's most spoken languages. Other motivations are language-specific: from the end-user's point of view, speech input would be particularly useful as a means to side-step keyboard input using the Chinese writing system. From a researcher's point of view, Mandarin presents unique challenges, such as the use of tones and the lack of word segmentation.

In this paper we detail recent developments in our Mandarin CTS recognition system. Our efforts had two main thrusts: porting technology that was successful for English CTS to the Mandarin CTS task (e.g. improved training and decoding techniques), and modifications to system components aimed at better representing tone in Mandarin (e.g. phone set improvements and adding pitch

features). In the following sections, we describe the baseline system used in this work, followed by descriptions for each improvement. We close with a summary of the performance gains and a discussion of potential future improvements.

2. BASELINE SYSTEM

Our baseline system started from SRI's 2003 Mandarin prototype, which used 39 dimension MFCC features plus its derivatives. It used the LDC released Mandarin CallHome dictionary phone set as its basis for pronunciations and for acoustic hidden Markov model (HMM) training. The 35 hours of Mandarin CallHome (CH) and CallFriend (CF) corpora were used for both acoustic model (AM) and language model (LM) training. Acoustic models were gender-independent maximum likelihood estimated (MLE) models that used within-word triphones and bottom-up genone clustering [1]. During training, mean and variance normalization, along with vocal tract length (VTL) normalization [2], were applied per speaker, using the speaker identities already provided in the training data release. During testing, since there may be multiple speakers on some conversation sides, we used a Gaussian mixture model to automatically cluster each conversation side into two speakers [3]. Mean/variance/ VTL normalization and MLLR adaptation [4] were then applied on a per-speaker basis using the automatically determined clusters. Two passes of MLLR unsupervised adaptation were applied: first a phone-loop adaptation [5] was applied; the adapted model was then used in first-pass decoding to generate hypotheses. Second, these hypotheses were used for the second-pass MLLR adaptation. Three MLLR transforms were used, specified based on phone identity. Final recognition was then performed using this adapted model.

The test set, denoted here as dev03, was created by BBN in 2003 from the CF corpus for development testing. It contains 12 conversations (phone calls), amounting to 55 minutes of speech which is manually segmented into

1,516 utterances. Note this set is excluded from our training data. Our 2003 system selected the most frequent 11k words from CH+CF corpora as the decoding lexicon. A 3-gram LM was trained on CH+CF transcripts. The recognition system achieved 56.7% CER on this dev03 set, using NIST scoring tool which allows optionally deletable regions [6]. Note that all of the CER numbers reported in this paper are on this dev03 test set.

To improve the 2003 system, we first rebuilt our training pronunciation dictionary by merging the LDC released CallHome dictionary, the SRI-ICSI Mandarin Broadcast News (BN) dictionary, and the BBN Mandarin CTS dictionary. In initial experiments, the 86 BBN tonal phones are mapped to SRI's 40 toneless phones. Similar to the 2003 system, we trained an MFCC, gender-independent, within-word triphone model based on genome clustering. However, we tuned the system performance by cutting the number of Gaussians in half from the 2003 system.

For word n-gram LM training, we manually added 1k Mandarin words and 1k English words, most of which are proper names of people and locations. Then we added Mandarin Broadcast News (BN) language model training texts into n-gram LM training. Both bigram and trigram language models were trained using modified Kneser-Ney smoothing [7]. This improved system achieved a CER of 53.8% on the dev03 set and is serving as the baseline for the work reported here.

The rest of the paper will describe our experimental results using techniques listed in the abstract, namely, phone set, pitch feature, web texts, speaker-adaptive training (SAT), maximum mutual information estimation (MMIE), decision tree (DT) based parameter sharing, and cross word triphones. Our ultimate goal is to combine all seven techniques into two subsystems based on the MFCC and PLP front-ends and then perform system combination using confusion network-based ROVER. All experiments reported here used gender-independent models.

3. PRONUNCIATION PHONE SET

Since Mandarin is a tonal language, we added tones into the phone set after the lexicon had been fixed. We first used BBN's tonal pronunciation and achieved 53.3% CER on dev03, without pitch features.

As the BBN phone set contains some rare phones, we decided to replace the rare phones with a single common phone or a sequence of frequent phones to make them more trainable. The mapping table is shown below:

- II, I, i \rightarrow i (知 Z i)
- AA, a \rightarrow a (丹 d a n)

- w o \rightarrow o (波 b o)
- w OO \rightarrow w o (握)
- U \rightarrow W u (玉)
- E n \rightarrow a n (安)
- E y \rightarrow ey (杯 b ey)
- yi[nN] \rightarrow i[nN] (因, 应)

For example, 知 /Z II/ and 兹 /z I/ are now /Z i/ and /z i/ respectively. Notice that phone names are case sensitive. Vowels usually have 5 tones, but tone labels are omitted here, assuming readers understand that each rule applies to all tones.

During genome clustering, we allowed parameter sharing across different tonal phones of the same toneless base phone. Different state locations (state 1 vs. state 2 of a triphone HMM) were also allowed to share parameters. Thus with this reduced set of 62 phones, we partitioned the sharing space into 33 classes. In addition, all systems reported in this paper had 3 additional classes for the three phones which modeled nonspeech sounds: silence, noises, and laughter. The MFCC model trained using the reduced phone set produced a CER of 52.0%, or 1.3% better than using the original phone set.

4. ADDING PITCH FEATURES

After the phone set and pronunciation lexicon were fixed, we then added pitch features. The ESPS pitch track [8] was passed to SRI's robust pitch filter, which uses a log-normal tied mixture model [9] to avoid halving and doubling of the pitch period, followed by a median filter for smoothing. Log pitch values were taken for the voiced sections, and IBM-style smoothing [10] was applied for the unvoiced sections. First and second order differences of pitch were computed from the smoothed sequence. The mean and variance of the three-dimensional pitch feature were normalized per speaker, similar to the MFCC features. Variance normalization is particularly important since the pitch feature values (especially its deltas) are often in a different range than the MFCC feature values, which would have caused the Gaussian density value in a different order had the features not been normalized. The added pitch features brought another 2.5% absolute gain on CER, reducing it to 49.5%.

5. WEB-AUGMENTED LM

Previously in our English CTS system, we showed that we could use keyword queries to retrieve relevant texts from internet web pages to improve language modeling [11]. We extended this work to our Mandarin CTS system. We chose the most frequent n-grams from CH and CF acoustic data transcripts and passed them to the Google search engine to retrieve web pages in GB encoding.

Some sample n-grams are: 对对对 (yes yes yes), 哦我知道 (ah, I know), 八零零 (800).

We then trained a new n-gram language model according to the following steps:

1. Remove HTML markers and email addresses.
2. Insert sentence boundary markers based on punctuation.
3. Convert GB digits and English letters into ASCII. Then convert digits into spoken forms.
4. Segment Chinese characters into word units, using the segmentation tool in [12].
5. Remove ~40% of the documents that have the highest CTS perplexity. This left about 84M words of training texts, which were used to train a web-based n-gram language model.
6. The web-based LM was interpolated with the baseline LM (trained on CH+CH+BN) with an interpolation weight of 0.2, estimated using a held-out subset of CF.

The web-based interpolated n-gram LMs reduced the CER rate by another 0.5%, to 49.0%.

6. SPEAKER ADAPTIVE TRAINING

SAT training has been found to be effective in reducing the variance in a speaker-independent model and thus making the model more discriminative [13]. In our implementation, we used 1-class constrained MLLR [4] with a full-matrix transformation to simulate SAT training. Since there is only one transformation, we can apply the inverse transformation in the feature domain. In this way SAT parameters are estimated via the current MLLR code and applied during both training and decoding in the feature domain. Combining SAT and the web-enhanced LM, we achieved a CER of 47.6% on dev03.

7. MMIE TRAINING

Recently maximum mutual information estimation (MMIE) training has been found to outperform MLE training, particularly for English CTS. Following [14], we trained one generalized MMIE iteration on top of an existing MLE model. Lattices used for training denominator counts were generated with the same MLE model using an unigram language model with a tree-organized dictionary. The same H -ratio tuned on SRI's English CTS system was applied to the Mandarin CTS system for parameter estimation. Our first MMIE model did not incorporate SAT transforms. The MMIE-trained, within-word, non-SAT model with the web enhanced trigram language model reduced the CER rate to 47.4%, compared to 49.0% with MLE models.

8. DECISION-TREE BASED PARAMETER SHARING

Decision-tree based parameter sharing [15] should be superior to bottom-up state clustering, primarily due to its prediction on unseen contexts. In our initial efforts to replace the bottom-up clustered genone models with top-down decision-tree based senone models, we manually created 48 linguistic categorical questions to be applied to each of the 36 subspaces of parameters (all triphone-states of the same toneless base phone formed one subspace). In a preliminary experiment on the within-word, non-SAT, MLE system, it achieved a 0.2% CER reduction: the error rate was 48.8% compared to 49.0% with bottom-up clustering. Without further tuning, we next moved to cross-word modeling, since we expected the modeling of unseen contexts to be more important when cross-word triphones are considered.

9. MODELING CROSS-WORD CO-ARTICULATION

All of the experiments so far used within-word triphones. To better model co-articulation across word boundaries, we trained cross-word (CW) triphone models. Again, 36 parameter sharing subspaces were used. A third pass MLLR adaptation was applied to the CW model based on the best hypothesis generated from the within-word model. The adapted within-word model was also used to generate a word lattice to constrain the search space for the CW model to speed up decoding. With 48 categorical questions in decision-tree clustering, the CER after MLE CW model decoding was 46.9%.

Since our decision-tree clustering did not utilize composite questions, we next added 18 additional categorical questions intended to simulate some composite questions. The additional questions reduced the CER by another 0.1% to 46.8%. Note that we had not yet incorporated SAT or MMIE training in this result.

10. N-BEST ROVER

It is usually beneficial to combine the N-best outputs of several different systems based on word posterior probabilities represented as confusion networks [16]. In a preliminary attempt, we combined the N-best lists from two MFCC-based and one PLP-based systems, all using MLE within-word triphone models. One MFCC system (49.0% CER) used MFCC+pitch; the second MFCC system (48.7% CER) used a slightly different smoothing technique for the pitch of unvoiced sections and more MLLR iterations during adaptation. The PLP+pitch

system had a CER of 50.5%. The total score of each the N-best hypotheses was defined as

$$\begin{aligned} \text{total score} = & (\text{acoustic score}) \\ & + (\underline{\text{language weight}}) * (\text{language score}) \\ & + (\underline{\text{word insertion penalty}}) * (\# \text{ hypothesis words}) \end{aligned}$$

and then normalized to obtain posterior probabilities. The two underlined parameters were optimized to minimize CER on a held-out set. Finally the posteriors from the three systems were interpolated linearly with yet another set of specified weights.

With uniform interpolation weights, the combination of the three systems yielded a CER of 47.2%, a 1.5% improvement over the single best system. To understand the upper bound on accuracy we optimized the language weight and word insertion penalty for each system based on the correct transcription of dev03, obtaining a CER of 46.8%.

11. DISCUSSION

All results are summarized in Table 1. Without using system combination, the improvement from the updated 53.8% baseline to the best-case result of 46.8% corresponded to a 13% relative reduction in CER.

System	CER
2004 Baseline	53.8%
BBN tonal phone set	53.3%
(a. Reduced tonal phone set)	52.0%
(a)+(b. pitch)	49.5%
(a)+(b)+(c. Web LM)	49.0%
(a)+(b)+(c)+ SAT	47.6%
(a)+(b)+(c)+ MMIE	47.4%
(a)+(b)+(c)+DT	48.8%
(a)+(b)+(c)+DT+CW	46.8%
ROVER among PLP and MFCC front ends with (a)+(b)+(c)	47.2%

Table 1: Progress on the character error rate on Mandarin CTS dev03 set.

Recently LDC released an additional 58 hours of training data and a 2.48-hour development set, dev04. With the added training data, we achieved 31.8% CER on dev04 after we combined all the components, which was competitive with other state-of-the-art speech recognition systems.

In the near future, we plan to integrate linear discriminative analysis (HLDA or LDA) with pitch features. Furthermore, we need to reduce the out-of-vocabulary rate of the decoding lexicon, which may improve the contribution of web data on the LM. We would also like to port other LM improvements from the

English CTS system. For example, we plan to explore questions such as whether we can import techniques for modeling disfluencies in English to Mandarin, as well as model Mandarin-specific aspects of conversational speech, such as the abundant interjection words.

ACKNOWLEDGMENTS

The authors would like to thank BBN for sharing their Mandarin pronunciation dictionary and the dev03 data set. This work was supported by DARPA under contract MDA972-02-C-0038. This work is approved for public release, distribution unlimited.

REFERENCES

- [1] V. Digalakis and H. Murveit, GENONES: Optimizing the degree of mixture tying in a large vocabulary hidden Markov model based speech recognizer, *Proc. IEEE ICASSP*, vol. 1, pp. 537-540, 1994.
- [2] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, Speaker Normalization on Conversational Telephone Speech, *Proc. IEEE ICASSP*, vol. 1, pp. 339-341, 1996.
- [3] A. Sankar, F. Beaufays, and V. Digalakis, Training Data Clustering for Improved Speech Recognition, *Proc. Eurospeech*, vol. 1, pp. 503-506, 1995.
- [4] M. J. F. Gales, Maximum Likelihood Linear Transformations for HMM-based Speech Recognition, *Computer Speech and Language*, vol. 12, pp. 75-98, 1998.
- [5] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, The SRI March 2000 Hub-5 Conversational Speech Transcription System, *Proc. NIST Speech Transcription Workshop*, 2000.
- [6] <http://www.nist.gov/speech/tests/rt/rt2004/fall/>. Click "RT Fall 2004 Evaluation Plan".
- [7] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling, *Computer Speech and Language*, vol. 13, No. 4, pp. 359-394, 1999.
- [8] ESPS Version 5.0 Programs Manual, Entropic Research Laboratory, Aug. 1993.
- [9] M. K. Sönmez, L. Heck, M. Weintraub, E. Shriberg, A Lognormal Model of Pitch for Prosody-based Speaker Recognition, *Proc. Eurospeech*, vol. 3, pp. 1391-1394, 1997.
- [10] C.J. Chen, R.A. Gopinath, M.D. Monkowski, M.A. Picheny and K. Shen, New methods in continuous Mandarin speech recognition, *Proc. Eurospeech*, vol. 3, pp. 1543-1546, 1997.
- [11] I. Bulyko, M. Ostendorf and A. Stolcke, Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures, *Proc. HLT/NAACL*, pp. 7-9, 2003.
- [12] W. Jin, A Case Study: Chinese Segmentation and its Disambiguation, Computing Research Laboratory, New Mexico State University, Las Cruces, Technical Report MCCA-92-227, 1992.
- [13] T. Anastasakos, J. McDonough, and R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training", *Proc. ICSLP*, vol. 2, pp. 1137-1140, 1996.
- [14] J. Zheng, J. Butzberger, H. Franco, and A. Stolcke, Improved Maximum Mutual Information Estimation Training of Continuous Density HMMs, *Proc. Eurospeech*, vol. 2, pp. 679-682, 2001.
- [15] M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting Unseen Triphones with Senones", *IEEE Trans. on Speech and Audio Processing*, vol. 4, No. 6, pp. 412-419, 1996.
- [16] L. Mangu and E. Brill and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks", *Computer Speech and Language*, vol. 14, No. 4, pp. 373-400, 2000.