Recent Improvements of the RWTH GALE Mandarin LVCSR System

Ch. Plahl[1], B. Hoffmeister[1], M.-Y. Hwang[2], D. Lu[3], G. Heigold[1], J. Lööf[1], R. Schlüter[1], H. Ney[1]

[1]Lehrstuhl für Informatik 6 - Computer Science Department
RWTH Aachen University, Aachen, Germany
[2] Microsoft Research, Redmond, WA, USA
[3] Computer Science Department, Southwest Forestry University, Kunming, China
{plahl,hoffmeister,heigold,loof,schlueter,ney}@cs.rwth-aachen.de
mehwang@microsoft.com

Abstract

This paper describes the current improvements of the RWTH Mandarin LVCSR system. We introduce a new reduced toneme set developed at RWTH. We are using different toneme sets and pronunciation lexica. For the purpose of discriminative training we will show a fast way to transform word lattices between systems using different toneme sets and pronunciation lexica. In addition to various acoustic front-ends, the current systems use different kinds of neural network toneme posterior features.

While different kinds of systems are developed, a two stage decoding framework for combining these systems is applied. We show detailed recognition results of the development cycle of the systems. Finally, two methods to integrate tonal features are compared.

Index Terms: Mandarin speech recognition, LVCSR, system combination, multiple feature streams

1. Introduction

Within the GALE project, we have developed a automatic speech recognizer for continuous Mandarin speech, handling broadcast news (BN) and broadcast conversations (BC). This paper summarizes the further development and improvement of our work presented in [1]. The final system is competitive to current Mandarin speech recognizers, [2, 3, 4].

We start by introducing two different toneme sets and dictionaries used. The LC-STAR toneme set (denoted as RWTH−83) and the corresponding pronunciation lexicon are taken from the preceding system, [1]. In addition, we use a toneme set and pronunciation lexicon based on the lexicon of the University of Washington (UW). Section 3 describes the acoustic models based on a MFCC and a PLP front-end in combination with two neural network posterior features [5, 6]. While the computationally most expensive part in discriminative training is the generation of word lattices for a huge amount of data, we present a fast and simple way for porting word lattices for discriminative training across different toneme sets.

In Section 4 we describe the training and testing corpora followed by the experimental sections. We commence by comparing the new toneme sets and an investigation to include tonal features in a single LDA, in Section 5.1 and 5.2 resp. Finally, Section 6 describes the definite decoding framework used for the GALE 2007 re-evaluation. The system consists of two decoding runs joined by a cross-adaptation step. We present detailed character error rates (CER) for the decoding process followed by the final result, competitive to other state-of-the-art decoders.

2. Pronunciation Dictionary and Language Model

The RWTH Mandarin LVCSR system follows a common approach for Mandarin LVCSR systems and uses word-based toneme pronunciation models [2, 4, 3].

We present two different pronunciation models, both follow the main-vowel principle as described in [7]. The first toneme set (RWTH−83) is the one used in [1], a subset of SAMPA-C [8] which contains 14 vowels (equivalently 55 tonal vowels), 26 consonants, one silence and one garbage phone. The main source for this pronunciation dictionary is the LC-Star Mandarin lexicon [9].

The second toneme set (RWTH-71) is an improved version of the first toneme set. There are three different glides in Mandarin, `/y, w, v/`. In RWTH−83 glides are tonal, making triphones fragmented. For example, syllable `dui4` in Pinyin is pronounced as `d w4 e4 y4`. It splits `/w[e4]y/` triphone into 16 possibilities, causing the Markov state clustering extra difficulty. Furthermore, RWTH-83 contains three additional toneless glides for those Mandarin syllables beginning with glides, such as `yan2` → `/y y2 a2 n/`, `wan2` → `/w w2 a2 n/` and `yuan2` → `/ v v2 a2 n/`. These extra glides make the HMM duration unnecessarily longer for that syllable, and limit the triphone context length.

In light of this flaw, we design RWTH-71 following [2]. Starting from the UW 72-phone set in Table 4 of [2], we add tonal diphthongs `/ey/` and `/ay/`, get rid of `Y`, merge tonal `a` and `A`, but keep tonal `/IH,I,i/` separate. Next, we replace tonal `yu` with tonal `v i` to get rid of `yu1−4`, resulting in RWTH-71.

Table 1: changes in the toneme set

(a) pronunciations			(b) syllables with `v`-glide		
dui4	→	d w ey4	jiong3	→	j v o3 NG
yan2	→	y a2 n	qiong2	→	q v o2 NG
wan2	→	w a2 n	xiong1	→	x v o1 NG
yuan	→	v a2 n	yong3	→	v o3 NG

The pronunciations of the above mentioned syllables are listed in Table 1.a. Finally, in RWTH-71 we use `v`-glide (instead of `y`-glide, as most Mandarin systems do) for the four syllables in Table 1.b.

The two language models (LMs) used in this work were kindly provided by UW and SRI. Both LMs share the same 60K vocabulary. The first 4-gram LM (LM.v1), used in all recognizers, is the same *pruned* LM as the one used in the GALE 2007 summer evaluation. The second 4-gram LM (LM.v2) is an improved version of LM.v1, using more data and no pruning. LM.v2 is used in lattice rescoring.

3. Acoustic Modelling

Similar to the systems presented in [10] and [1], the subsystems differ only in their acoustic front-ends, and the toneme set, the pronunciation dictionary, resp. The toneme set and the pronunciation dictionary are described in Section 2. The final system in the GALE 2007 re-evaluation consists of two subsystems labelled s1 and s2. While s1 is trained using RWTH−83, s2 is based on RWTH−71. The acoustic training is performed independently for each of the subsystems.

3.1. Acoustic Features

The acoustic front-ends of the (sub-)systems consist of MFCCs or PLPs as base features. In addition, a voicedness feature [11] is augmented to the PLP feature extraction of s2 while s1 consists of MFCCs only. The features are normalized by segment-wise mean and variance normalization and are fed into a sliding window of length nine. All feature vectors within the sliding window are concatenated and projected to a 45 dimensional feature space using a linear discriminative analysis (LDA).

In addition, a tonal feature and its first and second derivatives, represented by the first and second order regression coefficients, are augmented to the LDA-transformed baseline features. Tonal information is crucial for Mandarin ASR systems, because tonal patterns play an important role in distinguishing tonemes and words in the Mandarin language. The tonal feature used is described in [12].

For the experiments in Section 5.2, the setup is slightly different. Instead of augmenting the LDA-transformed baseline features to the tonal features, a common LDA for both feature streams is used, following [13]. In this case no derivatives of the tonal features are used.

Finally, the features of s1 and s2 each are concatenated with toneme posterior features produced by a neural network trained on a 1200h subset of the training corpus. S1 uses hierarchical MRASTA (HMRASTA) features, produced by a hierarchical neural network with multiple time resolution features (MRASTA) [5] as input. We use a hierarchy of three nets to produce the HMRASTA-features following [14]. The first and second net uses the higher and lower frequency parts of the MRASTA features, combined with PLP features in the last net. At the end, the toneme posterior features are transformed by a logarithm and reduced by a principal component analysis (PCA) to 51 dimensions. Overall, concatenation of all features leads to a feature dimension of 99 for s1.

In contrast to s1, s2 uses neural network features based on TANDEM, [15], and hidden activation temporal patterns phoneme posteriors (HATs) described in [6, 2]. Finally, the TANDEM and HAT features are combined using the Dempster-Shafer [16] algorithm, transformed by a logarithm and reduced by a PCA. Overall, s2 uses 80 feature components.

3.2. Acoustic Training

The acoustic models for all systems are based on triphones with cross-word context, modelled by a 3-state left-to-right hidden Markov model (HMM). A decision tree based state tying is applied resulting in a total of 4500 generalized triphone states. The acoustic models consist of Gaussian mixture distributions with a globally pooled diagonal covariance matrix. Both maximum likelihood (ML) and discriminative training are applied.

The filterbanks underlying the MFCC and PLP feature extraction undergo a vocal tract length normalization (VTLN). The warping factor classifier is trained beforehand on the complete training corpus. For the training with VTLN on 230 hours no new classifier was trained.

In order to compensate for speaker variations we have used constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR). While s1 uses the standard approach, for s2 a modified version of the SAT/CMLLR training is applied. The speaker adaptive training is combined with the LDA-transformation step, resulting in speaker-specific dimension reducing feature transformation matrices as introduced in [17]. In addition, during recognition, MLLR is applied to the means of the acoustic models.

Minimum phone error (MPE) [18] is applied to refine the ML trained acoustic models. For the MPE training of the two different systems we generate word-conditioned word lattices using the SAT/CMLLR model of s1 in combination with a bigram language model. System dependent alignments are produced for the accumulation and are kept fixed during the training iterations. The optimal number of training iterations is determined by recognition on the development corpus.

In order to save computation time, we use a fast way to convert the generated word lattices of s1 to the new toneme set of s2. Since the two toneme sets share the same lexicon we do not need to convert these words or to build up words from the pronunciations matching a second lexicon. In order to cope with alternative pronunciations we have to combine them first and split them afterwards. Due to that a weight factor is introduced to allow alternative pronunciations in the converted word lattices. After mapping the different words to all pronunciations provided by the pronunciation dictionary of s2, the weight factors are chosen uniformly w.r.t. the scores. The pronunciations are realigned afterwards and the word boundary times are kept fixed. During the whole procedure. Finally, the transformed word lattices are used for discriminative training of s2. As shown in Section 6, the new word lattices work very well.

4. Corpora

1534h of broadcast news (BN) and broadcast conversations (BC) of speech data collected by LDC are used for training. The corpus includes data from the Hub4 and TDT4 corpora and from the first three years of the GALE project (releases P1R1-4, P2R1-2, P3R1).

For the development cycle of the system, a 230h subset of the corpus has been created. The subset contains the HUB4 corpus (30h), 100h of BN and 100h of BC from the four releases of the first year of the GALE project. Table 2 gives detailed statistics for the corpora used.

Table 2: Acoustic data for training and testing

	Training Data		Testing Data	
	230h	1534h	eval06	dev07
total data	230h	1534h	2.2h	2.55h
# segments	206K	1.3M	1301	1985
# running words	2.2M	15.5M	22K	28K
# distinct words	40K	63K	5.3K	5.3K

For the final systems we use the GALE 2007 development corpus (dev07) for tuning and the GALE 2006 evaluation corpus (eval06) for testing. As shown in Table 2, the eval06 corpus contains 2.2 hours of BN and BC, while dev07 contains 2.55 hours. The two corpora used are manually segmented and provided by LDC. However, in the development cycle of the system, the segmentations provided by UW, labelled eval06.v1 and dev07.v1, are used. In addition, the training transcripts were pre-processed by UW-SRI as described in [19].

Table 3: Improvements of the RWTH Mandarin LVCSR System using different toneme sets and pronunciation lexica.

	CER[%]					
	dev07.v1			eval07.v1		
Toneme Set	VTLN	SAT/CMLLR	LM-rescore	VTLN	SAT/CMLLR	LM-rescore
RWTH-83	21.1	19.5	19.1	24.9	23.0	22.6
RWTH-71	20.7	19.0	18.5	24.3	22.4	21.7

Table 4: different combination of acoustic feature streams

		CER[%]					
		dev07.v1			eval07.v1		
Phoneme Set	Integration	VTLN	SAT/CMLLR	LM-rescore	VTLN	SAT/CMLLR	LM-rescore
mfcc + tone	concatenated	17.3	15.5	14.6	24.4	21.7	20.6
	common LDA	16.9	15.4	14.5	24.3	21.6	20.5
plp + tone	concatenated	17.4	15.6	14.6	24.2	21.9	21.0
	common LDA	17.0	15.6	14.6	24.4	21.9	21.0

5. System Development

5.1. Development of the Toneme Set

In this section we present the results concerning the improvements introduced by RWTH-71 in contrast to the old toneme set RWTH-83.

The acoustic models based on MFCC features are trained on the 230 hour subcorpus mentioned in Section 4. The recognition is performed on the two acoustic segmentations eval06.v1 and dev07.v1 of the evaluation and development corpus. The recognition is divided into three passes, starting with VTLN-warped features as the first pass. More information of the 3-pass recognition setup is given in Section 6.1 and Figure 1. The VTLN warping factor classifier is trained beforehand on the 1500h corpus. As the second step, SAT/CMLLR adaptation is applied, followed by a lattice rescoring with LM.v1.

Table 3 summarizes the improvements from the new toneme set and pronunciation lexicon. The toneme set RWTH-71 leads to an absolute reduction of the character error rates (CER) of about 0.4%-0.6% for dev07.v1 and more than 0.6% for eval06.v1. Overall, the relative improvement is up to 2%-4% for all three passes and corpora.

5.2. Acoustic Feature Combination

Our standard approach to integrate additional acoustic feature sets is to augment these features to the LDA-transformed base features. As an alternative, described in Section 3.1, we feed all features into a common LDA estimation. The two approaches are labelled ”concatenated” and ”common LDA” in Table 4.

The common LDA approach performs well for the first pass on dev07 and improves the MFCC and PLP system up to 0.4% absolute, but eval06 is improved only slightly. Furthermore, the decrease for all other steps is slightly less. Nevertheless, a common LDA results in a small improvement, but has to be investigated further.

6. Evaluation System

In this section, the final system used in the GALE 2007 re-evaluation is presented. The final system consists of two subsystems labelled s1 and s2, trained on the complete training corpus. The detailed acoustic front-ends used are introduced in Section 3.

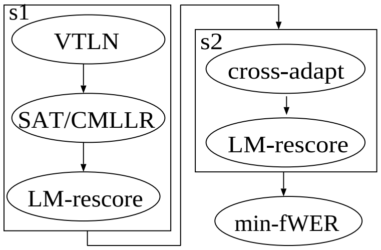

Figure 1: Two stage cross-adaptation: 3-pass stage for a single system followed by a 2-pass cross-adaptation stage

6.1. Decoding Architecture

Similar to [1], the decoding framework is divided into two main stages, starting with a multipass recognition stage. The first two passes are realized by a 4-gram Viterbi decoder, while the third pass uses lattice based LM rescoring. Figure 1 shows the complete decoding process for the final system.

While the first pass uses the ML model with VTLN normalisation, the SAT/CMLLR recognition is performed by the MPE trained model. The adaptation statistics for this step are collected from the previous recognition result. For VTLN normalisation, we estimate a classifier on the complete training corpus. Finally, the word lattices produced in the last recognition step are rescored with the full LM.v2. Experimental results on the tune and development sets are given in Table 5.

Table 5: recognition results for first decoding stage

		CER[%]		
corpus	system	VTLN	SAT/CMLLR	LM-rescore
dev07	s1	15.8	12.4	11.7
	s2	12.9	10.8	10.5
eval06	s1	22.0	19.4	18.6
	s2	19.4	16.6	16.1

Overall, the three passes of the first decoding stage result in an error reduction of more than 20% relative for the test corpora, compared to the VTLN baseline. Detailed CERs for each pass are listed in Table 5.

The second stage of the decoding pipeline is divided into 2 passes. The first pass consists of cross-adaptation which provides a simple and effective way to combine systems [20]. In particular, it allows to benefit from systems that show a significantly higher WER or CER than the target system.

As shown in Table 6, s2 clearly outperforms s1. The difference between these systems is more than 1% absolute for dev07 and about 2% for eval06. Overall, the best benefit is reached by cross adapting s2 with the output of s1, denoted by s1 → s2. As a last step, we applied the min.fWER decoding method [21]. As shown in Table 6, the CER decreases by 0.5% absolute for eval06 and 0.7% for dev07.

Table 6: Experimental results of the second decoding stage.

system	CER[%]	
	dev07	eval06
s1	11.7	18.6
s2	10.5	16.1
s1 →s2	9.8	15.6

7. Conclusion and Further Work

Recent improvements for the current RWTH LVCSR system for Mandarin are presented. An new toneme set, RWTH-71, is introduced, which decreases the character error rate by about 3% relative. Furthermore, we have presented a fast and simple technique to transform word lattices of another toneme set without repeating the complete word lattice generation by a full recognition. These word lattices have been used in the discriminative training of the acoustic models. In addition, we have compared two approaches to combine multiple feature streams. While concatenation is the simplest approach, a common LDA slightly improves the system. Research on this area will be continued at RWTH.

Finally the Mandarin system used in the GALE 2007 reevaluation is presented, consisting of two subsystems, differing in the acoustic baseline features. Two different neural network posterior features are used and cross-adaptation has been performed.

In order to further improve the RWTH Mandarin system, currently new methods for system and acoustic feature combination are investigated. Furthermore, we are planning to integrate new discriminative training criteria in the development cycle of the RWTH Mandarin system.

8. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA).

We want to thank University of Washington and SRI International for all the support and ICSI and IDIAP for providing the neural network toneme posterior features.

9. References

[1] B. Hoffmeister et. al., “Development of the 2007 RWTH mandarin LVCSR system,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Kyoto, Japan, Dec. 2007, pp. 455–460.

[2] M.-Y. Hwang et. al., “Building a highly accurate mandarin speech recognizer,” in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Kyoto, Japan, Dec. 2007, pp. 490–495.

[3] T. Ng et. al., “Progress in the BBN mandarin speech to text system,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, Apr. 2008, pp. 1537–1540.

[4] S. M. Chu et. al., “Recent advantages in the GALE mandarin transcription system,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, Apr. 2008, pp. 4329–4333.

[5] H. Hermansky and P. Fousek, “Multi-resolution RASTA filtering for TANDEM-based ASR,” in *Proc. European Conf. on Speech Communication and Technology*, Lisbon, Portugal, Sept. 2005, pp. 361–364.

[6] B. Chen et. al., “Learning long-term temporal features in LVCSR using neural networks,” in *Proc. Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, Oct. 2004.

[7] C. J. Chen et. al., “Recognize tone languages using pitch information on the main vowel of each syllable,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, May 2001, vol. 1, pp. 61–64.

[8] X. Chen et. al., “An application of SAMPA-C for standard Chinese,” in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 3147–3150.

[9] F. de Vriend, N. Castell, J. Gimnez, and G. Maltese, “LC-STAR: XML-coded phonetic lexica and bilingual corpora for speech-to-speech translation,” in *Proc. of Papillon 2004, Workshop on Multilingual Lexical Databases*, Grenoble, France, Aug. 2004.

[10] J. Lööf et. al., “The RWTH 2007 TC-STAR evaluation system for european English and Spanish,” in *Proc. Int. Conf. on Speech Communication and Technology*, Antwerp, Belgium, Aug. 2007, pp. 2145–2148.

[11] A. Zolnay, R. Schlüter, and H. Ney, “Robust speech recognition using a voiced-unvoiced feature,” in *Proc. Int. Conf. on Spoken Language Processing*, Denver, CO, USA, Sept. 2002, vol. 2, pp. 1065–1068.

[12] X. Lei et.al., “Improved tone modeling for Mandarin broadcast news speech recognition,” in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, Pennsylvania, USA, Sept. 2006, pp. 1237–1240.

[13] R. Schlüter, A. Zolnay, and H. Ney, “Feature combination using linear discriminant analysis and its pitfalls,” in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, USA, Sept. 2006, pp. 345–348.

[14] F. Valente and H. Hermansky, “Hierarchical and parallel processing of modulation spectrum for ASR applications,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, Apr. 2008, pp. 4168–4171.

[15] H. Hermansky, D.P.W. Ellis, and S. Sharma, “Tandem connectionist feature stream extraction for conventional hmm systems,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000, pp. 1635–1638.

[16] F. Valente and H. Hermansky, “Combination of acoustic classifiers based on dempster-shafer theory of evidence,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, Apr. 2007.

[17] J. Lööf et. al., “Efficient estimation of speaker-specific projecting feature transform,” in *Proc. Int. Conf. on Speech Communication and Technology*, Antwerp, Belgium, Aug. 2007, pp. 1557–1560.

[18] D. Povey and P. C. Woodland, “Minimum phone error and i-smoothing for improved discriminative training,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, May 2002, vol. 1, pp. 105–108.

[19] A. Venkataraman et. al., “An efficient repair procedure for quick transcriptions,” in *Proc. Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, Oct. 2004.

[20] D. Guiliani and F. Brugnara, “Acoustic model adaptation with multiple supervisions,” in *Proc.* TC-STAR *Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 151–154.

[21] B. Hoffmeister, T. Klein, R. Schlüter, and H. Ney, “Frame based system combination and a comparison with weighted ROVER and CNC,” in *Proc. Int. Conf. on Spoken Language Processing*, Pittsburgh, PA, USA, Sept. 2006.