Development of the GALE 2008 Mandarin LVCSR System

C. Plahl, B. Hoffmeister, G. Heigold, J. Lööf, R. Schlüter, H. Ney

Lehrstuhl für Informatik 6 - Computer Science Department
RWTH Aachen University, Aachen, Germany
{plahl,hoffmeister,heigold,loof,schlueter,ney}@cs.rwth-aachen.de

Abstract

This paper describes the current improvements of the RWTH Mandarin LVCSR system. We introduce vocal tract length normalization for the Gammatone features and present comparable results for Gammatone based feature extraction and classical feature extraction. In order to benefit from the huge amount of data of 1600h available in the GALE project we have trained the acoustic models up to 8M Gaussians. We present detailed character error rates for the different number of Gaussians.

Different kinds of systems are developed and a two stage decoding framework is applied, which uses cross-adaptation and a subsequent lattice-based system combination. In addition to various acoustic front-ends, these systems use different kinds of neural network toneme posterior features. We present detailed recognition results of the development cycle and the different acoustic front-ends of the systems. Finally, we compare the ultimate evaluation system to our last years system and can report a 10% relative improvement.

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

1. Introduction

Within the GALE project, we build a highly accurate automatic speech recognizer for continuous Mandarin speech, handling broadcast news (BN) and broadcast conversations (BC). This paper summarizes the current developments and improvements for the GALE 2008 evaluation and continues the work of [1]. We add a new type of neural network based toneme posterior features and train three systems using different acoustic front-ends. Two of them are used in the final evaluation system. The decoding framework includes a cross-adaption followed by a lattice-based system combination, which gives a further improvement. The system shows to be competitive to current Mandarin speech recognizers [2, 3, 4].

Section 2 introduces the pronunciation lexicon and Section 3 describes the acoustic models based on a MFCC, and a PLP, and a Gammatone (GT) front-end in combination with two different kinds of neural network posterior features [5, 6]. In Section 4 we describe the training and testing corpora followed by the development of the evaluation system in Section 5. We introduce vocal tract length normalization (VTLN) for Gammatone features and report detailed character error rates (CER) for models with increased parameter size.

Finally, Section 6 describes the definite decoding framework used for the GALE 2008 evaluation. The system consists of two decoding runs joined by a cross-adaptation and a system combination step. Overall, we present CERs for most of the decoding processes followed by a comparison of the last two evaluation systems, competitive to other state-of-the-art decoders.

2. Pronunciation Dictionary and Language Model

The RWTH Mandarin LVCSR system follows the common approach for state-of-the-art Mandarin LVCSR systems [2, 3, 4] and uses a word-based pronunciation dictionary. The dictionary maps words to phoneme sequences, whereas the phoneme carries the tone information, usually referred to as a toneme.

The design of the pronunciation dictionary follows the main-vowel principle as described in [7]. The toneme set (RWTH-71) is an improved version of the toneme set used in the last GALE evaluation and is based on SAMPA-C [1]. We derived RWTH-71 from University of Washington's (UW) 72-phone set by adding several tonal diphthongs like `/ey/` and using `v`-glide for some syllables. In addition, we merged tonals like `a` and `A`, but keep tonal `/IH,I,i/` separate. Finally, we end up with RWTH-71, containing 69 tonemes, augmented by a garbage phone and a silence model.

The language model (LM) used in this work was kindly provided by UW and SRI. The vocabulary size of the LM is 60K. The full 4-gram LM is used in lattice rescoring only, while a pruned version is applied in all other recognition steps.

3. Acoustic Modelling

Similar to the systems presented in [1, 8], the subsystems differ only in their acoustic front-ends. The final system built for the GALE 2008 evaluation consists of two subsystems labelled s1 and s2. S1 is based on MFCCs and s2 is based on PLPs. In addition, after the evaluation we started to train a third subsystem labelled s3 using the GT features.

The acoustic training is performed independently for each of the three subsystems.

3.1. Acoustic Features

The acoustic front-ends of the systems consist of MFCCs, PLPs, and GTs as base features. The GT features are extracted by auditory filter banks realized by Gammatone filters [9]. Vocal tract length normalization (VTLN) for the GT features is described in Section 5.1.

The features are normalized by segment-wise mean and variance normalization and concatenated with a tonal feature. 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 detail in [10]. The concatenated feature streams are fed into a sliding window of length nine frames and all feature vectors within the sliding window are concatenated and projected to a 45 dimensional feature space using a linear discriminant analysis (LDA). Using a common LDA to integrate the tonal features continues the work started in [1, 11].

3.2. Toneme Posterior Features

The feature streams of s1 and s2 are concatenated with toneme posterior features which are produced by a neural network (NN). The NNs are trained on a 1500h subset of the acoustic training corpus. S1 uses hierarchical multiple time resolution (HMRASTA) features. An augmented hierarchical NN processing procedure with MRASTA features [5] as input used to produce the features. We use a hierarchy of two nets to produce the final posterior features, following [12, 13]. While the first net receives the fast modulation frequencies as input, the second net uses the slow modulation frequencies, tonal information and PLP features as additional information to the net. A detailed description of the NN features can be found in [13]. Afterwards, the toneme posterior features are transformed by a logarithm and reduced by a principal component analysis (PCA) to 35 dimensions. Overall, concatenation of all features leads to a feature dimension of 80 for s1.

In contrast to s1, s2 uses neural network features build by parallel processing of the 1500h of training data. The neural network features are based on TANDEM, [14], and hidden activation temporal patterns phoneme posteriors (HATs) described in [6, 2]. The feature generation consists of several steps. The first step calculates the TANDEM phoneme posterior and in the next step, a separate 2-stage layer NN is trained to identify the posterior probabilities, followed by the combination of different critical bands to estimate the HATs, [6]. Finally, the TANDEM and HAT features are combined using the Dempster-Shafer algorithm [15], transformed by a logarithm and reduced by a PCA. Overall, s2 uses 77 feature components to train the acoustic model.

Compared to the last Evaluation the NN features used for s2 are trained on 200h additional hours, while the structure of the NN features for s1 completely changed.

3.3. 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 filter banks 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, estimated by a grid search in the range of 0.8 - 1.2. In order to compensate for speaker variations we use constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR). In addition, during recognition, MLLR is applied to the means of the acoustic models.

Modified Minimum Phone Error (MPE) is applied to refine the ML trained acoustic models [16]. For the modified MPE training of the two different systems we generate word-conditioned word lattices using the SAT/CMLLR model of s2 in combination with a unigram language model. System dependent alignments are produced for the accumulation and are kept fixed during all training iterations. The optimal number of training iterations is determined by recognition on the development corpus.

4. Corpora

1600h of BN and BC of speech data collected by LDC are used for training. The corpus includes data from all years of the GALE project (releases P1R1-4, P2R1-2, P3R1-2, P4R1).

Table 1: Acoustic data for training and testing

	Train set	Test set		
		dev07	dev08	eval07-seq
total data	1600h	2.55h	1.0h	1.63h
# segments	1.3M	1985	619	1013
# running words	16.5M	28K	11K	17K
# distinct words	63K	5.3K	3K	4.1K

For the final systems, we use the GALE 2007 development corpus (dev07) for tuning and the GALE 2008 development and the sequestered data of the GALE 2007 evaluation (eval07-seq) for testing. As shown in Table 1, the development and test data sum up to 5h of BN and BC. The GALE 2006 evaluation corpus contains 2.2h of audio data and is used for the GT-VTLN experiments only. The corpora used are manually segmented and provided by LDC. The training transcripts are pre-processed by UW-SRI as described in [17].

5. System Development

5.1. VTLN for Gammatone

Gammatone features [9] have been shown to be competitive to standard features like MFCCs or PLPs. The Gammatone features are extracted by auditory filter banks realized by Gammatone filters.

The filter is defined in the time domain by the following impulse response:

$$h(t) = k \cdot t^{n-1} \exp(-2\pi \cdot B \cdot t) \cdot \cos(2\pi \cdot f_c \cdot t + \phi).$$

Here, k defines the output gain, B defines the bandwidth, n is the order of the filter, f_c is the filter's center frequency, and ϕ the phase.

The center frequencies of the filters are distributed over the frequency range according to the Greenwood function with human parameters [18]

$$\rho(x) = A \cdot (10^{a \cdot x} - k)\text{Hz}$$

where $A = 165.4$, and $a = 0.88$, and $k = 2.1$.

In order to transfer the VTLN to Gammatone features, the center frequencies of the filters have to be scaled explicitly, here, similar to the standard case, piecewise-linear. After scaling the frequency space using the Greenwood function, the center frequencies are equally spaced. In order to get the warped center frequencies, the following steps have to be applied:

- transform the minimum (100 Hz) and the maximum (7500 Hz) center frequencies to obtain the minimum and maximum warped greenwood scales:

$$\begin{bmatrix} x_{min} \\ x_{max} \end{bmatrix} = \rho^{-1} \left(\begin{bmatrix} f_{min} \\ f_{max} \end{bmatrix} \right),$$

  where x_{min}, x_{max} are the values of the warped frequencies f_{min} and f_{max}.

- transform the greenwood values back to the non-warped frequency axis,

$$f_{center} = w^{-1}(\rho(x))$$

where f_{center} defines the center frequencies of the Gammatone filters in *Hz*, $x \epsilon [x_{min}, x_{max}]$, and w the warping function.

Table 2: Results with and without VTLN warping for MFCC, PLP and GT, trained on a subcorpus of 230h.

system	CER[%]			
	dev07		eval06	
	unwarped	warped	unwarped	warped
s1 (MFCC)	17.8	17.3	24.9	24.4
s2 (PLP)	17.8	17.4	24.9	24.5
s3 (GT)	17.8	17.5	24.9	24.7

Unlike VTLN for MFCCs or PLPs, VTLN for GT features is carried out in the time domain, not in the frequency domain. Due to that, VTLN for Gammatone is done by warping the center frequencies of the Gammatone filters, instead of a redistribution in the frequency domain. As shown in Table 2, all systems perform equal to each other, achieving a character error rate (CER) of 17.8% on dev07 and 24.9% on eval06. When VTLN is applied to the GT features, an improvement of about 0.3% absolute for dev07 and 0.2% for eval06 is achieved. Even though the reduction is not as high as for the MFCCs or PLPs, applying VTLN for GT decreases the CER.

5.2. Increasing the Number of Gaussians

In order to tap the full potential of the big amount of data used to train the speech recognition subsystems we increase the number of Gaussians used in the acoustic model. Most state-of-the-art speech recognition systems use up to 1M Gaussian parameters, as in the presented evaluation system.

Since we use a globally pooled diagonal covariance matrix for our mixture models we have to use a large number of Gaussians. This is necessary to cope with variations in the data which cannot be modelled by the single covariance matrix. We have trained our acoustic models with up to 8M Gaussians parameters, resulting in 1K Gaussians for each of the triphone states. Using an 80 dimensional feature vector this means we have to train up to 640M free parameters. So far, the models are maximum likelihood estimates, discriminative training has not been performed yet.

Table 3 reports the results using different number of Gaussians. When estimating more than 1M Gaussians, the error rate is reduced by 0.2% for each step the number of Gaussians is doubled. Overall, a reduction of 0.5% after rescoring is achieved when 8M Gaussian parameters are used instead of 1M. Nevertheless, in order to cope with the huge number of Gaussians, more than 4GB of memory is needed to train the GMMs.

6. Evaluation System

In this section, the final system built for the GALE 2008 evaluation is presented. The 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. A complete discriminative training of s3, using the GT features, could not be finished.

Table 3: First decoding stage recognition results for dev07 of subsystems s1 of the GALE 2008 evaluation system using different numbers of Gaussians. No discriminative training has been performed.

# of Gaussians	dev07 (CER[%])		
	pass1	pass2	pass3
4.5K	17.8	16.3	15.9
9K	16.8	15.1	14.7
18K	15.6	13.9	13.6
36K	14.8	13.1	12.7
70K	13.9	12.6	12.1
140K	13.3	11.8	11.5
270K	12.8	11.4	11.0
510K	12.3	10.9	10.5
1M	11.9	10.6	10.3
2M	11.6	10.6	10.2
4M	11.3	10.4	10.0
8M	11.1	10.2	9.8

6.1. Decoding Architecture

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

Moreover, the first pass uses the ML model with VTLN normalization, the SAT/CMLLR recognition is performed by the modified MPE trained model. The adaptation statistics for this step are collected from the previous recognition result. For VTLN normalization, we estimate a classifier on the complete training corpus. Finally, the word lattices produced in the last recognition step are rescored with the full 4-gram LM. Experimental results for the GALE 2007 re-evaluation and the GALE 2008 evaluation on the tune and testing sets are given in Table 4. A detailed description of the GALE 2007 re-evaluation system can be found in [1]. While s2 outperforms s1 in the GALE 2007 re-evaluation, the subsystems for the GALE 2008 evaluation perform equally well.

Table 4: Final recognition results for first decoding stage for the two subsystems s1 and s2 of the GALE 2007 re-evaluation and GALE 2008 evaluation system, including discriminative training.

system	CER[%]		
	dev07	dev08	eval07-seq
s1 (2007)	12.9	11.6	14.1
s1 (2008)	9.6	9.2	11.0
s2 (2007)	10.0	9.5	11.0
s2 (2008)	9.9	9.3	11.0

The second stage of the decoding pipeline is divided into 2 passes followed by a system combination. The first pass consists of cross-adaptation which provides a simple and effective way to combine systems [19]. In particular, it allows to benefit from systems that show a significantly higher WER or CER

than the target system. Adapting s1 to the out s2 will be written as s2 $\rightarrow$ s1. Since s2 outperforms s1 in the GALE 2007 re-evaluation, the improvement for s1 by adapting to s2 could not result in an overall improvement. The systems for the GALE 2008 evaluation perform equally well and also after adaptation the final results are comparable, Table 5. The single system could be improved by about 5-6% relative for all three corpora. In addition, the final combined evaluation systems improved by 0.3% leads to an overall improvement of 5% relative for eval07-seq and about 10% for dev07 and dev08 compared to the last GALE 2007 re-evaluation system.

Table 5: Final recognition results of the for second decoding stage for the two subsystems s1 and s2 of the GALE 2007 re-evaluation and GALE 2008 evaluation system.

		CER[%]		
eval	system	dev07	dev08	eval07-seq
2007	s2 $\rightarrow$s1	11.1	10.4	12.3
	s1 $\rightarrow$s2	9.8	9.4	10.9
	final (min.fWER)	9.8	9.4	10.9
2008	s2 $\rightarrow$s1	9.1	8.6	10.7
	s1 $\rightarrow$s2	9.3	8.7	10.6
	final (min.fWER)	8.8	8.5	10.4

7. Conclusion and Further Work

Recent improvements of the current RWTH LVCSR system for Mandarin were presented. We have introduced vocal tract length normalization for the Gammatone filter based feature extraction, which decreases the character error rate by about 2% relative. Each of the filter's center frequency can be modelled independently: exploiting this observation, we expect to gain further improvements in VTLN for Gammatone features.

Furthermore, we gave detailed results for increasing the number of Gaussians to 8M. Doubling the Gaussians starting from 1M leads each time to an improvement of 0.2% to 0.3%.

Finally, the Mandarin system used in the GALE 2008 evaluation was presented, consisting of two subsystems that differ in their base features and their toneme posterior features. In order to have comparable results for the GT features, we will perform a complete training of the GT based system. Two different kinds of neural network based toneme posterior features are used, the HMRASTA and the TANDEM/HAT features. The features are important for a low error rate and for this year's evaluation both features performed well resulting in two subsystems of comparable performance. In discriminative training we applied the recently developed modified-MPE criterion.

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] C. Plahl et. al., "Recent improvements of the RWTH GALE mandarin LVCSR system," in *Proc. Int. Conf. on Speech Communication and Technology*, Australia, Aug. 2008, pp. 2426–2429.

[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] 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.

[4] 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.

[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, Sep. 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*, vol. 1, Salt Lake City, USA, May 2001, pp. 61–64.

[8] 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.

[9] R. Schlüter et. al., "Gammatone features and feature combination for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, Honolulu, HI, USA, Apr. 2007, pp. 649–652.

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

[11] 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, Sep. 2006, pp. 345–348.

[12] 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.

[13] F. Valente et. al., "Hierarchical processing of the modulation spectrum for GALE LVCSR systems," in *submitted to INTERSPEECH*, 2009.

[14] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2000, pp. 1635–1638.

[15] 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.

[16] G. Heigold, R. Schlüter, and H. Ney, "Modified mpe/mmi in a transducer-based framework," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, Apr. 2009.

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

[18] D. D. Greenwood, "A cochlear frequency-position function for several species–29 years later," *Acoustical Society of America Journal*, vol. 87, pp. 2592–2605, Jun. 1990.

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