DEVELOPMENT OF THE 2007 RWTH MANDARIN LVCSR SYSTEM

Björn Hoffmeister, Christian Plahl, Peter Fritz, Georg Heigold, Jonas Lööf, Ralf Schlüter, Hermann Ney

Human Language and Pattern Recognition
Computer Science Department 6, RWTH Aachen University, Germany
{hoffmeister,plahl,heigold,loof,schluter,ney}@cs.rwth-aachen.de

ABSTRACT

This paper describes the development of the RWTH Mandarin LVCSR system. Different acoustic front-ends together with multiple system cross-adaptation are used in a two stage decoding framework. We describe the system in detail and present systematic recognition results. Especially, we compare a variety of approaches for cross-adapting to multiple systems.

During the development we did a comparative study on different methods for integrating tone and phoneme posterior features. Furthermore, we apply lattice based consensus decoding and system combination methods. In these methods, the effect of minimizing character instead of word errors is compared. The final system obtains a character error rate of 17.7% on the GALE 2006 evaluation data.

Index Terms— Mandarin speech recognition, system combination, multiple feature streams

1. INTRODUCTION

In the course of the GALE project, an Arabic and a Mandarin LVCSR system have been set up at RWTH. This paper describes in detail the development of the Mandarin system; a description of the Arabic system can be found in [1].

We start by introducing the phoneme set and the pronunciation dictionary, which are based on SAMPA-C and the LC-Star Mandarin pronunciation lexicon [2, 3]. The final system uses four different acoustic front-ends: MFCCs, PLPs, gammatone cepstral coefficients [4], and neural network (NN) based phoneme posterior features [5]. In addition, a single tonal feature is used [6]. Section 3 describes the acoustic front-ends, the tonal feature, and the four acoustic models used in the final system. It is followed by Section 4 describing the training and testing corpora.

In Section 5 we show results from the development cycle of the final system. In particular, we present results on the integration of the tonal and the NN features with the MFCC features. In the literature, several ways to combine multiple feature streams are proposed [7, 8, 9]. We compare the approaches and motivate the integration method used for the final system. Also, we study lattice based consensus decoding and system combination methods for Mandarin ASR. The standard evaluation metric for Mandarin is the character error rate (CER), whereas the common consensus decoding and combination methods minimize the word error rate (WER). We present comparative results for both decision rules.

Section 6 describes our decoding framework. The final decoder consists of two stages: two parallel 5-pass decoding paths followed by a 2-pass cross-adaptation path. The choice of the cross-adaptation method is based on the comparison of a variety of approaches for cross-adapting to multiple systems. We present error rates for the GALE 2006 evaluation and 2007 development corpora. Finally, we give a conclusion and an outline of our future work.

2. PRONUNCIATION DICTIONARY AND LANGUAGE MODEL

The RWTH Mandarin LVCSR system follows the common approach for Mandarin LVCSR systems and use a phoneme based pronunciation model [7, 10, 11]. The phoneme set is a subset of SAMPA-C [2] and contains 14 vowels and 26 consonants. Tone information is included following the main-vowel principle described in [12]. We merge tone 3 and 5 for all vowels [13]. Furthermore, for the phoneme @' we merge tone 1 and 2, because we have only very few observations for tone 1. The resulting phoneme set consists of 81 tonemes, an additional garbage phone, and silence.

The main source for our pronunciation dictionary is the LC-Star Mandarin lexicon [3]. It defines a mapping from Pinyin to SAMPA-C and it contains pronunciations for 96K words. In addition, we use the word to Pinyin mappings from the Hub5/4 Mandarin lexicon (LDC96L15 and LDC97E7) and from the CEDICT dictionary. Noise, hesitations, laughter, and unintelligible words are mapped to the garbage word, which is modeled by the garbage phone. Pronunciations for words not contained in any of the lexica are produced as follows: we segment the unknown word into a sequence of known words by applying a longest-match segmenter and concatenate the pronunciations of the segmentation result.

All the language models (LMs) used in this work are kindly provided by the University of Washington (UW) and SRI. The decoding vocabulary is determined by the language model. For all recognition experiments we use a pruned 4-gram derived from the full LM; the full LM is only applied in word graph rescoring. For the final system we use two LMs sharing the same 60K vocabulary: a 5-gram (LM.v1) and an improved 4-gram (LM.v2) [14].

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), or SRI.

3. ACOUSTIC MODELING

The final system consists of four independent subsystems labeled s1-s4. They differ in their acoustic front-ends similar to the systems described in [15]. Acoustic training is performed independently for each of the systems.

3.1. Acoustic Features

The acoustic front-end of System s1 consists of MFCC features. The features are normalized by segment-wise mean and variance normalization and are fed into a sliding window of length nine. All feature vectors in the window are concatenated and projected to a 45 dimensional feature space by applying a linear discriminative analysis (LDA). Finally, a tonal feature and its first and second derivatives are augmented to the feature vector. Tonal information is crucial for Mandarin ASR systems, because tonal patterns play an important role in distinguishing phonemes and words in the Mandarin language. The tonal feature used is described in [6].

System s2 and s3 are equal to s1 beside the base features: instead of MFCCs, s2 uses PLPs and s3 uses gammatone cepstral coefficients. The gammatone features are described in [4] and were shown to be competitive to standard features like the MFCCs or PLPs.

System s4 starts with the same acoustic front-end as s1, but the features are augmented with phoneme posterior features produced by a neural network. The input of the net are multiple time resolution features (MRASTA) [5]. The output layer corresponds to the phonemes in the phoneme set used. It is trained on a phoneme alignment which is produced by System s1. We use a single net with one hidden layer. The dimension of the phoneme posterior features is reduced by a principal component analysis (PCA) to 24 yielding an overall feature dimension of 72.

3.2. Acoustic Training

The acoustic models for all systems are based on triphones with cross-word context. They are modeled by a 3-state left-to-right hidden Markov model (HMM). A decision tree based state tying is applied resulting in a total of 4,500 generalized triphone states. We use Gaussian mixture distributions with a globally pooled diagonal covariance matrix. Due to the rigor variance modeling our systems require many Gaussians and we use up to 2M Gaussians for maximum likelihood (ML) trained acoustic models. For computational reasons, we restrict the number to about 1M for discriminative training.

The filterbanks of the MFCC and PLP feature extraction are normalized by applying a 2-pass vocal tract length normalization (VTLN). The warping factors are estimated by a grid search in the range of 0.8 - 1.2. For the gammatone cepstral coefficients in s3 we do not apply VTLN.

We compensate for speaker variations by using constrained maximum likelihood linear regression speaker adaptive training (SAT/CMLLR) for all systems. Additionally, in recognition MLLR is applied to the means of the acoustic models.

Discriminative training is performed to refine the ML trained acoustic models. In order to improve the models we use the discriminative minimum phone error (MPE) criterion [16]. For the MPE training of the different systems we generate word-conditioned word graphs with the s4 SAT/CMLLR system in combination with a bigram language model. System dependent alignments within the word boundaries are produced for the accumulation. These alignments are kept fixed during the training iterations. The optimal number of training iterations is determined by recognition on the development corpus.

Table 1. Acoustic data for training and testing

	Training Data		
	440h	870h	1230h
total data	465h	872h	1230h
# segments	500K	850K	1.1M
# running words	4.6M	8.9M	12.6M
# distinct words	50K	49K	51K

	Development and Testing Data			
	dev04	eval04	eval06	dev07
total data	0.48h	0.96h	2.15h	2.35h
# segments	283	560	1302	1701
# running words	4.8K	9.2K	22K	28K
# distinct words	1.8K	2.8K	5.3K	5.3K

4. CORPORA DESCRIPTION

We use data from the Hub4 and TDT4 corpora which only contain broadcast news (BN), whereas the corpora taken from the GALE project contain a mix of broadcast news and broadcast conversations (BC). We take the speech data from the quarterly releases of the first year and the first two releases of the second year (P1R-4, P2R1-2) to set up the training material. The 440 hours consist of the Hub4 corpus, the TDT4 corpus, and the GALE releases P1R1 and P1R2. The Hub4, TDT4, and P1R1-4 data yield the 872 hours of speech. In order to build the last corpus the 358 hours from P2R1 and P2R2 are added to the 872 hours. Table 1 gives detailed statistics for the corpora used. The silence ratio is about 13% for all the corpora.

For the final system we use the GALE 2006 evaluation corpus (eval06) as tune set and the GALE 2007 development corpus (dev07) for testing. As shown in Table 1 the eval06 corpus contains 2.15 hours of BN and BC data and the dev07 corpus 2.35 hours.

During the development cycle we use a second tune and test set: the RT04 development (dev04) and evaluation (eval04) corpora containing 0.5 hours and 1.0 hours of Mandarin broadcast news. In contrast to dev06 and eval07, the RT04 corpora include a large amount of English words: 3% for dev04 and 0.6% for eval04. Since our decoding vocabularies contain almost no English words, all the English words in the recognition corpora directly cause insertion errors.

The training transcripts are pre-processed by UW-SRI as described in [17]. As well, UW supplies the acoustic segmentation for all corpora. For eval06 and dev07 they provide two segmentations, seg.v1 and an improved seg.v2 [14].

Table 2. Progress of the RWTH Mandarin LVCSR System. Log-linear integration of tone and NN features for 30h and 100h, integration via concatenation for 440h and 870h.

Setup	CER[%] dev04	CER[%] eval04
30h (Hub4)	8.5	19.2
100h (Hub4 + 70h from TDT4)	8.5	19.0
440h (Hub4 + TDT4 + GALE P1R1-2)	8.0	17.5
870h (Hub4 + TDT4 + GALE P1R1-4)	8.5	16.7

5. DEVELOPMENT OF THE SYSTEM

In this section we give an overview of the methods we use during the development cycle of the final system. We start with an MFCC system trained on the 30 hours of Hub4 data. The training data is later augmented by 70 hours from the TDT4 corpus. Both setups use 107 phonemes and a 49K decoding vocabulary. We train a first neural net on the 100 hours and produce 42 dimensional NN based phoneme posterior features for both systems. With these systems we give a comparative study on different approaches for integrating tone and NN features, see 5.1.

For the 440 hours system the phoneme set is reduced to 81 phonemes and we switch to the LM.v1 with a 60K decoding vocabulary. On the 440 hours we train a new MFCC model and the neural network as described in Section 3.

The final system is based on the 870 hours setup. From this setup we train the four Systems s1-s4. The NN features used by s4 are produced by the net trained on the 440 hours. Using this setup, we compare word graph and character graph based consensus decoding and system combination, see 5.2.

Recently, we got an additional 358 hours of acoustic training data. We use them in order to re-train the s4 MPE model by performing another ten iterations of MPE training on the complete 1230 hours.

Table 2 shows recognition results for the different setups. All presented results are obtained with MFCC based models and VTLN.

5.1. Acoustic Feature Combination

All our subsystems use two or three feature streams: cepstral features (MFCCs, PLPs, or gammatones), the tone feature, and optionally the NN based phoneme posterior features. The literature contains several ways to combine these feature streams. The most simple approach is to concatenate the individual feature vectors [7, 11]. An alternative is to feed the feature streams into a single LDA [8, 10]. The third approach investigated is to perform the integration in a log-linear model [9]. A similar method is proposed in [18].

Table 3 summarizes the results for the different integration methods. In the case of concatenation and log-linear model combination, the three feature streams are the LDA transformed MFCCs, the tonal feature and its first and second derivatives, and the PCA transformed NN features. In the LDA approach we estimate a single LDA on the MFCC feature vector augmented with the tonal feature.

Table 3. Integration of tone and NN based phoneme posterior features with MFCC features.

Features	Integration	CER[%] dev04	CER[%] eval04
30h training (Hub4)			
MFCC + tone	concatenated	9.3	20.5
MFCC + tone	log-linear	8.7	20.0
MFCC + tone	LDA	9.3	20.0
MFCC + tone + NN	log-linear	8.5	19.2
100h training (Hub4 + 70h from TDT4)			
MFCC + tone	concatenated	8.3	19.5
MFCC + tone	log-linear	8.7	19.2
MFCC + tone + NN	log-linear	8.5	19.0
870h training (Hub4 + TDT4 + GALE P1R1-4)			
MFCC + tone	concatenated	8.3	16.8
MFCC + tone	log-linear	8.2	16.9
MFCC + tone + NN	concatenated	8.5	16.7
MFCC + tone + NN	log-linear	8.0	16.7

For the concatenation and the LDA approach we train a single acoustic model from the resulting features. In contrast, for the log-linear model combination we train separate models for each feature stream. In the latter case, the single models are estimated from the same fixed alignment. We optimize the log-linear weights for the individual models on the development set by word graph rescoring. The initial alignment and the word graphs are obtained from a system trained on concatenated features.

For fewer data, the log-linear model combination gives some nice improvement over the simple concatenation approach. But with more training data the benefit declines and for the 870 hours setup we see no improvement at all.

5.2. Consensus Decoding And System Combination

The common evaluation metric for Mandarin ASR tasks is the character error rate (CER), instead of the word error rate (WER) used e.g. for European or Arabic languages. The RWTH Mandarin recognizer produces word sequences and word graphs. For Viterbi decoding the best word and character sequence are equal, because the minimized cost function is the sentence error. On the other hand, a confusion network or a minimum time frame word error rate (min.fWER) [19] lattice-decoder minimizes the WER instead of the desired CER. The same holds for system combination techniques like ROVER [20] or min.fWER lattice-combination [21]. In order to minimize CER we have to split the arcs of the word graphs into characters first.

We compare the results of consensus decoding and system combination on word and on character graphs. The investigated methods are min.fWER based consensus decoding, min.fWER combination, and ROVER with confidence scores.

The min.fWER approaches as well as our confidence score calculation require word or character graphs with boundary times. The word graphs produced do contain word boundary times, but character boundaries have to be computed in a post-processing step. We test two strategies for assigning start and

Table 4. Consensus decoding and system combination results. For character graphs with "correct times" the character time boundaries are derived from a forced alignment.

		word graphs [CER%]		char. graphs [CER%]		char. graphs [CER%]	
				(correct times)		(approx. times)	
System(s)		eval06	dev07	eval06	dev07	eval06	dev07
s1	Viterbi	22.0	19.4	22.0	19.4	22.0	19.4
	min.fWER	21.8	19.3	21.8	19.2	21.8	19.2
s2	Viterbi	22.3	19.6	22.3	19.6	22.3	19.6
	min.fWER	22.1	19.5	22.2	19.5	22.2	19.5
s3	Viterbi	21.5	19.0	21.5	19.0	21.5	19.0
	min.fWER	21.3	18.9	21.4	18.8	21.5	18.8
s1+s2+S3	ROVER	20.2	17.9	20.0	17.7	20.1	17.7
	min.fWER	20.1	17.5	20.0	17.5	20.1	17.5
s4(MPE+360h,	Viterbi	17.8	14.5	17.8	14.5	17.8	14.5
cross-adapted)	min.fWER	17.7	14.3	17.8	14.4	17.8	14.4

end times to characters: for each arc in the word graph we run a forced alignment and reconstruct the boundaries for each character. In our second approach, we approximate the character times by distributing the word duration equally over all characters. While the first approach needs 0.5 to 1.0 real time (RT), depending on the graph density, the second approach needs less than 0.01 RT.

From the results in Table 4 we see that the approximated character boundary times effectively work as good as the boundaries derived from a forced alignment. Furthermore, for almost all experiments we observe no difference in minimizing WER or CER. Only ROVER seems to benefit from switching to characters.

6. DECODING FRAMEWORK

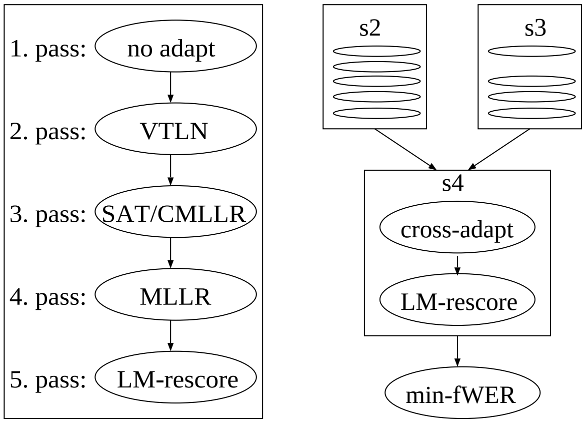
6.1. Multi-Pass Recognition

The first stage of the final decoding framework is divided into five passes as illustrated in Figure 1. The first four passes are realized by a fourgram Viterbi decoder, while the fifth pass is an LM rescoring:

- 1. pass: no adaptation
- 2. pass: 2-pass-VTLN
- 3. pass: SAT/CMLLR
- 4. pass: MLLR
- 5. pass: LM rescoring

In the first pass the ML estimated model with no adaptation is used. From the recognition output we estimate warping factors for the VTLN of the MFCC and PLP features; for gammatones we do not apply VTLN. The next two passes use SAT/CMLLR and MLLR resp., in which adaptation statistics are collected from the previous recognition result. In pass four word graphs are produced, which are rescored with the full LM in the fifth and final pass.

The five passes result in an overall reduction in CER of about 10% relative for eval06 and about 9% for dev07. Detailed results for each system and pass are given in Table 5. The MPE trained models give a further reduction in the CER resulting in a 12% to 15% relative decrease over all passes. Adding the 358 hours of extra data to the MPE training slightly decreases the CER and the total relative improvement is about 16% for both corpora.



(a) 5-pass recognition (b) 2-pass cross-adaptation

Fig. 1. Two stage decoding framework: (a) 5-pass stage for a single system followed by (b) the 2-pass cross-adaptation stage with multiple systems.

A direct comparison of the results of System s1 and s4 shows that the NN based phoneme posterior features reduce the CER by about 15% relative. Interestingly, the gains from VTLN and MLLR are much smaller for s4 than for s1. Experiments with using a fast VTLN warping factor estimator even give an increase in CER for System s4.

Finally, we see that LM.v2 outperforms LM.v1 by about 0.8% absolute in CER consistently for all systems and passes.

6.2. Cross-Adaptation

Cross-adaptation proved to be a simple and effective way to combine systems [22]. In particular, it allows to benefit from systems that show a significantly higher WER or CER than

Table 5. Results for the 5-pass recognition setup.

System		1. pass	2. pass	3. pass	4. pass	5. pass
		eval06 CER[%]				
LM.v1						
s1	ML	24.3	23.8	22.7	22.2	22.0
s2	ML	24.5	24.0	22.9	22.6	22.3
	MPE			21.9		
s3	ML	24.6	-	23.0	22.5	22.2
	MPE			22.1	21.7	21.5
s4	ML	22.7	22.6	20.8	20.7	20.5
	MPE			20.0	19.9	19.6
	MPE+360h			19.7	19.4	19.3
LM.v2						
s4	ML	21.9	21.7	19.9	20.0	19.7
	MPE			19.1	19.0	18.7
	MPE+360h			18.8	18.7	18.3
		dev07 CER[%]				
LM.v1						
s1	ML	21.1	21.4	19.9	19.6	19.4
s2	ML	21.6	21.5	20.1	19.9	19.6
	MPE			19.0		
s3	ML	21.6	-	20.1	19.9	19.7
	MPE			19.6	19.2	19.0
s4	ML	19.7	19.9	17.7	17.7	17.4
	MPE			16.9	16.8	16.6
	MPE+360h			16.9	16.7	16.5
LM.v2						
s4	ML	19.0	18.9	17.1	17.0	16.7
	MPE			16.1	16.0	15.7
	MPE+360h			16.0	15.8	15.5

the target system. In our final decoding framework s4 clearly outperforms s1-s3. Experiments with ROVER give only very small improvements over s4.

We have tried a variety of approaches to find the best way to adapt s4 to s1-s3. Our baseline is the best cross-adaptation result obtained by adapting to the single system Viterbi hypotheses. An alternative is proposed in [23], where the authors adapt to the minimum phoneme error hypothesis. We try a similar approach by applying a modified min.fWER decoder [19]: we derive the time frame phoneme error rate (fPER) as an approximation of the PER in the same way as the fWER. The modified decoder minimizes a linear interpolation of the fPER and the fWER with weights 0.8 and 0.2. Decoding remains on the level of words. This is done, because we want to be able to use the result for cross-adaptation even in the case that the target system uses a different phoneme set. The interpolation weights used let the decoder choose the WER minimizing word sequence in the case that two phoneme sequences are likely.

For adapting s4 to two or three systems simultanously we investigate three approaches. In the first approach we average the adaptation statistics of the individual systems. Technically, we use each utterance multiple times for adaptation by concatenating the results from the individual systems. The second approach is to combine the hypotheses via ROVER. ROVER tends to produce many deletions which might harm adaptation. Thus, we optimize ROVER once for minimal WER (ROVER.1) and once for balanced deletions/insertions (ROVER.2). Finally, we compare the two ROVER combinations with the min.fWER combination method [21]. Results are presented in Table 6.

We try to increase the effect of cross-adaptation by using different acoustic segmentation and LMs for the decoding of s1-s3 and the decoding of the cross-adapted s4. The word graphs produced by the cross-adapted s4 system are subsequently rescored with the full LM. For dev07 we get improvements in CER of about 6% relative, for eval06 only about 3%. Adapting to multiple instead of the best single system only gives small improvements. The best way to adapt to multiple systems is the simplest approach: taking the average of the individual systems' adaptation statistics. Using the minimum phoneme error hypothesis instead of the Viterbi result does not give any improvements at all. We suspect s1 and s4 to be too similar for optimal cross-adaptation results because both systems use the MFCC features. The results support our assumption and we get slightly better error rates when adapting s4 to s2 and s3 only.

7. CONCLUSIONS AND FUTURE WORK

We presented the current RWTH LVCSR system for Mandarin. In the development cycle of the final system we studied the integration of additional feature streams such as tone and NN-based posteriors and the combination of multiple systems. The integration of the additional feature streams via a simple concatenation performs equally well as the more sophisticated approaches. For lattice based system combination we hardly see any improvements in minimizing the character error rate instead of the word error rate. We compared several methods for cross-adapting to multiple systems and gain up to 6% relative in CER by multiple system cross-adaptation in the final decoding framework.

In order to further improve the RWTH Mandarin system a new posterior feature estimation using hierarchical NNs is in progress [24]. We are currently investigating new discriminative training criteria at the RWTH and plan to integrate them into the acoustic training of the Mandarin system.

Acknowledgments

We want to thank UW and SRI for all the support we got, especially Mei-Yuh Hwang, Wen Wang for the language models, Xin Lei for the tonal feature, Gang Peng, and Andreas Stolcke. We want to thank IDIAP for contributing the NN feature training, especially Fabio Valente.

8. REFERENCES

[1] D. Rybach et. al., "Advances in Arabic broadcast news transcription at RWTH," submitted to *IEEE Automatic Speech Recognition and Understanding Workshop*, Kyoto, Japan, Dec. 2007.

[2] 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.

[3] F. de Vriend, N. Castell, J. Gimnez, and G. Maltese, "LC-STAR: XML-coded phonetic lexica and bilingual corpora

Table 6. Cross-Adaptation results for the s4 MPE system trained on 1230 hours (s4 MPE+360h).

		reference CER[%]		s4 MPE+360h CER[%]			
adapt to		LM.v1, seg.v1		LM.v2, seg.v1		LM.v2, seg.v2	
reference		eval06	dev07	eval06	dev07	eval06	dev07
s1	Viterbi	22.1	19.6	18.5	15.4	18.2	14.8
	min.fPER	22.0	19.4	18.6	15.4	18.2	14.8
s2	Viterbi	22.5	19.9	18.5	15.2	18.0	14.6
	min.fPER	22.3	19.7	18.3	15.1	18.1	14.6
s3	Viterbi	22.1	19.6	18.2	15.4	18.0	14.7
	min.fPER	22.0	19.5	18.3	15.5	18.0	14.7
s2+s3	average	-	-	18.0	15.1	17.8	14.5
	min.fPER/avg.	-	-	18.0	15.1	17.9	14.5
	ROVER.1	21.4	19.1			17.9	14.5
	ROVER.2	22.1	19.3			18.1	14.5
	min.fWER comb.	20.9	18.4			18.2	14.6
s1+s2+s3	average	-	-	18.1	15.1	17.8	14.6
	min.fPER/avg.	-	-	18.1	15.1	17.9	14.5
	ROVER.1	20.6	18.4			18.2	14.6
	ROVER.2	22.3	19.3			18.2	14.6
	min.fWER comb.	20.3	18.0			18.2	14.6

for speech-to-speech translation," in *Proc. of Papillon 2004, Workshop on Multilingual Lexical Databases*, Grenoble, France, Aug. 2004.

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

[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] 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.

[7] M.-Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, Pitsburgh, PA, USA, Sept. 2006, pp. 1233–1236.

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

[9] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, Mar. 2005, vol. 1, pp. 457–460.

[10] R. Sinha et. al., "The CU-HTK Mandarin broadcast news transcription system," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, vol. 1, pp. 1077–1080.

[11] B. Xiang, L. Nguyen, X. Guo, and D. Xu, "The BBN Mandarin broadcast news transcription system," in *Proc. European Conf. on Speech Communication and Technology*, Lisboa, Portugal, Dec. 2005, pp. 1649–1652.

[12] 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.

[13] L. Chen, L. Lamel, G. Adda, and J.-L. Gauvian, "Broadcast news transcription in Mandarin," in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000, vol. 2, pp. 1015–1018.

[14] M.-Y. Hwang et. al., "Advances in Mandarin broadcast speech recognition," in *Proc. Int. Conf. on Speech Communication and Technology*, Antwerp, Belgium, Aug. 2007, p. accepted for publication.

[15] 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, p. accepted for publication.

[16] 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.

[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] F. Seide and N. Wang, "Two-stream modeling of Mandarin tones," in *Proc. Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000, vol. 2, pp. 867–870.

[19] F. Wessel, R. Schlüter, and H. Ney, "Explicit word error minimization using word hypothesis posterior probabilities," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Beijing, China, May 2001, vol. 1.

[20] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Santa Barbara, California, USA, Dec. 1997.

[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*, Pitsburgh, PA, USA, Sept. 2006.

[22] 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.

[23] J. Ogata and Y. Ariki, "Unsupervised acoustic model adaptation based on phoneme error minimization," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, CO, USA, Sept. 2002, pp. 1429–1432.

[24] F. Valente et. al., "Hierarchical neural networks feature extraction for LVCSR system," in *Proc. Int. Conf. on Speech Communication and Technology*, Antwerp, Belgium, Aug. 2007, p. accepted for publication.