IMPROVED MODELS FOR MANDARIN SPEECH-TO-TEXT TRANSCRIPTION

Lori Lamel, Jean-Luc Gauvain, Viet Bac Le, Ilya Oparin, Sha Meng

Spoken Language Processing Group, LIMSI-CNRS
91403 Orsay, FRANCE
{lamel,gauvain,levb,oparin,mengsha}@limsi.fr

ABSTRACT

This paper describes recent advances at LIMSI in Mandarin Chinese speech-to-text transcription. A number of novel approaches were introduced in the different system components. The acoustic models are trained on over 1600 hours of audio data from a range of sources, and include pitch and MLP features. N-gram and neural network language models are trained on very large corpora, over 3 billion words of texts; and LM adaptation was explored at different adaptation levels: per show, per snippet, or per speaker cluster. Character-based consensus decoding was found to outperform word-based consensus decoding for Mandarin. The improved system reduces the relative character error rate (CER) by about 10% on previous GALE development and evaluation data sets, obtaining a CER of 9.2% on the P4 broadcast news and broadcast conversation evaluation data.

Keywords: Mandarin, speech-to-text transcription, speech recognition, character error rate

1. INTRODUCTION

This paper describes recent advances in Mandarin speech-to-text (STT) transcription at LIMSI, for which all system components have been improved. Compared with the most state-of-art systems for the same purpose [3, 14, 21, 17], a number of the novel and difference approaches are introduced and described. Developments in preparation for the Phase 4 of the Global Autonomous Language Exploitation (GALE) program addressing transcription of varied broadcast news (BN) and conversation (BC) data are highlighted.

On the acoustic modeling side, there has been growing interest of the use of discriminative features produced by a multi layer perceptron (MLP) in speech-to-text transcription systems [5, 6]. The MLP features developed in our system are based on a recently proposed Bottle-Neck architecture [9] with long-term speech representation at the input. Pitch features (F0) are also investigated since it is generally assumed that pitch features are important to capture the tone differences characteristic of the language. Several pitch extraction methods were compared in order to determine which method gives the best STT performance for our system.

Concerning language modeling, the standard n-gram LMs used for both decoding and lattice rescoring are obtained by interpolating up to 48 unpruned component LMs. Language model adaptation has been explored at multiple levels, based on each show, each snippet or each cluster. Additional Neural network LMs (NNLMs) used for final lattice rescoring are developed to make use of continuous representation of words, instead of the discrete space in conventional N-gram LMs. Some changes have been made to the pronunciation dictionary in particular to capture modifications due to the Sandhi tone [23]. Previous work has shown the advantage of an explicit WER minimization using an N-best list or a confusion network instead of using MAP decoding [16]. Since for Mandarin tasks, the STT results are typically measured in terms of character error rate (CER), the use of a character confusion network (similar to [7, 11]) was investigated to replace word-based confusion network decoding.

2. LANGUAGE MODELS

This section describes updating the standard N-gram backoff LMs, training the connectionist LMs, and experiments with LM adaptation. The first step is to extract 4-grams for each individual source. Then Kneser-Ney 4-gram LMs are estimated without any cutoff. The interpolation coefficients for the component LMs are tuned on the 13-hour dev09 set and the final LM generated. Four Neural Network LMs (NNLM) were also trained on a subset of the data, and interpolated with the 4-gram LM.

2.1. Backoff Language Models

Since words in written Chinese are not separated by white spaces, one either has to make use of character-based LMs or perform word segmentation as a pre-processing step. Since the former was found to be inferior to the latter [10, 15], the segmentation approach was adopted. The longest-match algorithm was used with a 56k word vocabulary (including 6k characters) to segment all of the training texts, so there are no out-of-vocabulary words in the text after segmentation. We also explored using automatic methods similar to [12] and available resources such as the OntoNotes corpus to increase the recognition word list, but thus far results have been inconclusive and the original vocabulary was kept.

About 590M words of segmented texts were added to the prior data, resulting in a pool of 48 text sources. This represents a 22% increase over the quantity of LM training texts previously used with a total of 3.2 billion segmented words in the full LM training data.

Language model training is performed with LIMSI STK toolkit. This toolkit allows efficient handling of huge language language models without any pruning or cutoff. This is an important feature as normally different cutoffs are applied to fit an LM trained on large data in memory. In our case, all information in the training data is kept, even though there are over 3 billion words. The N-gram hit rates and perplexity of the different development data sets with the P4 4-gram LM are shown in Table 1. The transcriptions of the 13-hour dev09 data set were used to tune the interpolation weights for the component LMs. The perplexity of the dev09 and dev09s data sets are reduced by about 10% with updated LM relative to the previous one. The component LMs with the largest interpolation weights are listed in Table 2, with the top 8 sources accounting for over 70% of the weight (4 of them estimated on transcripts).

2.2. Neural Network Language Models

In contrast to conventional N-gram LMs in which words are represented in a discrete space, Neural network LMs (NNLMs) make use

Table 1. N-gram hit-rates and perplexity on different GALE data sets with the updated 56K 4-gram LM.

Dev set	*1g(%)*	*2g(%)*	*3g(%)*	*4g(%)*	*ppx*
dev09s	1.6	26.0	37.7	34.7	207.8
dev08	1.3	25.2	37.9	35.6	192.5
dev07	1.3	25.5	37.9	35.4	184.7
eval07	1.5	25.8	38.2	34.5	206.6
dev07/08+eval07	1.4	25.5	38.0	35.1	194.5

Table 2. Top P4 component LMs ordered by interpolation weight.

CompLM	*type*	*weight*	*CompLM*	*type*	*weight*
bcm	transcr	0.176	bcm_P4	transcr	0.070
bnm	transcr	0.108	ibm_sina	news	0.067
ng	webdata	0.086	giga_cns	news	0.065
giga_xin	news	0.081	bcm_dev09train	transcr	0.050

of continuous-space representation of words, which enables a better estimation of unseen N-grams. The neural network deals with two tasks: projection of words with history to continuous space and calculation of LM probabilities for the given history. Both these tasks can be performed with a NN that contains two hidden layers [19]. NNLMs have been shown to improve over the N-gram baseline for different languages and tasks [20]. Recent evaluations have shown that NNLMs bring improvements even when the N-gram LM is trained on very large amounts of data without any cutoff and pruning, and with a well-tuned STT system.

Neural network LMs are used to rescore lattices generated with conventional N-gram LMs. In the current system four different neural networks were generated with different number of nodes in the hidden layer. These individual NNLMs were subsequently interpolated into one general NNLM for more robust probability estimation. The networks vary in the size of the hidden layer (500, 450, 500, 430), and the projection size of P-dimensional continuous space (300, 250, 200, 220). Three previous words form an input to the NN, and the 8K most frequent words are used as a shortlist to estimate the probabilities at the output layer as described in [19, 20].

Since it is not feasible to train a NNLM on all the available data, the data used to train the NNLMs were selected according to the interpolation weights assigned to the component N-gram models for the different data sources. The subset of data includes all BC and BN sources and the four text sources with the highest interpolation weights. The latter text sources are quite large, and were therefore downsampled as shown in Table 3). The application of the NNLM leads to a perplexity reduction of about 15% on different test sets (for example, the dev09 perplexity is reduced from 211 to 186), and consistent reductions in CER of 0.4-0.5% are observed.

2.3. Language Model Adaptation

Following work on LM adaptation reported in [1] and [13], we experimented adapting the mixture weights of the component language models in the STT system. For a given test set, the system's 1-best hypothesis obtained using the unadapted LM is used as target data to estimate new LM component weights and the word lattices are rescored using the new weights. Results obtained on the dev08 data set are given in Table 4 using a system without a neural network LM. The perplexities shown in the table are computed using the reference transcripts. Three adaptation levels have been explored: with a set of weights for each show, each snippet, or each speaker cluster. The snippets are obtained manually whereas the speaker clusters are produced automatically using our audio partitioner. Since the manual transcripts cannot be easily mapped to the automatic speaker clusters, the corresponding perplexity is not reported but it is expected to be between the show-based and the snippet-based perplexities. It can be seen in Table 4 that the best results are obtained using cluster-based adaptation, which results in a CER of 9.3% to be compared to 9.5% without adaptation.

Table 3. Neural Network LM sampling parameters. The total number of words/subset are given, along with the sampling factor, and the number of words after sampling.

Corpus	*#words*	*sample factor*	*#sampled words*
bcm/bnm	19.4M	0.9	17.4M
ng	315.5M	0.006	1.89M
giga_xin	367.0M	0.005	1.83M
ibm_sina	279.9M	0.004	1.11M
giga_cns	76.7M	0.01	0.76M

Table 4. Perplexities and CERs with different LM adaption levels (dev08).

Adaptation level	*PPX*	*CER (%)*
None	195.4	9.5
show-based	168.1	9.4
snippet-based	146.1	9.4
cluster-based	NA	9.3

3. RECOGNITION LEXICON

The recognition vocabulary of our P4 system contains 56k entries, including (and composed of) 6482 characters. There are a total of 7462 distinct pronunciations associated with the characters for about 1.15 pronunciations/character. (Although there are more characters in Mandarin Chinese, these are very rare in modern texts and have not been included in the vocabulary). It is important that all possible pronunciations are associated with each character, since word pronunciations are formed by concatenation of their component characters. In order to verify this, pronunciations for all single characters were extracted from an online source (www.weeeeb.com/w/). Additional pronunciations were found for 20 frequent characters, and added into dictionary for both single-character words and the multi-character words containing those characters. canonical F0 contour pattern. Tones in continuous speech undergo modifications due to the Sandhi tone [23]. These were not fully represented in the dictionary, so pronunciations were added according to two Sandhi rules. First, the pronunciations of two frequent characters, 'no' and 'one', depend on their context. Second, all characters with tone 3 should change to the tone 2 when preceding a character with tone 3 [22]. These changes had a negligible (under 0.1%) effect on the CER.

4. ACOUSTIC MODELING

There has been increasing use of discriminative features produced by a multi-layer perceptron in STT transcription systems [5, 6]. While the MLP features have never been shown to consistently outperform cepstral features (PLP), the performance of state-of-the-art STT systems has been improved when both types of features are used in conjunction. In this work, the MLP features are based on the Bottle-Neck architecture [9]. Pitch features (F0) are also investigated since it is generally assumed that pitch features are important to capture the tone differences characteristic of the language. Several pitch extraction methods were compared in order to determine which method gave the best STT performance for our system. This section presents an overview of recent studies with pitch and MLP features sets. The

Table 5. Comparison of pitch methods and smoothing on 4 data sets.

Features	*bndev06*	*bcdev05*	*bndev07*	*bcdev07*
PLP	11.0	25.1	8.8	26.4
PLP+F0 LIMSI	10.3	24.8	6.0	24.5
PLP+F0 CU	11.7	26.0	7.6	27.8
PLP+F0 linear	09.9	22.9	5.7	24.1

PLP analysis has been used in all LIMSI STT systems since 1996 and is described in [8]. The acoustic models (AM) were trained on over 1600 hours of manually transcribed broadcast news and broadcast conversation data distributed by LDC, using both standard PLP and concatenated MLP+PLP+F0 features. For the PLP models, a maximum-likelihood linear transform (MLLT) is also used. The model sets cover about 49k phone contexts, with 11.5k tied states and 32 Gaussians per state. Silence is modeled by a single state with 2048 Gaussians. Speaker-independent models are trained on all available data (1422 hours, 684h female/734h male of speech data after eliminating silence), and serve as priors for Maximum *a Posteriori* (MAP) estimation of gender-specific models.

4.1. Pitch features

Different pitch detection methods and smoothing techniques were explored. An in-house autocorrelation based method with linear interpolation for unvoiced segments was compared to the ESPS pitch extraction with both linear and CU mode interpolation methods. A 3-dimensional pitch feature vector (pitch, Δ and $\Delta\Delta$ pitch) is added to the original PLP feature, resulting in a 42-dimension feature vector (PLP+F0). Character error rates (CER) are shown in Table 5 for the different methods for a single decoding pass and a 4-gram language model. The best results for all test sets were obtained using the ESPS pitch extraction with linear interpolation. Another smoothing, "Piecewise Cubic Hermite Interpolating Polynomial" was also considered, however the performance degraded by about 0.5%.

4.2. MLP features

MLP features have been successfully used in the LIMSI STT systems since 2007. Combined with classical PLP features (and pitch features), these probabilistic features significantly reduce the word error rate for Mandarin and other languages. The MLP features are generated in two steps. The first step is raw feature extraction which constitutes the input layer to the MLP. In this work, the wLP-TRAP (Time-warped linear predictive TRAP [4]) and the TRAP-DCT (TD) [18] features are used. The wLP features are costly to calculate since they use very large FFT transformations. The wLP features contain 25 LPC coefficients in 19 frequency bands $\rightarrow 19 \times 25 = 475$ raw features. The TRAP-DCT features, a promising alternative to the wLP features, were shown to have similar performance but are cheaper to compute than the wLP-TRAP [18]. The TD features are obtained from a 19-band Bark scale spectrogram, using a 30 ms window and a 10 ms offset. A discrete cosine transform (DCT) is applied to each band, also resulting in 475 raw features. The raw features are the input to a 4-layer MLP with the bottle-neck architecture [5, 6, 9]. The size of the third layer (the bottleneck) is equal to the desired number of features (39). In a second step, the raw features are processed by the MLP and the features are not taken from the output layer of the MLP but from the hidden bottleneck layer and decorrelated by a PCA transformation. The STT system thus uses a 81-parameter feature vector formed by concatenating the MLP, PLP and pitch features (MLP+PLP+F0).

The MLP network was trained using the simplified training scheme proposed in [24] on about 928 hours of BN+BC data from a variety of sources. The training data are randomized and split in three non-overlapping subsets, used in 6 training epochs with fixed learning rates. The first 3 epochs use only 13% of data, the next 2 use 26%, the last epoch uses 52% of the data, with the remainder used for cross-validation to monitor performance. The MLP has 256 targets, corresponding to the individual states for each phone and one state for the additional pseudo phones (silence, breath, filler). The frame accuracies for the MLP$_{TD}$ and the MLP$_{wLP}$ are shown in Table 6. It can also be seen that the frame classification is higher for the wLP features than the TD ones.

Table 6. Frame accuracies with TD and wLP raw features

Features	*train data*	*#hidden nodes*	*%trn acc*	*%CV acc*
TD	928hrs	3.5k	53.7	53.3
wLP		3.5k	56.0	55.4

Table 7. CER on dev07 data with automatic partitioning and 4-gram LM for various feature combination schemes.

Features	PLP+F0	MLP$_{TD}$+ PLP	MLP$_{wLP}$+PLP	MLP$_{wLP}$+PLP+F0
CER (%)	12.3	12.3	12.0	11.4

Table 7 reports CERs on the dev07 data set using a 2-pass decoding scheme with unsupervised acoustic model using the CMLLR and MLLR techniques and a 4-gram word LM. The PLP+F0 models and the MLP$_{TD}$+PLP without pitch both give a CER of 12.3%. The combination of MLP$_{wLP}$ and PLP performs slightly better, with an absolute error reduction of 0.6% when pitch is included in the feature vector (MLP$_{wLP}$+PLP+F0).

4.3. Initial-Final Models

It has been recently reported that initial-final systems outperform phone-based models on a large training corpus [2]. An initial-final phone set was defined, combining prevocalic glides with initial consonants as the initial (C or C+G), and the vowel (optionally followed by a nasal) as the final (V or V+N). Most syllables are made of an initial unit and a final unit, with only 45 out of 1348 syllables comprising of only a final unit. A total of 53 initial units and 83 final units were selected according to the rules and classification in standard Mandarin Pinyin (www.pinyin.info/rules/initials_finals.html).

Several sets of initial-final AMs have been developed, and tested using a single decoding pass system, with pronunciation probabilities. Table 8 gives CERs for different experimental configurations. The first row reports the baseline CER (14%), obtained with gender-dependent, phone-based AMs. The second row corresponds to gender-independent, initial-final models covering 57k contexts with 11.5k states, with a CER of 15%. Increasing the number of model contexts by almost a factor of 3 (to 147k) only gives a tiny reduction of 0.1% if the number of tied states is kept the same (compare S1 and S2), with a more moderate improvement when the number of states is increased (S3 vs S4). The CER with the S4 system is 0.2% better than the phone baseline, and Rover of the two system outputs

Table 8. CER for Initial-Final PLP+F0 systems on dev09s data.

System	*Description*	CER	Del	Ins
Phone	Baseline, 49k ctx, 11.5k tied state	14.0	3.6	1.1
S1	57k ctx, 11.5k tied states	15.0	3.3	1.5
S2	147k ctx	14.9	3.3	1.4
S3	+ gender	13.9	3.0	1.4
S4	+ 35k tied states	13.6	2.8	1.6
ROVER	S4 + phone	12.9	3.0	1.5

Table 9. Comparing of word and character consensus decoding.

Dev09s	*MLP+PLP+F0*	*PLP+F0*	*ROVER (c0.4)*
Word CN	9.9	10.7	9.9
Char CN	9.8	10.5	9.7

Table 10. CER on 5 Mandarin sets with LIMSI P3.5 and P4 systems.

System	*dev07*	*eval07ns*	*dev08*	*eval08ns*	*dev09s*
P3.5	9.7	9.1	9.2	13.7	10.4
P4	9.3	8.4	8.5	12.5	9.4

gives an additional error reduction of 0.8%. The initial-final models are still under development and have not yet been incorporated in the evaluation system.

5. WORD VS CHARACTER CONSENSUS DECODING

Previous work has shown the advantage of an explicit WER minimization using an N-best list or a confusion network instead of using MAP decoding. For consensus decoding [16], a word lattice is converted to a word confusion network and the 1-best word consensus hypothesis is obtained by taking the word with the highest confidence score in each confusion network slot. For Mandarin the STT results are measured in term of character error rate instead of WER, so instead of using word confusion network we investigated the use of character confusion network (similar to [7, 11]). To generate the character confusion network, each edge of the word lattice is simply split into individual characters, then character consensus decoding is performed.

Results for the MLP-based (MLP+PLP+F0), the PLP-based and the ROVER combination systems are shown in Table 9 for the dev09s data. We observed that the CER reduced by an absolute 0.1% for the individual systems and about 0.25% with ROVER.

6. EXPERIMENTAL RESULTS

The STT system has one decoding chain with three steps. Each decoding step generates a word lattice with cross-word, position-dependent, gender-independent acoustic models, followed by consensus decoding with a 4-gram LM and pronunciation probabilities [8]. Unsupervised acoustic model adaptation is performed for each segment cluster using the CMLLR and MLLR techniques prior to the next decoding pass. Different AMs are used in successive decoding passes. The MLP+PLP+F0 model is used in the first and third one, while the PLP+F0 based model is used in the second pass. The interpolated connectionist 4-gram LM is used in the final pass.

Table 10 summarizes the STT results on five BN+BC Mandarin data sets used in the GALE community, with the LIMSI P3.5 system and the P4 systems. The P4 systems gives a relative CER reduction of about 8% for most test sets. Table 11 gives the CER of the LIMSI Mandarin STT component of the AGILE system on the GALE 2009 evaluation data. The overall CER is about 9%, with a very large difference in performance on the BN and BC data subsets, the BC CER being almost 5 times larger than that for BN.

7. SUMMARY

This paper has highlighted recent improvements in the LIMSI Mandarin STT system which has state-of-the-art performance on varied broadcast news and broadcast conversation data. The system is trained on large audio and text corpora available in the GALE program, and was a component system in the AGILE team participation to the GALE Phase 4 evaluation. Although the CER of well prepared BN speech is quite low (3.4%), there is clearly room for improvement on more varied BC data.

Table 11. CER on GALE P4 eval data, overall and BN/BC subsets.

	BN+BC	*BN*	*BC*
eval09	9.2	3.4	15.3

ACKNOWLEDGMENTS

The authors acknowledge the help of Jun Luo, who contributed to earlier LIMSI systems. This work was in part supported by the DARPA GALE program, and by OSEO, the French State agency for innovation under the research program Quaero.

8. REFERENCES

[1] L. Chen et al. Dynamic Language Modeling for Broadcast News, *ICSLP'04*.

[2] S.M. Chu et al., Recent advances in the IBM GALE Mandarin Transcription System, *ICASSP'08*.

[3] S. Chu, et al. The IBM 2009 Mandarin Broadcast Transcription System, *ICASSP'10*.

[4] P. Fousek, *Extraction of Features for Automatic Recognition of Speech Based on Spectral Dynamics*, PhD, Czech Tech Univ, Prague, 2007.

[5] P. Fousek, L. Lamel, J.L. Gauvain, On the Use of MLP Features for Broadcast News Transcription, *TSD08*.

[6] P. Fousek, L. Lamel, J.L. Gauvain, Transcribing Broadcast Data Using MLP Features, *Interspeech'08*.

[7] Y.S. Fu, Y.C. Pan, L.S. Lee, Improved Large Vocabulary Continuous Chinese Speech Recognition by Character-Based Consensus Networks, *ISCSLP'06*.

[8] J.L. Gauvain, L. Lamel, G. Adda, *The LIMSI Broadcast News Transcription System*, Speech Communication, **37**(1-2):89-108, 2002.

[9] F. Grézl, P. Fousek, Optimizing Bottle-Neck Features for LVCSR, *ICASSP'08*.

[10] J.L. Hieronymus et al., Exploiting Chinese Character Models to Improve Speech Recognition Performance, *Interspeech'09*.

[11] V.B. Le et al., Word/sub-word lattices decomposition and combination for speech recognition, *ICASSP'08*.

[12] X. Lei, W. Wang, A. Stolcke, Data-driven Lexicon Expansion for Mandarin Broadcast News & Conversation Speech Recognition, *ICASSP'09*.

[13] X. Liu, M. J. F. Gales, P.C. Woodland, Context Dependent Language Model Adaptation, *Interspeech'08*.

[14] X. Liu, M.J.F. Gales, P.C. Woodland, Language Model Cross Adaptation For LVCSR System Combination, *Interspeech'09*.

[15] J. Luo, L. Lamel, J.L. Gauvain, Modeling Characters Versus Words for Mandarin Speech Recognition, *ICASSP'09*.

[16] L. Mangu, E. Brill, A. Stolcke, *Finding consensus in speech recognition: word error minimization and other applications of confusion networks*, Computer, Speech & Language, **14**(4):373-400, 2000.

[17] T. Ng et al., Progress in the BBN 2007 Mandarin Speech to Text System, *ICASSP'08*.

[18] P. Schwarz, P. Matějka, J. Černocky, Towards Lower Error Rates In Phoneme Recognition, *TSD'04*.

[19] H. Schwenk, *Continuous Space Language Models*, Computer, Speech & Language, **21**:492-518, 2007.

[20] H. Schwenk, J.L. Gauvain, Training Neural Network Language Models On Very Large Corpora, *JHLT/EMNLP*, 2005.

[21] L. Xin et al., Development of the 2008 SRI Mandarin Speech-to-text System for Broadcast News and Conversation, *Interspeech'09*.

[22] H. Yu, et al.,The ISL RT04 Mandarin broadcast news evaluation system, *EARS Rich Transcription Workshop*, 2004.

[23] S. Zhang et al., Main Vowel Domain Tone Modeling with Lexical and Prosodic Analysis for Mandarin ASR, *ICASSP'09*.

[24] Q. Zhu et al., Using MLP features in SRI's conversational speech recognition system, *Interspeech'05*.