IMPROVING SPEECH TRANSCRIPTION FOR MANDARIN-ENGLISH TRANSLATION

M. Tomalin, M.J.F. Gales, X.A. Liu, K.C. Sim, R. Sinha, L. Wang, P.C. Woodland, & K. Yu*

Engineering Department, Cambridge University, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {mt126, mjfg, xl207, kcs23, rs460, lw256, pcw, ky219}@eng.cam.ac.uk

ABSTRACT

This paper describes the development of the CU-HTK Mandarin Speech-To-Text (STT) system and assesses its performance as part of a transcription-translation pipeline which converts broadcast Mandarin audio into English text. Recent improvements to the STT system are described and these give Character Error Rate (CER) gains of 14.3% absolute for a Broadcast Conversation (BC) task and 5.1% absolute for a Broadcast News (BN) task. The output of these STT systems is then post-processed, so that it consists of sentence-like segments, and translated into English text using a Statistical Machine Translation (SMT) system. The performance of the transcription-translation pipeline is evaluated using the Translation Edit Rate (TER) and BLEU metrics. It is shown that improving both the STT system and the post-STT segmentations can lower the TER scores by up to 5.3% absolute and increase the BLEU scores by up to 2.7% absolute.

***Index Terms*—** Speech Recognition, Sentence Boundary Detection, Machine Translation

1. INTRODUCTION

This paper explores a transcription-translation pipeline that converts Mandarin audio into English text. The system has three main components:

- **STT Component**: this converts Mandarin audio into Mandarin text.
- **Integration Component**: this post-processes the STT Mandarin output by mapping spoken numbers to digits and by subdividing some of the STT segments into smaller 'sentence-like' units.
- **SMT Component**: this translates the post-processed Mandarin text into English text.

Since the SMT system used in the experiments was trained using standard text rather than speech transcription data, the Integration component was designed to convert the STT token sequences into the kind of sentence-like groupings that are encountered in standard text.

The performance of the STT component was assessed using the CER metric, while the performance of the whole system was evaluated using TER and BLEU scores (these metrics are defined in section 4). This paper investigates some of the interactions that occur between the three main system components, and it focuses primarily on two aspects of the pipeline. First, the impact of improvements to the STT component are explored both at the STT stage (using CER) and at the translation stage (using TER/BLEU). Second, the impact on SMT performance of resegmenting the STT output into sentence-like units is investigated.

2. THE MANDARIN STT SYSTEMS

Two Mandarin STT systems will be discussed in this paper. The first system, STT-05, is described in [1] and the architecture is shown in Figure 1. This system used a multi-pass/multi-branch framework to convert Mandarin audio into Mandarin text. The P1 stage uses gender independent models to provide adaptation supervision for the P2 lattice generation stage. The P3 stage performs Constrained Maximum Likelihood Linear Regression (CMLLR) and lattice-based Maximum Likelihood Linear Regression (MLLR) using the 1-best adaptation supervision and lattices from the P2 stage. The final system output was derived by combining various P3 outputs using Confusion Network Combination (CNC).

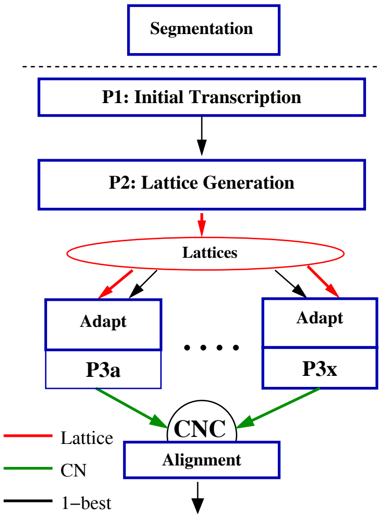

Fig. 1. The multi-pass and multi-branch framework for the STT-05 system.

The STT-05 system used 148 hours of BN acoustic model training data. 28 hours of this was Hub-4 data released by the Linguistic Data Consortium (LDC) with accurate transcriptions. For the remaining 120 hours of TDT4 Mandarin BN data, only closed-captions references were provided, so light supervision techniques

K.C. Sim is now with the Institute for Infocomm Research in Singapore.
This work was in part supported by DARPA under the GALE programme via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred. We would like to thank BBN Technologies and Language Weaver for their assistance with this research.

were used [2]. Approximately 1 hour of the TDT4 Mandarin data consisted of English, and a further 10 hours of TDT4 English training data were folded in to boost the system's English output. Minimum Phone Error (MPE) trained triphone models were built using the available acoustic data.

The STT-05 system used a Language Model (LM) that was trained using 366M words from five sources all released by LDC: the correct acoustic transcripts for Hub-4 Mandarin data, China Radio, Mandarin TDT[2,3,4], GigaWord (Xin Hua) and People's Daily. In addition, downloaded webdata for up to March 2004 was included. A 55K word-list was used for both the language and acoustic model training and testing, and this word-list covered all English and Mandarin words in the acoustic training data.

The second system, STT-06, used the same basic architecture as the STT-06 system, although, rather than using the 1-best supervision from the P2 stage, cross-adaptation was used. Specifically, the CU multi-pass system generated test data lattices, and BBN provided system output as the supervision to adapt the CU acoustic models. The multi-pass BBN system was trained on similar training data to the CU system and it produced an alternative acoustic segmentation [3]. The BBN confidence scores for each recognised word were used when confidence score based adaptation was performed. The adapted models were used to rescore the lattices and generate the final output. The STT-06 system architecture is shown in Figure 2.

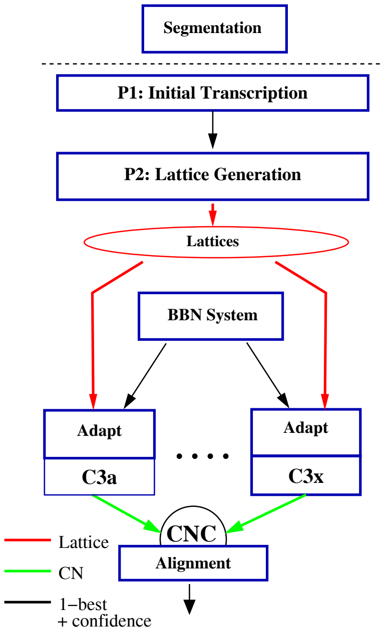

Fig. 2. The multi-pass and multi-branch evaluation framework for the STT-06 system.

The STT-06 system acoustic models were trained using 504 hours of data – the 148 hours used by the STT-05 system, plus an additional 356 hours of which 156 hours consisted of BC data while the rest was BN speech. This additional data covers all the incremental LDC GALE releases of Chinese broadcast training data from December 2005 to May 2006. Once again, MPE triphones were built.

The STT-06 LM was trained using 1.2G words, the 355M words used by the STT-05 LM plus an additional 770M words. The new LM training data included all the transcriptions of the additional acoustic data, more downloaded webdata collected at CU up to the end of January 2006, extra data from the LDC Chinese GigaWord II release, and Phoenix TV downloaded webdata from June 2004 to January 2006 collected by SRI, University of Washington (UW) and National Taiwan University (NTU). A 3-way interpolation between the BN, BC and Phoenix TV data based component LMs was used to build the final LM, and a tuning set consisting both BN and BC data was use to optimise the interpolation weights to obtain a more balanced CER performance for both genres. The acoustic and language models used an expanded list of 52k multiple character Chinese words for character to word segmentation. The revised recognition vocabulary contains a total of 58k words approximately, which includes an additional 5k single character Chinese words, and 269 highly frequent English words observed in all the acoustic training transcriptions.

The STT-05 and STT-06 systems both used the same pre-STT segmentation-clustering scheme. In this scheme, a GMM classifier splits the data into wideband speech, telephone speech, speech with music and pure music regions, a gender dependent phone recogniser is then run to locate gender-change points and silence, then a symmetric divergence based change point detector and BIC agglomerative clustering stage is used to refine the segmentation [1]. The purpose of this segmentation stage was to divide the audio data into homogeneous acoustic blocks, and a maximum segment length limit of 30 seconds was used.

CER results for the STT-05 and STT-06 systems for a BC and a BN test set are given in Table 1. The bcm test set consists of 2.5 hours of Mandarin BC data taken from 5 shows (2 VOA and 3 PHX) that were broadcast in March 2005. The bnm test set contains 3.5 hours of BN Mandarin data, selected from 14 shows (7 CCTV, 3 PHX, 1 VOA, 1 CNR, 1 RFA, 1 NTDTV) that were broadcast between February 2001 and October 2005. It should be noted that the purpose of this section is not primarily to explore in detail the various contributions of the differences that distinguish the STT-05 and STT-06 systems. Rather, it is simply to present two different STT systems which achieve different levels of performance so that the consequences of STT improvements for SMT can be investigated.

Test	System	# Segs	CER
bcm	STT-05	836	32.4
	STT-06	836	18.1
bnm	STT-05	934	17.4
	STT-06	934	12.3

Table 1. CERs for the STT-05 and STT-06 systems for the bcm and bnm test sets.

The CERs for the bcm test set show the largest reduction (14.3% absolute). This is mainly due to the fact that the STT-06 system was trained using BC data, while the STT-05 system did not use BC data. However, the CER reduction of 5.1% absolute for the bnm test set indicates that the STT-06 system generally outperforms the STT-05 system.

3. POST-PROCESSING STT MANDARIN OUTPUT

In recent years it has become desirable to produce STT output that contains information concerning sentence boundaries or 'Slash Units' (SUs) [4] [5] [6]. SUs are sentence-like units, not traditional 'grammatical' sentences, and recent studies have demonstrated that text readability is improved if information about SU boundaries is included in STT transcriptions [7]. However, it is useful to establish whether SMT systems benefit from taking sentence-like STT segmentations as input.

Two post-STT segmentation strategies are explored in this paper. Both take (segmented) transcriptions of Mandarin audio as input and split the STT segments into smaller units using sentence boundary information. The guiding assumption is that the STT segmenter undersegments the data, hence the need for a subsequent segment splitting stage. These sentence-like segments are then passed to the SMT component to be translated into English text. The two post-processing systems compared here are the following:

- **SSP-NGRAM**: this system uses audio silence splitting and a fourgram Hidden Event Language Model (HELM) [8] in a sequential configuration.
- **PFM-NGRAM**: this system combines a CART-Style Decision Tree Prosodic Feature Model (PFM) and a fourgram in a decoding framework.

In the SSP-NGRAM system, the silence splitting occurs if the inter-word silence gaps are longer than 0.9 seconds. Also, a maximum segment length of 20 seconds is imposed. These values were both determined empirically. The splitting process is applied recursively until all segments are shorter than the maximum segment length. The resulting segments are then further split using a word-based HELM fourgram trained on 500M words of training data. This data was a subset of the STT-06 LM training data, and all sentence boundaries were marked by a unique token. Since the HELM gives the posterior probability of a given token functioning as an SU boundary, $p(T)$, a constant threshold value can be set which determines the frequency with which the SU boundaries are inserted by the HELM. The threshold value was defined as *thresh* = 1 - p(T) and an empirically determined *thresh* value of 0.75 was used.

The PFM-NGRAM system described here is based on the system discussed in [6]. A task-specific fourgram and PFM were combined in a lattice-based 1-Best Viterbi decoding framework. A word-based fourgram was constructed using exactly the same sentence-boundary marked training data as the HELM. Since the initial purpose of this work was to contrast the SSP-NGRAM and PFM-NGRAM systems, the PFM was built using only a single prosodic feature – namely, the absolute duration of the pause that followed each lexical item. The PFM was created as described in [6] and it was built using 28 hours of Hub-4 Mandarin data. The fourgram was combined with the PFM in a lattice-based 1-Best Viterbi decoding framework with an empirically determined grammar scale factor of 0.8. The 1-Best decoder output produced token sequences for each file in the test sets, and these contained the STT lexeme token sequence and SU boundary tokens that had been inserted automatically during the decoding process.

It is useful at this point to stress the main differences between these two systems. For instance, while the SSP-NGRAM system implements the acoustic and lexical stages separately, in sequence, the PFM-NGRAM system combines them in an integrated decoding framework. Consequently, while the acoustic information incorporated into the SSP-NGRAM system is not dependent on lexical information of any kind, this is not true of the PFM-NGRAM system since, in that framework, each lexical item is associated with the probability of an acoustic event.

4. THE STATISTICAL MACHINE TRANSLATION COMPONENT

The post-processed STT output was translated using the Language Weaver Chinese-English v4.0 SMT system. This is a high-performance real-time phrase-based statistical translation system [9]. The same SMT system was used for all the experiments discussed in this paper.

The SMT output was evaluated using TER and BLEU scores. The TER score measures the ratio of the string edits between a word sequence in the target language (E) and the word sequence in the reference (E_r) to the total number of words in the reference [10]. Permissable edits include Insertion (Ins), Deletion (Del), Substitution (Sub) and Phrase Shift (Shft)[1]:

$$\texttt{TER}(E, E_r) = \frac{\text{Ins} + \text{Del} + \text{Sub} + \text{Shft}}{N} \times 100 \qquad (1)$$

where N is the total number of words in the reference. In all the experiments described here, lower case texts were used. The NIST BLEU score is a variant of the metric detailed in [11]. It computes the geometric mean of the precision of ngrams and includes a brevity penalty ($\gamma(E, E_r) \leq 1$) if the hypothesis is shorter than the reference:

$$\texttt{BLEU}(E, E_r) = \left(\exp\left[\frac{1}{N}\sum_{n=1}^{N}\log p_n(E, E_r)\right]\gamma(E, E_r)\right) \times 100 \qquad (2)$$

where $p_n(E, E_r)$ is the precision of ngrams in the hypothesis, E, given the reference, E_r. In this paper, $N = 4$ was used.

5. RESULTS

The performance of the composite transcription-translation system described in this paper was assessed using the bcm and bnm test sets. Results for the STT-05 system with post-processing are given in Table 2.

Test	Post-Processing	#Segs	TER	BLEU
bcm	STT-05	836	79.89	7.40
	+ SSP-NGRAM	2148	79.21	8.05
	+ PFM-NGRAM	1817	**78.44**	**8.18**
bnm	STT-05	934	74.37	11.60
	+ SSP-NGRAM	2172	73.02	12.71
	+ PFM-NGRAM	2151	**72.62**	**12.86**

Table 2. TER and BLEU scores for STT-05 systems for the bcm and bnm test sets.

The two post-processing schemes both produce more than double the number of segments in the STT output for the BC and BN tasks, and the results indicate that subdividing the STT segments into sentence-like units can improve system performance by up to 1.5% absolute (TER) and 0.8% absolute (BLEU) for the BC task, and 1.8% absolute (TER) and 1.3% absolute (BLEU) for the BN task.[2] In all cases, the system that used the PFM-NGRAM in the integration component achieved the best TER and BLEU scores.

[1] Phrase shift is the movement of a contingent block of words from one location in the hypothesis to another

[2] All results are given as 'absolute' rather than 'relative' numbers unless stated otherwise.

Test	Post-Processing	#Segs	TER	BLEU
bcm	STT-06	836	76.47	9.01
	+ SSP-NGRAM	2296	75.02	9.66
	+ PFM-NGRAM	2259	**74.58**	**10.06**
bnm	STT-06	934	73.21	12.39
	+ SSP-NGRAM	2007	72.07	13.21
	+ PFM-NGRAM	2298	**71.38**	**13.57**

Table 3. TER and BLEU scores for STT-06 systems for the bcm and bnm test sets.

The results for the STT-06 system with post-processing are given in Table 3, and, once again, as for the STT-05 system, both post-processing schemes drastically increase the number of segments in the STT output. The results show that the PFM-NGRAM scheme achieves the best TER and BLEU scores, with gains over the STT baseline of 1.9% (TER) and 1.1% (BLEU) for the BC task, and 1.8% (TER) and 1.2% (BLEU) for the BN task. It is felt that the PFM-NGRAM post-processing scheme achieves better performance than the SSP-NGRAM scheme primarily because it enables the acoustic and lexical information to be combined within a single decoding framework, rather than being processed separately in sequence.

In order to highlight some of the main contrasts that these various systems provide, Table 4 indicates the performance differences between the STT-05, STT-06, and PFM-NGRAM post-processed STT-06 systems in terms of CER, TER, and BLEU.

Test	Post-Processing	#Segs	CER	TER	BLEU
bcm	STT-05	836	32.4	79.89	7.40
	STT-06	836	18.1	76.47	9.01
	+ PFM-NGRAM	2259	18.1	**74.58**	**10.06**
bnm	STT-05	934	17.4	74.37	11.60
	STT-06	934	12.3	73.21	12.39
	+ PFM-NGRAM	2298	12.3	**71.38**	**13.57**

Table 4. TER and BLEU scores for the STT-05 and STT-06 systems for the bcm and bnm test sets.

For the BC task, improvements to the STT component which result in CER gains of 14.3% produce corresponding TER and BLEU gains of 3.4% and 1.6% respectively. However, if the PFM-NGRAM post-processing stage is incorporated into the pipeline then total TER and BLEU improvements of 5.3% and 2.7% can be gained over the STT-05 baseline. The basic pattern is similar for the BN task, though the CER gains are smaller (mainly for the training data reasons discussed earlier). In the BN case, the pipeline that uses STT-06 output and PFM-NGRAM post-processing improves the TER and BLEU scores by 2.9% and 1.9% over the STT-05 baseline.

6. CONCLUSION

This paper has discussed various interactions between STT and SMT in a transcription-translation pipeline that converts Mandarin audio into English text. Two Mandarin STT systems have been contrasted, and it has been shown that the STT-06 system achieves CER gains of 14.3% for the BC task and 5.1% for the BN task compared to the baseline STT-05 system. When TER and BLEU scores are obtained for the (un-post-processed) translated output of these two systems, improvements of 3.4% (TER) and 1.6% (BLEU) for the BC task and 1.2% (TER) and 0.8% (BLEU) for the BN task are observed.

In addition, it has been shown that breaking the STT segmentations down into smaller sentence-like units can also help to improve the TER and BLEU scores. Two post-processing schemes have been discussed, and the PFM-NGRAM scheme consistently produces the best results, achieving TER and BLEU gains of up to 5.3% and 2.7% respectively.

The results presented in this paper quantify the extent to which SMT performance benefits from improvements in STT performance for a specific experimental set-up. Further, the impact of different post-STT segmentation strategies has been explored, and it has been shown that SMT systems produce more accurate output if they take sentence-like segmentations as input.

7. REFERENCES

[1] R. Sinha, M.J.F. Gales, D.Y.Kim, X.A.Liu, K.C.Sim, and P.C.Woodland, ‘The CU-HTK Mandarin Broadcast News Transcription System’ *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006

[2] H.Y. Chan and P.C. Woodland, ‘Improving Broadcast News Transcription by Lightly Supervised Discriminative Training’, *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[3] B. Xiang, L. Nguyen, X. Guo, and D. Xu, ‘The BBN Mandarin Broadcast News Transcription System’, *Proc. of EuroSpeech* 2005

[4] E. Shriberg and A. Stolke, ‘Prosody Modeling for Automatic Speech Recognition and Understanding’, Mathematical Foundations of Speech and Language Processing, IMA Volumes in Mathematics and Its Applications, Vol 138, Springer-Verlag, New York 105-114, 2004

[5] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. C. Woodland, and M. Harper, ‘Structural Metadata Research in the EARS Program’, *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005

[6] M. Tomalin and P.C.Woodland, ‘Discriminatively Trained Gaussian Mixture Models for Sentence Boundary Detection’ *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006

[7] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, ‘Measuring the readability of automatic speech-to-text transcripts’, *Proc. of EuroSpeech*, 2003

[8] A. Stolke, E. Shriberg, D. Hakkani-Tür, and G. Tür ‘Modeling The Prosody of Hidden Events for Improved Word Recognition’, *Proc. of EuroSpeech* 1999

[9] http://www.languageweaver.com

[10] M. Snover, B.J. Dorr, R. Schwartz, J. Makhoul, L. Micciulla, R. Weischedel, ‘A Study of Translation Edit Rate with Targeted Human Annotation’, *Proc. of Association for Machine Translation in the Americas*, 2006.

[11] K. Papineni, S. Roukos, T. Ward, and W. Zhu, ‘BLEU: A Method for Automatic Evaluation of Machine Translation’ Tech. Rep. RC22176 (W0109-022), IBM Research Division, 2001