IMPROVEMENTS IN ACCURACY AND SPEED IN THE HTK BROADCAST NEWS TRANSCRIPTION SYSTEM

P.C. Woodland[†‡], *J.J. Odell*[‡], *T. Hain*[†], *G.L. Moore*[†], *T.R. Niesler*[†], *A. Tuerk*[†] *& E.W.D. Whittaker*[†]

[†]Cambridge University Engineering Department,Trumpington Street, Cambridge CB2 1PZ, UK.
Email: {pcw,th223,glm20,trn,at233,ewdw2}@eng.cam.ac.uk
[‡]Entropic Ltd., Compass House, 80-82 Newmarket Road, Cambridge, CB5 8DZ, UK.
Email: {pcw,jo}@entropic.co.uk

ABSTRACT

This paper describes a number of recent improvements to the HTK Broadcast News Transcription System. Changes to the system include the use of more acoustic training data; use of cluster-based variance normalisation and vocal tract length normalisation; the use of interpolated language models and enhanced adaptation using a full variance transform. These changes produce an reduction in word error rate of 13%. A simplified version of the system has also been constructed that runs in less than 10 times real-time and gives a 2.3% absolute higher error rate than the 300xRT full system.

1. INTRODUCTION

There is currently much interest in the automatic transcription of found speech and general audio sources. This paper takes as its starting point the HTK Broadcast News Transcription System used in the 1997 DARPA/NIST Hub4 evaluation. Changes to the system are described which improve both transcription accuracy and allow a simplified version to operate in less than 10 times real-time on commodity hardware.

The paper first briefly outlines the 1997 system and then describes experiments in the use of vocal tract length normalisation and variance normalisation; increased acoustic training data; improved language models to combine data from different types of source rather than simply pooling the texts; and full-variance transform adaptation. The changes to increase the speed of operation of the system include careful choice and optimisation of the platform, decoding parameters and model sets together with a significantly faster segmentation step. For both the improved accuracy and improved speed system, the results of the system on the 1998 DARPA Hub4 evaluation are presented.

2. OVERVIEW OF 1997 SYSTEM

The HTK Broadcast News system runs in a number of stages. The input audio stream is first segmented and classified as wide or narrow band; a first pass recognition is performed using triphone HMMs and a trigram language model (LM) to get an initial transcription for each segment; the speaker gender for each segment is found; the segments are clustered and unsupervised maximum likelihood linear regression (MLLR) transforms [2, 6, 7] estimated for each segment cluster. This is followed by generating a lattice for each segment using the adapted triphone models with a bigram LM, expanding these lattices using a word 4-gram interpolated with a category trigram LM, and performing iterative lattice rescoring and MLLR adaptation with a set of quinphone HMMs. Finally hypotheses from the quinphone and triphone stages are combined to form the final output. System details can be found in [11].

The data segmentation [4] aims to generate acoustically homogeneous speech segments and discard non-speech portions such as pure music. It uses a set of Gaussian mixture models to classify the data as either wideband speech, narrow-band speech, pure music, or speech and music and then any pure music is discarded. A gender dependent phone recognition stage then generates a stream of gender labelled phone units. Using a clustering procedure and a set of smoothing rules the final segments to be processed by the decoder are generated.

For recognition, each frame of input speech is represented by a 39 dimensional feature vector that consists of 13 MF-PLP cepstral parameters (including c_0) and their first and second differentials. Cepstral mean normalisation (CMN) is applied over each segment. The cross-word triphone HMMs were estimated using 72 hours of broadcast news data (BNtrain97) and contained 6684 states each with 12 Gaussians while the quinphone models used 8180 states and 16 Gaussians per state. The HMMs were initially trained on all the wide-band analysed training data and then narrow-band sets were estimated by using a version of the training data with narrow-band analysis (125-3750Hz). Gender dependent versions of each were made and the reduced bandwidth models used for data classified as narrow band.

The system uses a 65k word dictionary, based on LIMSI'93 dictionary, with many additions and corrections. The 1997 system used N-gram language models trained on 132 million words of broadcast news texts, the LDC-distributed 1995 newswire texts, and the transcriptions from BNtrain97 (LMtrain97). This corpus was used to estimate both word N-grams and a category N-gram based on 1000 automatically generated word classes [8].

The final hypothesis combination uses word-level confi-

dence scores based on an N-best homogeneity measure. These are used with the NIST ROVER [1] program to produce the final output. The system gave an error rate of 15.8% in the 1997 DARPA/NIST Hub4 evaluation.

3. IMPROVING ACCURACY

The 1997 system described above formed the basis of the improved accuracy "Hub" 1998 system. The 1997 Hub4 evaluation test set (BNeval97) was used for system development.

3.1. Increased Acoustic Training Data

We first compared the effect of using the additional training data available for the 1998 evaluation which contained a further 71 hours of usable data (total set denoted BNtrain98). Versions of the acoustic models (triphones and quinphones) used in the 1997 system were trained with BNtrain98 (16 mixture components per state for both triphones and quinphones). Experiments with no adaptation (or cluster-based normalisation) showed that the word error rate (WER) was reduced by up to 0.9% absolute. However when MLLR adaptation and VTLN were applied (see below) the WER gain was reduced to 0.4% absolute. However it was noted that the gains were across all speech conditions with the largest gains being for non-native speakers. Similarly, when using quinphone models with MLLR a gain of 0.5% in WER was achieved with increased training data.

3.2. Vocal Tract Length Normalisation

We applied the vocal tract length normalisation (VTLN) approach developed for conversational telephone speech transcription [5] to broadcast news. To give robust warp factor selection, the technique uses feature variance normalisation.

The VTLN and the variance normalisation is done on a segment cluster basis. We found an overall improvement in WER with cluster-based variance normalisation of 0.3% absolute and a further 0.6% absolute by applying VTLN in both training and testing without adaptation. However with mean and variance MLLR adaptation the separate beneficial effect of variance normalisation and VTLN is much reduced.

Data Type	BNtrain97		BNtrain98	
	non-VTLN	VTLN	non-VTLN	VTLN
F0	10.4	9.8	9.8	9.5
Overall	17.5	16.8	16.7	16.4

Table 1: %WER on BNeval97 for different training/normalisation. Mean+variance MLLR is used with the 1997 4-gram LM and triphone HMMs. Non-VTLN systems use segment-based cepstral mean normalisation.

A summary performance on BNeval97 (MLLR adapted triphones) for increased training data and the use of VTLN is shown in Table 1 for the prepared studio speech (F0) and overall error rates. Furthermore, in line with the triphone figures, the overall gain for 1998 trained MLLR adapted quinphone models was 0.4% absolute due to VTLN.

3.3. Language Modelling

For the 1998 system, the additional transcriptions from the 1998 acoustic training were available. Furthermore we processed additional transcriptions of broadcast news texts supplied by Primary Source Media (from late 1996, 1997 and early 1998) so that we had a total of 190MW of such data available. Finally, we decided to use a different (though similarly sized) portion of newspaper texts covering 1995 to February 1998 (about 70MW in total). All these sources excluded data from the designated test epochs. This corpus was denoted LMtrain98.

Previously we have constructed LMs by simply pooling the texts and weighted the acoustic data transcription counts. Here, as others have done previously (e.g. [10]), we experimented with building separate language models for each of the 3 data sources and then interpolating the language models. For efficiency and ease of use in decoding, a model merging process was employed using tools supplied by Entropic Ltd., that gives a similar effect to explicit model interpolation but saves run-time computation and storage. The interpolation weights were chosen to minimise perplexity.

Data Type	LMtrain97 pooled	LMtrain98 pooled	LMtrain98 interp.
F0	11.0	10.4	10.4
F1	18.7	18.0	17.1
Overall	18.4	17.7	17.2

Table 2: %WER on BNeval97 for different trigram LMs with VTLN unadapted triphone HMMs with either pooled data or (merged) interpolated LMs.

The effect of using three different LMs on BNeval97 with VTLN data and 1998 unadapted triphone HMMs is shown in Table 2. Note that the LMtrain98 models also used a revised vocabulary which reduced the out-of-vocabulary rate on BNeval97 by about 0.1%. It can be seen that the new training corpus reduces the WER by 0.7% absolute and a further 0.5% absolute reduction was obtained by using a merged interpolated language model. The merged interpolated models gave most improvement on the spontaneous speech portions of the data (F1 data). Later experiments with adapted quinphone models showed that a total improvement of 0.9% absolute was gained from using the the new LM data and estimation procedure.

3.4. Full Variance Transform / SAT

The basic adaptation approach in our system remains MLLR for both means and variances [3]. In addition, for the quinphone stage of iterative unsupervised adaptation, the effect of a single full variance (FV) transform [3] was investigated.

This FV transform was used with, for the wideband data, HMMs estimated with a single iteration of speaker adaptive training (SAT) [9] to update the mean parameters. The effect of these changes is shown in Table 3. It can be

Data Type	-FV +SAT	+FV +SAT	+FV -SAT
F0	9.0	8.7	8.8
Overall	14.6	14.3	14.4

Table 3: %WER on BNeval97 for BNtrain98 VTLN MLLR adapted quinphones using the 1998 fgintcat LM with / without a full-variance (FV) transform and SAT mean estimated models.

seen that the FV transform reduces the error rate by 0.3% absolute with SAT training contributing 0.1%. The word error rate on BNeval97 of 14.3% (including FV and SAT) represents a 13% reduction relative to the same stage of the 1997 evaluation system [11].

3.5. 1998 Evaluation Results

The accuracy improvements shown above were included in the HTK "Hub" system for the 1998 DARPA/NIST Broadcast News Evaluation. The system now operated in six passes: the first pass (P1) uses gender independent triphones without VTLN; gender dependent VTLN triphones are used in P2 and lattices with MLLR-adapted triphones generated in P3 with a bigram language model and then expanded to use the (merged) 4-gram interpolated with a category trigram. Passes 4-6 use quinphone models gradually increasing the number of MLLR transforms and also the FV transform.

Stage	LM	MLLR /FV	% WER	
			Overall	F0
P1	tg	N/N	19.9	10.9
P2	tg	N/N	17.5	10.2
P3	bg	1/N	19.1	11.9
P3	fgintcat	1/N	15.3	8.7
P4	fgintcat	1/N	14.9	8.3
P4	fgintcat	1/Y	14.2	8.0
P6	fgintcat	4/Y	14.2	8.0
ROVER	fgintcat	4/Y+1/N	13.8	7.8

Table 4: Word error rates for each stage of the 1998 HTK broadcast news evaluation system (also P4 FV contrast). Only P1 uses gender independent non-VTLN HMMs. P1 to P3 use triphones and P4-P6 quinphones.

The results (over the complete 1998 evaluation set) for each of these stages, together with additional contrasts, is shown in Table 4. There is a 12% reduction in error by using gender dependent models and VTLN (P1 to P2) and a further 7% from using MLLR. This is a rather smaller MLLR gain than previously observed which we believe is due to the more extensive input data normalisation. There is a 7% gain moving from adapted triphones to adapted quinphones: most of which (5%) was due to the full variance adaptation. This gain from the FV transform was rather greater than observed on the BNeval97 data.

4. IMPROVING SPEED

A faster version of the system capable of running in less than 10xRT (rather than 300xRT) was required. A simpler architecture using fewer decoding passes was needed and the benefits of each of the various stages was examined. The initial decoding pass allowing adaptation of gender dependent models increased accuracy significantly. However further iterations and the use of quinphone models produced relatively little benefit at considerable computational cost. It was also decided that VTLN would not be included due to the complexity of the current implementation.

An initial version of the system was developed for the 1998 TREC 7 SDR system which was a development of the 1997 system. It used tighter pruning with the same triphone acoustic and 4-gram language models but made no use of the quinphone models. On a Sun Ultra 2300 the execution time of the HTK SDR system was roughly: segmentation 4xRT; first pass decode 10xRT; clustering and adaptation 1xRT; second pass decode/lattice generation 30xRT; and language model application 3xRT. Although these times were dominated by decoding, a system operating in under 10xRT requires faster operation of all components.

4.1. Platform Choice

The Sun Ultra was compared to Intel-based compute platforms and it was found that a 450MHz Intel Pentium II Xeon running Windows NT gave the best decoding performance, in part due to the processor itself (released 18 months later than the 300MHz Sun Ultra) and in part due to the highly tuned Intel compiler available for NT. This improved decoding times by about a factor of two. The final system used this platform for most compute intensive parts but also used a 450MHz Pentium II running Linux and the Ultra 2300 for some other stages as convenient.

4.2. Segmentation and Classification

The segmenter for the 10xRT was a simplified version of the full system segmenter but used a faster classification phase (fewer passes of adaptation) and smaller gender dependent phone models and improved decoding. It was found that the revised segmentation approach ran in about 1xRT and for the first pass system on BNeval97 data gave a 0.3% absolute increase in error rate.

4.3. Decoding Parameters

The recogniser used for both the first and second pass is the LVX decoder which forms part of the version 2 release of Entropic's HAPI programming interface. This is a single pass time synchronous decoder incorporating cross word triphones and a trigram language model into a single lattice generating pass. The 4-gram language model is applied to the generated lattice to produce the single most likely hypothesis for each segment.

A number of experiments were conducted to determine optimal decoding parameters for this data. It was found that while the real-time factor for individual segments is highly variable, the processing time for the second pass decode was much more closely correlated with the first pass time than the first pass time is with segment duration. This allows the second pass parameters to be set

dependent on the computational requirements estimated from the time taken for the first pass and the relationship between first and second pass times derived from experiments on other data.

4.4. Results

A first set of models (HMM1) was used for the initial decoding pass and consisted of 8893 distinct models sharing 4011 tied states each represented by a 16 component Gaussian mixture distribution. This was used with a 60k trigram merged language model trained on the LMtrain98 data. The output of the first pass was then expanded to a 4-gram LM. The results of the first pass decode (running in about 2xRT) is given in Table 5 for both the BNeval97 and the BNeval98 sets along with the 1996 Hub4 development set, BNdev96ue (which contains complete broadcast episodes and is substantially more difficult).

Data Type	% Word Error Rate		
	BNdev96ue	BNeval97	BNeval98
F0	11.1	12.9	12.1
Overall	26.8	21.4	21.2

Table 5: First pass results from 10xRT system

The next stage used gender dependent models and consisted of 13428 distinct models sharing 5606 tied states each represented by a 20 component mixture Gaussian. Gender dependent and bandwidth dependent versions of these models were produced. Simplified MLLR transforms were produced for each set of clustered segments and the second pass decode run again using a trigram language model and expanded to use the final 4-gram language model interpolated with the category trigram from the full system. The gender determination, clustering and adaptation took about 0.5xRT and the second pass decoding about 5xRT. The final results of the 10xRT system are shown in Table 6 and show that across test sets there is an overall reduction in WER of about 5% absolute by including the second pass.

Data Type	% Word Error Rate		
	BNdev96ue	BNeval97	BNeval98
F0	8.6	9.4	9.7
Overall	22.1	15.8	16.1

Table 6: Final results for various test sets for the full 10xRT system

On BNeval98, the 10xRT system had a word error rate 2.3% absolute (16% relative) higher than the full improved system (which ran in approximately 300xRT) and the same error rate on the 1997 evaluation data as the full system from a year earlier [11]. The 10xRT system had the lowest overall word error rate for such systems in the 1998 DARPA evaluation.

5. CONCLUSIONS

This paper has described recent progress in improving both the accuracy and speed of the HTK broadcast news transcription system. The accuracy of the system improved by 13% relative over the 1997 system. A 10xRT version of the system is as accurate as the full 1997 HTK system and can be used for fairly accurate transcription of large amounts of broadcast news data.

Acknowledgements

This work is in part supported by an EPSRC grant on "Multimedia Document Retrieval" reference GR/L49611 and by a grant from DARPA.

6. REFERENCES

1. Fiscus, J.G. (1997) A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347-354, Santa Barbara.
2. Gales M.J.F. & Woodland P.C. (1996). Mean and Variance Adaptation Within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249-264.
3. Gales M.J.F. (1997). Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. Technical Report, CUED-F-INFENG TR.291, Cambridge University Engineering Dept.
4. Hain T, Johnson S.E., Tuerk A., Woodland P.C. & Young S.J. (1998) Segment Generation and Clustering in the HTK Broadcast News Transcription System. *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pp. 133-137.
5. Hain T., Woodland P.C., Niesler T.R. & Whittaker E.W.D. (1999). The 1998 HTK System for Transcription of Conversational Telephone Speech. *Proc. ICASSP'99*, pp. 57-60, Phoenix.
6. Leggetter C.J. & Woodland P.C. (1995). Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. *Proc. ARPA Spoken Language Technology Workshop*, pp. 104-109. Morgan Kaufmann.
7. Leggetter C.J. & Woodland P.C. (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech & Language*, Vol. 9, pp. 171-185.
8. Niesler T.R., Whittaker E.W.D & Woodland P.C. (1998). Comparison of Part-Of-Speech and Automatically Derived Category-Based Language Models for Speech Recognition. *Proc. ICASSP'98*, pp. 177-180, Seattle.
9. Pye, D. & Woodland P.C. (1997). Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition. *Proc. ICASSP'97*, pp. 1047- 1050, Munich.
10. Wegmann S., Scattone F., Carp I., Gillick L., Roth R. & Yamron J. (1998) Dragon Systems' 1997 Broadcast News Transcription System. *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne.
11. Woodland P.C., Hain T., Johnson S.E., Niesler T.R., Tuerk A., Whittaker E.W.D. & Young S.J. (1998). The 1997 HTK Broadcast News Transcription System. *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pp. 41-48, Lansdowne.