The 1998 HTK Broadcast News Transcription System: Development and Results

P.C. Woodland, T. Hain, G.L. Moore, T.R. Niesler, D. Povey, A. Tuerk & E.W.D. Whittaker

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
e-mail: {pcw,th223,glm20,trn,dp10006,at233,ewdw2}@eng.cam.ac.uk

ABSTRACT

This paper presents the development of the HTK broadcast news transcription system for the November 1998 Hub4 evaluation. Relative to the previous year's system The system a number of features were added including vocal tract length normalisation; cluster-based variance normalisation; double the quantity of acoustic training data; interpolated word level language models to combine text sources; increased broadcast news language model training data; and an extra adaptation stage using a full-variance transform. Overall these changes to the system reduced the error rate by 13% on the 1997 evaluation data and the final system had an overall word error rate of 13.8% for the 1998 evaluation data sets.

1. Introduction

Significant progress in the accurate transcription of broadcast news data has been made over the last few years so that we are now at a point where such systems can be used for a variety of tasks such as audio indexing and retrieval. However there is still much interest in reducing the error rate of such systems further which will increase the potential for further applications as well as establishing techniques for the accurate transcription of general audio material.

The HTK Broadcast News Transcription System used in the 1997 DARPA/NIST Hub4 evaluation had an overall word error rate of 15.8%. This paper describes a number of experiments with, and developments of, that system. Some of these were included in 1998 HTK Hub4 evaluation system.

The main areas of development that were used in the 1998 evaluation system were the use of vocal tract length normalisation on a segment cluster basis and cluster-based variance normalisation; the use of an increased quantity of acoustic training data from about 70 hours to 140 hours; the use of interpolated word level language models to combine data from different types of source rather than simply pooling the texts; the use of more broadcast news language model training data; and an extra acoustic adaptation stage using a full-variance transform to supplement the normal mean and variance MLLR transform. Other experiments which didn't lead to overall word error rate reductions include discriminative training using the frame discrimination method and use of the soft-clustering technique.

The paper is arranged as follows. We first give details of the broadcast news data used in the experiments, then give an outline of the overall system used in the 1997 evaluation. The subsequent sections give the details of a number of experiments that we performed in system development. This is followed by a description of, and the results from, the 1998 Hub4 evaluation system. The full recognition results from the various stages of operation are included.

2. Broadcast News Data

This section describes the various data sets that have been used in the experiments reported in the paper.

The baseline acoustic corpus available in 1997 used recorded audio from various US broadcast news shows (television and radio). This amounted to a total of 72 hours of usable data (BNtrain97). This data was annotated to ensure that each segment was acoustically homogeneous (same speaker, background noise condition and channel). The LDC released a further tranche of data of similar size in 1998 (about 71 hours of data). This was similarly transcribed at the speaker turn level but didn't distinguish between background conditions which meant that marked training segments were no longer necessarily homogeneous. The combined set of 1997 and 1998 data is denoted BNtrain98.

Focus	Description
F0	baseline broadcast speech (clean, planned)
F1	spontaneous broadcast speech (clean)
F2	low fidelity speech (typically narrowband)
F3	speech in the presence of background music
F4	speech under degraded acoustical conditions
F5	non-native speakers (clean, planned)
FX	all other speech (e.g. spontaneous non-native)

Table 1: Broadcast news focus conditions.

System development mainly used the 1997 Hub4 evaluation data, BNeval97. BNeval97 was taken from a number of sources broadcast in October/November 1996 and was presented to the system as a single 3 hour file. The 1998 evaluation data, BNeval98, consisted of two 1.5 hour data sets:

the first drawn from a similar epoch as the 1997 data and the second drawn from June 1998. The evaluation results are presented for each of the NIST "focus" conditions which are shown in Table 1.

Focus Cond	Proportion of data	
	BNeval97	BNeval98
F0	45.0%	30.6%
F1	20.0%	19.3%
F2	16.1%	3.4%
F3	5.1%	4.3%
F4	4.9%	28.2%
F5	2.3%	0.7%
FX	6.3%	13.5%

Table 2: Proportion of test data of different audio type

The proportion of data of each type in BNeval97 and BNeval98 is given in Table 2. It can be seen that there is a rather different distribution data type between the two sets: particularly for F0, F2, F4 and FX.

3. Overview of 1997 system

The HTK Broadcast News system runs in a number of stages. The input audio stream is first segmented; 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, 7, 8] 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 [16].

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 to type (wideband speech, narrow-band speech, pure music, 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 (including c_0) MF-PLP cepstral parameters and their first and second differentials. Cepstral mean normalisation (CMN) is applied over each segment. The triphone HMMs were estimated using BNtrain97 and contained 6684 decision-tree clustered states [17], each with 12 Gaussians per state 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. Narrow-band sets were estimated by using a version of the training data with narrow-band analysis (125-3750Hz), and gender dependent versions of each were made. The reduced bandwidth models are used for data classified as narrow band.

The system uses the LIMSI 1993 WSJ pronunciation dictionary augmented by pronunciations from a TTS system and hand generated corrections for a 65k word vocabulary. 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 [6, 10, 11].

The final hypothesis combination uses word-level confidence scores based on an N-best homogeneity measure. These are used with the NIST ROVER program [1] to produce the final output.

4. System Developments

The architecture of the 1997 system described above formed the basis of the 1998 system. During system development the BNeval97 test set was used.

4.1. Increased Acoustic Training Data

We first compared the effect of using the additional training data in the BNtrain98 set. 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, using quinphone models with MLLR a gain of 0.5% in WER was achieved with increased training data.

We also did some experiments that used automatic segmentation of the extended training data to try and ensure that the segments used in training were acoustically homogeneous but this provided no additional improvements.

4.2. Vocal Tract Length Normalisation

We have previously worked on robust vocal tract length normalisation (VTLN), most recently in the context of conversational telephone speech transcription[5].

We use a maximum likelihood technique to select the best data warp factor via a parabolic search. It is important when comparing the warped data likelihoods to properly take into account the effect of the transformation. We have done this implicitly by performing variance normalisation on the data. 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.

A summary performance on BNeval97 (MLLR adapted triphones) for increased training data and the use of VTLN is shown in Table 3. Furthermore, in line with the triphone figures, the overall gain for 1998 trained MLLR adapted quinphone models was 0.4% absolute due to VTLN.

Data Type	BNtrain97		BNtrain98	
	non-VTLN	VTLN	non-VTLN	VTLN
F0	10.4	9.8	9.8	9.5
F1	17.3	16.4	16.8	16.3
F2	21.7	20.6	21.2	20.2
F3	28.6	29.3	27.8	30.0
F4	21.2	20.0	20.1	19.7
F5	24.4	23.8	21.4	20.0
FX	32.7	32.7	30.9	30.9
Overall	17.5	16.8	16.7	16.4

Table 3: %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.

4.3. 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. [15]), 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 interplotation 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
F2	22.9	21.6	19.9
F3	32.0	30.0	28.6
F4	22.1	22.1	22.2
F5	23.5	22.3	22.7
FX	33.1	32.5	31.4
Overall	18.4	17.7	17.2

Table 4: %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 4. 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 a 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. 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.

4.4. Full Variance Transform / SAT

The basic adaptation approach in our system remains MLLR for both means and variances [2]. 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) [14] to update the mean parameters. The effect of these changes is shown in Table 5. It can be seen that the FV transform reduces the error rate by 0.3% absolute with SAT training contributing 0.1%. The word error rate

Data Type	-FV +SAT	+FV +SAT	+FV -SAT
F0	9.0	8.7	8.8
F1	13.7	13.6	13.8
F2	17.6	17.1	17.1
F3	25.9	26.5	26.4
F4	17.8	17.2	17.4
F5	20.2	17.5	17.8
FX	27.9	27.2	27.1
Overall	14.6	14.3	14.4

Table 5: %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.

on BNeval97 of 14.3% (including FV and SAT) represents a 13% reduction relative to the same stage of the 1997 evaluation system [16].

4.5. Frame Discrimination

We have recently experimented with discriminative training of large vocabulary systems and using the frame discrimination (FD) technique [13]. FD is related to maximum mutual information estimation (MMIE), but uses all Gaussians (rather than all words) in the system to model confusion data. We developed a fast implementation technique to make FD training on large HMM sets practical and on the WSJ/NAB task FD gives similar reductions in word error rate (about 5% relative) to lattice-based MMIE with a much smaller computational cost.

Our experiments with FD on broadcast news data show that overall we get very similar results to maximum likelihood training, although the training procedure gives a sizeable improvement in the FD criteria. We therefore did not include FD modelling in the 1998 evaluation system.

4.6. Soft Clustering

The soft-clustering technique developed at JHU [9] had shown worthwhile reductions in word error rate on the Switchboard corpus and we performed a preliminary evaluation on Broadcast News data. The technique works by increasing the number of Gaussian components in each state distribution while not increasing the overall number by increased Gaussian sharing so that the strict context to "tied-state" relationship given by decision tree state-clustering [17] is not enforced.

Initially we used bandwidth independent, gender independent triphones to evaluate the technique and under these conditions it gave a 1% absolute reduction in WER. However, when bandwidth dependent, gender dependent models with variance normalisation and MLLR adaptation were used, there was no WER advantage and hence soft clustering was not used in the 1998 evaluation system.

5. 1998 DARPA Evaluation System

This section describes the HTK system used in the 1998 evaluation. The system takes the 1997 system and includes the additional acoustic training data in BNtrain98; cluster-based normalisation and VTLN; the revised language modelling data and build procedure and full variance adaptation with SAT training.

5.1. Language Models

The word N-grams were trained by interpolating (and merging) component LMs trained on the acoustic transcriptions, the broadcast news texts and the newspaper texts. The resulting LMs had 5.6 million bigrams, 9.9 million trigrams and 7.4 million 4-grams. The category-trigram used 1000 automatically derived word classes and was trained using LMtrain98. Category bigrams and trigrams were added only if the leave-one training set likelihood improved and the final category model contained 0.85 million bigrams and 9.4 million trigrams.

The 65k wordlist was chosen by combining the word frequency lists from the different LM training sources with suitable weightings and choosing the most frequent words for which we already had pronunciations.

	BNeval97	BNeval98_1	BNeval98_2
OOV rate	0.43%	0.38%	0.40%
tg pplex	145.2	140.4	157.1
fg pplex	131.6	127.8	143.1
fgintcat pplex	128.6	125.2	139.8

Table 7: OOV rate and perplexities of the 1998 evaluation LMs. Perplexities shown for trigram (tg), 4-gram (fg) and word 4-gram interpolated with category trigram (fgintcat).

The out-of-vocabulary (OOV) rate and perplexity of these language models on BNeval97 and the two halves of the BNeval98 set is shown in Table 7. It was noted that compared to the use of the 1997 language models all OOV rates had been reduced slightly, the most being by 0.1% on BNeval98_2. Furthermore the 1998 4-gram language model gave a constant 15% improvement (over all test sets) in perplexity over the equivalent model used in the 1997 evaluation.

Stage	LM	MLLR /FV	% Word Error							
			Overall	F0	F1	F2	F3	F4	F5	FX
P1	tg	N/N	19.9	10.9	20.5	29.6	20.2	20.6	26.0	34.8
P2	tg	N/N	17.5	10.2	17.9	26.5	19.1	17.2	24.7	30.9
P3	bg	1/N	19.1	11.9	20.4	27.6	24.2	18.6	24.3	30.5
P3	tg	1/N	16.2	9.5	17.4	22.4	18.8	16.0	20.4	27.3
P3	fg	1/N	15.5	8.8	16.6	22.9	18.6	15.1	20.0	26.8
P3	fgintcat	1/N	15.3	8.7	16.3	22.6	18.1	14.8	21.3	26.5
P4	fgintcat	1/N	14.9	8.3	15.7	21.5	16.8	14.8	20.4	26.2
P4	fgintcat	1/Y	14.2	8.0	15.2	20.2	16.4	14.1	16.6	24.7
P6	fgintcat	4/Y	14.2	8.0	15.4	20.3	16.5	14.0	16.6	24.6
ROVER	fgintcat	4/Y+1/N	13.8	7.8	15.1	20.1	15.8	13.6	16.6	24.1

Table 6: 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.

5.2. Decoding Passes

The overall decoding process proceeds as for the 1997 system, but with a couple of additional stages. The first pass (P1) uses gender independent triphone HMMs to get an initial transcription with a trigram LM. This transcription is used for both gender selection as well as VTLN warp selection for each segment cluster. Gender dependent VTLN models are then used (P2) to provide a revised transcription which is used to estimate global mean and variance MLLR transforms for each cluster. These adapted models are then used to generate lattices (P3/bg) which are expanded to use the 4-gram word LM interpolated with the category-based trigram model (P3/fgintcat).

The system then uses quinphone models (VTLN/SAT trained) and MLLR with an additional FV transform to process the data (P4). This stage is repeated twice more while increasing the number of MLLR transforms (P5/P6). The confidence-annotated output of P6 is combined with P3/fgintcat output with ROVER.[1]

5.3. System Performance

The results (over the complete 1998 evaluation set) for each of these stages, together with additional contrasts, is shown in Table 6. 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 6% gain from employing the category trigram and 4-gram over the trigram alone, and 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.

Finally, the effect of the automatic segmentation procedure on the BNeval98 set was investigated. On BNeval97 we had found that automatic segmentation had produced very similar overall accuracy to manually defined segments. However on BNeval98 the automatic segmenter faired more poorly. The WER for a first pass with wideband models was 0.7% absolute higher with the automatic segments than with the manual segments. This poorer performance was also reflected in the number of frames assigned to multiple speaker segments: 1.6% for BNeval97 but 4.3% for BNeval98.

[1] Before ROVER combination an alignment pass was run to get exact word timings. Due to the effects of automatic segmentation this process reduces the WER by about 0.1% absolute.

6. Conclusion

This paper has described the development and performance of the 1998 HTK broadcast news transcription system. A number of improvements to the systems accuracy have been described and, as in previous years, the system gave the lowest error rate overall on the main F0 focus condition. While the system produces good results it is computationally expensive: a companion paper [12] discusses a version of the system that runs in less than ten times real-time on commodity hardware.

7. Acknowledgements

This work is in part supported by an EPSRC grant on "Multimedia Document Retrieval" reference GR/L49611 and by a grant from DARPA. Entropic Ltd. supplied software to aid in decoding and language model estimation. X. Luo of JHU supplied code to help with the soft-clustering experiments.

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. Kneser R. & Ney H. (1993). Improved Clustering Techniques for Class-Based Statistical Language Modelling. *Proc. Eurospeech'93*, pp. 973-976, Berlin.
7. 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.
8. 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.
9. Luo X. & Jelinek F. (1998) Non-reciprocal Data Sharing in Estimating HMM Parameters. JHU Center for Language and Speech Processing, Research Note No. 32.
10. Martin S., Liermann J. & Ney H. (1995). Algorithms for Bigram and Trigram Clustering. *Proc. Eurospeech'95*, pp. 1253-1256, Madrid.
11. 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.
12. Odell J.J., Woodland P.C. & Hain T. (1999) The 1998 CUHTK-Entropic 10xRT Broadcast News Transcription System. *Proc 1999 DARPA Broadcast News Transcription and Understanding Workshop*, Dulles.
13. Povey D. & Woodland P.C. (1999) Frame Discrimination training of HMMs for Large Vocabulary Speech Recognition. *Proc. ICASSP'99*, pp. 333-336, Phoenix.
14. Pye, D. & Woodland P.C. (1997). Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Speech Recognition. *Proc. ICASSP'97*, pp. 1047- 1050, Munich.
15. 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.
16. 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.
17. Young S.J., Odell J.J. & Woodland P.C. (1994). Tree-Based State Tying for High Accuracy Acoustic Modelling. *Proc. ARPA Human Language Technology Workshop*, pp. 307-312. Morgan Kaufmann.