THE DEVELOPMENT OF THE 1996 HTK BROADCAST NEWS TRANSCRIPTION SYSTEM

P.C. Woodland, M.J.F. Gales, D. Pye & S.J. Young

Cambridge University Engineering Department,
Trumpington Street,
Cambridge, CB2 1PZ, UK

ABSTRACT

This paper describes our efforts in extending a large vocabulary speech recognition system to handle broadcast news transcription. Results using the 1995 DARPA H4 evaluation data set are presented for different front-end analyses and for the use of unsupervised model adaptation using maximum likelihood linear regression (MLLR). The HTK system for the 1996 H4 evaluation is then described. It includes a number of new features compared to previous HTK large vocabulary systems including decoder-guided segmentation, segment clustering, cache-based language modelling, and combined MAP and MLLR adaptation. The system makes multiple passes through the data and the detailed results of each pass are given. The overall word error rate obtained by the 1996 evaluation system was 27.5%, and a bug-fixed version reduced this to 26.6%.

1. INTRODUCTION

Large vocabulary continuous speech recognition (LVCSR) systems have traditionally been developed for read speech with a close talking microphone. Recently there has been increasing interest in using such systems in less restricted audio conditions such as for speech with high background noise and multiple microphone conditions; for transcription of conversational speech over the telephone and for transcription of broadcast news.

Broadcast news transcription poses a number of challenges for large vocabulary transcription systems. The data in broadcasts is not homogeneous and includes a number of data types for which current speech recognition systems have high error rates. A typical news broadcast may include data from different speech styles (read, spontaneous and conversational); native and non-native-speakers; high bandwidth and low bandwidth channels either with or without background music or other background noise.

Recently we have begun working on the problem of broadcast news transcription with a number of aims: to generalise the capabilities of our system for transcription of a general audio stream; to complement our other work in conversational and spontaneous speech; to provide a speech front-end to a broadcast news database system that has recently been developed in Cambridge [1]; and to allow us to participate in the 1996 ARPA Hub 4 evaluation which concerns the transcription of US television and radio news programmes.

Previously [11, 10] we have developed a 3-stage approach to recognition of unknown channel data. First the data is classified as to broad type, an appropriate "approximately-matched" HMM system is used followed by fine tuning the models on the actual test data using maximum likelihood linear regression (MLLR) [5, 6, 2] adaptation. For general audio transcription we are extending this approach using several basic HMM sets, and adding components for classification and clustering of audio segments into single speaker and audio type "sessions" within a broadcast.

This paper begins with a brief overview of the HTK LVCSR system. Since the broadcast news task often requires that a recogniser operates with poorly-matched acoustic models some experiments to test the robustness of the front-end parameterisation are then reported. Some preliminary experiments on the DARPA 1995 Hub 4 Marketplace radio data are described and it is shown that a combination of a robust front end analysis and MLLR adaptation can provide fairly good performance even when based on HMMs trained only from wideband clean acoustic data.

Finally the HTK system for the 1996 DARPA H4 evaluation is described. The system was developed by taking our HMM systems trained on the Wall Street Journal dataset and adapting them to the different acoustic conditions found in broadcast news data using the supplied acoustic training data. The detailed performance of this system on the 1996 evaluation data is given.

2. HTK LVCSR SYSTEM OVERVIEW

This section gives an overview of the standard HTK LVCSR system. The system uses state-clustered, cross-word mixture Gaussian context-dependent acoustic models and a back-off N-gram language model. More details of the system can be found in [9, 11].

In the standard system, each speech frame is represented by a 39 dimensional feature vector that consists of 12 mel frequency cepstral coefficients, normalised log energy along with the first and second differentials of these values. Cepstral mean normalisation (CMN) is applied. We have also investigated the use of PLP-based parameterisations [4] for improved robustness (see Sec. 3 and [11]).

The system uses the LIMSI 1993 WSJ pronunciation dictionary for pronunciations. This is augmented by pronunciations from a TTS system and hand generated corrections. The HMMs are cross-word context-dependent and use decision tree state clustering [13]. The context for clustering can either be a single phone (triphone context) or use longer range contexts (e.g. quinphones).

The standard gender independent triphone model set (the HMM-1 set of [9]) is trained on the 36,493 sentences from SI-284 WSJ0+1 data set and has 6,399 speech states, with each state having a 12 component Gaussian mixture output distribution. More detailed acoustic models can also be used in lattice-rescoring mode. In particular a set of models which depend on the neighbouring 2 phones and the locations of word boundaries has been used (the HMM-2 set of [9]). The HMM-2 set has 9354 speech states, each state characterised by a 14 component mixture Gaussian distribution.

Models are adapted to new speakers and environments using MLLR. Sets of transformation matrices for the Gaussian means [5], and optionally variances [2], are estimated to increase the likelihood of the adaptation data. If only a small amount of adaptation data is available, or very robust transform estimation is required, a single global transformation matrix may be used. If more data is available a regression class tree [6] can be used to define a set of transformations for the HMM set.

The HTK LVCSR system uses time-synchronous decoders that can either operate in a single pass or can be used to produce or rescore word lattices [7, 9] which compactly store multiple sentence hypotheses.

3. 1995 HUB4 EXPERIMENTS

In this section we describe some initial experiments using the 1995 Hub 4 evaluation data. The aim was to calibrate the difficulty of the broadcast news transcription problem and determine the effectiveness of both different front-end analyses and unsupervised adaptation on this data.

All experiments reported in this section were based on the HMM-1 model set with the 65k trigram language model used in the 1995 HTK H3 system [10]. The decoder was run with fairly tight pruning beamwidths and so some search errors will have occurred.

3.1. 1995 H4 Data

The 1995 Hub 4 evaluation data [8] consisted of portions (one complete show and 2 "heads" and 2 "tails") from 5 episodes of the NPR "Marketplace" business news radio programme transmitted during August 1995. NIST had also made available 10 complete "Marketplace" shows that could be used for training purposes, although none of this material has been used in the experiments presented here. For the November 1995 Hub 4 evaluation, only episode boundaries were given as side information and hence systems had to perform their own data segmentation and labelling of the audio data type.

To simplify the experiments reported here, we have used the segmentation boundaries, and where appropriate the segment type labels, provided by NIST for the evaluation data. This means that, for instance, no pure music segments are included[1]. The NIST transcriptions label each segment with a talker identifier; the presence (BM) or absence (NM) of background music and whether the speech signal is a full 8kHz bandwidth signal (FB) or of reduced telephone bandwidth (RB). Each segment contains just one audio type and data from a single speaker. Segments range in length from less than 1 second to several minutes.

In the experiments in the following sections, the results are given for the three audio conditions with a significant amount of data (NM/FB, BM/FB and NM/RB) along with the overall word error rate. In each case the official transcriptions and mapping files were used in scoring.

3.2. H4 Front End Comparison

In this section the environmental robustness of two front-end parameterisations are compared with the aim of selecting a robust front-end for the broadcast news task.

The standard HTK V2.0 MFCC front-end was used to produce baseline performance figures. We have previously found that cepstral parameters based on a perceptual linear prediction (PLP) [4] speech parameterisation were more robust to mismatched environments, although results were somewhat mixed. However we had found on other data that a modified form of PLP using the the MFCC filter-bank (MF-PLP) analysis is more effective than the use of the standard PLP filterbank, and so MF-PLP was compared to standard MFCC analysis.

Test Data Subset	Front-End Type	
	MFCC	MF-PLP
NM/FB	31.2	26.7
BM/FB	43.9	41.2
NM/RB	65.3	58.5
Overall	41.3	36.4

Table 1: % word error rate for MFCC and MF-PLP parameterisations on Nov'95 H4 data

The recognition results for the two analyses for different au-

[1] Segmentation/classification experiments that we have performed indicate that using Gaussian mixture models for different audio types enables pure music segments to be detected with a high degree of accuracy

dio types is given in Table 1. Overall there is 12% reduction in error rate using MF-PLP with the largest error rate reduction (14%) comes with the NM/FB data and the smallest reduction (6%) from the BM/FB data. It is clear that the MF-PLP front-end analysis provides significant performance gains under the mismatched conditions found in the H4 data when using models trained on clean speech. It would be expected that much smaller gains would be achieved with MF-PLP if the test and training data were more closely matched.

3.3. Unsupervised Adaptation

To try and reduce the mismatch between the test data and the models we applied 2 iterations of MLLR adaptation in transcription mode to the MF-PLP system. For the purposes of adaptation each Marketplace episode was split into a number of sessions with each session containing a single speaker and a single audio type. If the session contained less than 10 seconds of data then no adaptation was performed—this applied to less than 1% of the test data.

Both iterations of MLLR used block-diagonal transforms and only updated the Gaussian mean parameters. The first iteration used a global transformation and then the second iteration used multiple adaptation classes chosen using a regression class tree.

Test Data Subset	Adaptation Classes		
	None	Global	Multiple
NM/FB	26.7	24.8	22.3
BM/FB	41.2	32.8	31.3
NM/RB	58.5	40.7	36.5
Overall	36.4	29.8	27.0

Table 2: % word error rates for MF-PLP with unsupervised MLLR adaptation on 1995 H4 data.

The results of these two adaptation passes along with the unadapted MF-PLP error rates are shown in Table 2. Overall the use of a global transform reduces the error rate by 18%. The largest gains come from the most severely mismatched conditions: NM/RB improves by 30% and BM/FB by 20%, while the NM/FB data improves by just 7%. The second MLLR iteration improves the error rate by a further 9% with again the largest improvement being for the NM/RB data (10%).

Since the adaptation experiments used known segment boundaries and labels rather than an automatic system, it is impossible to directly compare the recognition results to those obtained in the 1995 Hub 4 evaluation. However, the figures show that the approach is effective even when no acoustic or language model training data from the broadcast news domain is available.

4. 1996 DARPA EVALUATION

In this section the 1996 Hub 4 task is discussed. The HTK system for the 1996 Hub 4 evaluation is described and the results of each stage of the system are given.

4.1. 1996 Hub 4 Data

The data for the evaluation consisted of U.S. television and radio broadcast news programmes recorded "off-air". For the primary partitioned evaluation (PE), the data was presegmented into portions that were acoustically homogeneous: i.e. a single speaker in a single audio condition. These segments varied in length from under one second to several minutes.

The labelling for each segment provided a fairly detailed description of the data. For convenience, the audio was divided into a number of "focus conditions" labelled F0 to F5 and FX. These are listed in Table 3.

Focus	Description
F0	baseline broadcast speech (clean, planned)
F1	spontaneous broadcast speech (clean)
F2	low fidelity speech (wideband/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 3: 1996 H4 focus conditions

A number of broadcast news shows transmitted prior to June 30th 1996 were recorded and labelled by the LDC for acoustic training. The evaluation data (broadcast in September 1996) contained some material from programmes used for training. In total there was about 35 hours of labelled broadcast news acoustic training data. When analysed by focus condition the amount available varied from 12 hours for F0 to 16 minutes for F5. Due to numerous transcription problems with this data, we had only six weeks to work with the training data before we actually ran the evaluation.

The LDC also supplied commercially available transcriptions of various broadcast news programmes produced by Primary Source Media, Inc. covering the period from January 1992 to April 1996 and containing approximately 132 million words of text.

The evaluation data consisted of 4 half-hour segmented broadcast news programmes (two television, two radio). The proportions of the test data in each focus condition was as follows: F0 29.7%; F1 32.7%; F2 8.7%; F3 7.0%; F4 9.1%; F5 1.5 % and FX 11.4%.

Processing Stage	LM Type	% Word Error							
		Overall	F0	F1	F2	F3	F4	F5	FX
Prelim. 1	tg	33.4	23.0	31.5	39.8	30.3	39.5	28.1	58.7
Prelim. 2	tg	31.1	21.3	30.1	38.7	29.9	33.9	27.1	52.7
Lattice Gen.	bg	34.1	25.2	33.9	41.2	32.5	36.4	27.8	52.4
Lattice Gen.	fg	29.4	20.7	29.4	34.6	25.0	32.4	23.7	49.2
HMM-2 (noadapt)	fg	30.3	20.7	27.9	37.3	25.8	36.5	25.8	55.3
Global HMM-2	fg	27.5	19.0	26.4	32.7	23.7	29.3	21.1	50.7
Multiple HMM-2	fg	27.7	19.1	26.6	33.1†	23.6†	29.1†	21.7†	51.0†
Multiple HMM-2	fg/cache	27.5†	18.7†	26.5†					

Table 4: % Word error rates on 1996 H4 evaluation data at various stages of processing. † denotes the system actually submitted for the evaluation.

4.2. 1996 HTK H4 System Overview

The overall style of processing adopted was broadly similar to that used by the 1995 HTK H3 system [10, 11]. However there were a large number of detailed differences including the use of the MF-PLP representation throughout. The system was run in multiple passes first starting with the most appropriate models available and at each stage using unsupervised test-data adaptation to refine the transcriptions.

After two "preliminary" passes through the data word lattices were generated. The preliminary passes and lattice generation used triphone models based on the HMM-1 set and adaptation used a global speech MLLR transform along with a separate silence transform. After lattices have been generated (using a bigram language model) they were expanded using a 4-gram and the HMM-2 models used. The HMM-2 models were initially adapted with a global transform and then a final pass run with more detailed adaptation. It was hoped that further adaptation/transcription passes could have been run with these models but the time available for the evaluation precluded this. For the F0 and F1 focus conditions, the lattices generated from the final stage were rescored with a cache language model.

4.3. Acoustic Model Training

For each focus condition a set of "initial" models were estimated using the HMM sets (both HMM-1 and HMM-2) trained on the secondary channel Wall Street Journal data (as in [10]). These sets were then adapted for each focus condition using mean and variance MLLR with the broadcast news training data for each focus.

For F2 there was a mix of narrow-band and wideband data, so the F2 training data was automatically classified as either narrowband or wideband using a simple high/low frequency energy ratio approach, and two sets of HMMs were adapted. The test data for F2 was also automatically labelled as narrow or wide band and the appropriate set of models used for each F2 segment.

Only a very small amount of the training data was labelled as F5 (planned, clean, non-native) and so this data was combined for both training and testing with the portion of FX that was labelled as spontaneous, clean, non-native data. In training, the F0 adapted models were further adapted using all of the clean non-native data.

For the other portions of FX, a "global" model was used which was formed by adapting the WSJ secondary channel models on all the broadcast news data. A number of other approaches to tackling the variety of data types present in FX were considered but lack of development time precluded a full investigation.

For the focus conditions with, what was judged to be a reasonable amount of training data (F0, F1 and F4), the HMM-2 MLLR-adapted models were further adapted using forward-backward MAP [3]. On the development data it was found that applying MAP after MLLR yielded between a 1% and 3% reduction in word error for these conditions.

4.4. Decoder Guided Segmentation

As described above, the data was pre-segmented as to audio type but there was no limit the on length of individual segments. For several processing stages (including lattice generation and manipulation), it is more convenient if the data contains segments no longer than 30s in duration.

For the first pass through the data, the recogniser was allowed to make "sentence-end" to "sentence-start" transitions midway through a segment. This information combined with the length of silences at these sentential transitions and other points in the segment (found from a forced alignment of the decoded output) was used to generate a new segmentation for the further decoding passes.

4.5. Segment Clustering

In order to perform test-data adaptation on broadcast news data when speaker identities are unknown, it is necessary to group segments that are "similar" so that sufficient data is available for robust unsupervised adaptation. For this purpose a within-focus-condition bottom-up segment clustering technique was adopted.

Each segment (before CMN was applied) is represented by its mean and variance and then segments are iteratively merged with the nearest segment group (as measured by a modified divergence measure) until all segment groups contain enough speech frames. The number of clusters is controlled by an occupation count threshold. Experiments on the development data showed that the system was relatively insensitive to the the cluster threshold and that this very simple clustering scheme could yield performance that is similar to that obtained using clusters based on known speaker identity.

4.6. Static Language Models

The evaluation system language model (LM) had a word list containing 65423 words chosen from the most frequent words in the broadcast news training texts, with the most frequent words in a number of other text corpora also added. There was an OOV rate of 0.74% on the evaluation data.

For the evaluation, bigram, trigram and 4-gram LMs were estimated by combining data from the LDC supplied broadcast news texts, the LDC 1995 newswire texts (non-financial and financial), the acoustic training data transcriptions (added twice) and the 1995 Marketplace transcriptions (added 3 times). The language models contained 6.9 million bigrams, 8.3 million trigrams and 8.6 million 4-grams.

Focus	proportion	% OOV	Perplexity	
Cond	of test %	rate	3-gram	4-gram
F0	29.7	1.6	188	172
F1	32.7	0.2	124	115
F2	8.7	0.8	145	135
F3	7.0	1.3	227	182
F4	9.1	0.5	123	111
F5	1.5	0.3	274	269
FX	11.4	0.3	164	157
Overall	100	0.74	154	141

Table 5: Hub 4 perplexities (1996 eval data)

The overall perplexity of the 4-gram LM was 141 and the trigram 154 on the evaluation data. The detailed perplexities and OOV rates, by focus condition are given in Table 5.

It was noted that these perplexities, although about 20% higher than typically observed on read newspaper texts, were also significantly below that of the development data.

4.7. Cache LM

A unigram and bigram cache model of the form used in [10] was interpolated with the static 4-gram language model for the F0 and F1 conditions. The cache was based on the output from the final acoustic pass and operated on a per-show, per-focus-condition basis. The cache includes future and previous words. Words with the same stem were also added to the unigram cache and common words were excluded.

4.8. System Results

Results from the system at various stages of processing are shown in Table 4. The table also contains the entry HMM-2 (noadapt) which provides a contrast for using the HMM-2 models with lattices with no test-data adaptation. Contrasting the first two passes shows that global test-data adaptation gives a 7% improvement in word error whereas for the HMM-2 models (using somewhat smaller clusters) yields a 9% improvement in word error. While there was a 14% reduction in word error moving from bigram to 4-gram this is rather smaller than we have previously observed and is probably in part due to lattice errors.

Overall the HMM-2 models gave about 6% fewer word errors. However the FX models performed rather more poorly and this was caused by a bug in the lattice pruning procedure which resulted in large lattices, such as those for FX, being incorrectly pruned.

Another somewhat surprising result was that, overall, the use of multiple adaptation classes for the HMM-2 models increased the word error rate by 0.2% absolute. This was also found to be due to a bug in the iterative adaptation procedure and hence the lowest error rate result for each focus condition was not necessarily submitted as output for the evaluation. For this type of task we have previously found it preferable to perform adaptation in a number of stages [12]. However for the evaluation run itself, there was not enough time to run these extra passes.

Finally the use of a cache language model reduced the error rate on F0 by 2% and on F1 by less than 0.5%. The lattices used were rather small and so there was rather limited scope for improvement.

4.9. Effect of Evaluation System Bugs

As mentioned above, after the evaluation it was discovered that there were two bugs in the system used. The first affected the quality of large lattices and had most impact on the FX condition. The second problem was in the application of multiple iterations of MLLR. A bug in a script caused the unadapted models to be used to find state-frame align-

Processing Stage	LM Type	% Word Error							
		Overall	F0	F1	F2	F3	F4	F5	FX
HMM-2/1 trans	fg	27.2	18.8	26.4	32.6	24.1	29.6	21.1	48.2
HMM-2/2 trans	fg	26.9	18.7	26.2	31.7	23.6	28.8	21.4	47.9
HMM-2/4 trans	fg	26.7	18.5	26.2	31.4	23.9	28.0	20.7	47.5
HMM-2/4 trans	fg/cache	26.6	18.1	26.2					

Table 6: % Word error rates for corrected system on 1996 H4 evaluation data.

ments rather than the previously adapted models which resulted in an increase in word error with the application of multiple transformations.

To correct both of these bugs the bigram lattices for all conditions were re-expanded to 4-g lattices and re-pruned using corrected software. In fact, the resultant lattices for many conditions were smaller than those actually used in the evaluation. Three iterations of MLLR were then run on these lattices with the HMM-2 model sets: the first used a global speech transform, the second pass a maximum of two speech transforms and the third pass a maximum of four transforms. This was the approach that we would have taken in the evaluation had time permitted. Finally the cache language model was again applied to the data from the F0 and F1 conditions.

From the results in Table 6 it can be seen that correcting the combined effect of the two bugs has reduced the overall error rate from 27.5% to 26.6%. By comparing the results of a single speech transform with those in Table 4, it can be seen that 1/3 of the extra errors were caused by the lattice pruning problem and that FX in particular was severely effected. Furthermore, the difference in error rate between the HMM-1 and HMM-2 results with a single speech transform is 7% with all focus conditions improving. Other than for FX, it can be seen that the performance on F4, F2, and to a smaller extent F0, has noticeably improved with the corrected system.

5. CONCLUSION

This paper has described our initial efforts to develop systems for broadcast news transcription. A number of new features have been added to our system and it has been shown that it is viable to adapt a system based on read speech using either supervised or unsupervised adaptation and obtain reasonable transcription accuracy.

6. ACKNOWLEDGMENTS

This work is in part supported by an EPSRC grant reference GR/K25380. Mark Gales is supported by a Research Fellowship from Emmanuel College, Cambridge. Rachel Morton and Rishi Nag helped with training data and dictionary preparation. Julian Odell gave assistance with decoders.

References

1. Brown M.G., Foote J.T., Jones G.J.F., Sparck-Jones K. & Young S.J. (1995). Automatic Context-Based Retrieval of Broadcast News. *Proc. ACM Multimedia'95*, pp. 35-43, San Francisco.
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. Gauvain J.L. & C.H. Lee (1994). Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Trans. SAP*, Vol. 2, No. 2, pp. 291-298.
4. Hermansky H. (1990). Perceptual Linear Prediction (PLP) Analysis for Speech. *J. Acoust. Soc. Amer.*, Vol. 87, pp. 1738-1752.
5. 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.
6. Leggetter C.J. & Woodland P.C. (1995). Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression. *Proc. Eurospeech'95*, pp. 1155-1158, Madrid.
7. Odell J.J., Valtchev V., Woodland P.C. & Young S.J. (1994). A One Pass Decoder Design For Large Vocabulary Recognition. *Proc. ARPA Human Language Technology Workshop*, pp. 405-410. Morgan Kaufmann.
8. Pallett D.S., Fiscus J.G., Garofolo J.S. & Przybocki M.A. (1996). 1995 Hub-4 "Dry-Run" Broadcast Materials Benchmark Tests. *Proc. DARPA Speech Recognition Workshop*, Harriman, New York.
9. Woodland P.C., Leggetter C.J., Odell J.J., Valtchev V. & Young S.J. (1995). The 1994 HTK Large Vocabulary Speech Recognition System. *Proc. ICASSP'95*, Vol. 1, pp. 73-76, Detroit.
10. Woodland P.C, Gales M.J.F., Pye D. & Valtchev V.(1996) The HTK Large Vocabulary Recognition System for the 1995 ARPA H3 Task. *Proc. DARPA Speech Recognition Workshop*, pp. 99-104, Harriman, New York.
11. Woodland P.C, Gales M.J.F., & Pye D. (1996) Improving Environmental Robustness in Large Vocabulary Speech Recognition. *Proc. ICASSP'96*, pp. 65-68, Atlanta.
12. Woodland P.C, Pye D. & Gales M.J.F. (1996) Iterative Unsupervised Adaptation Using Maximum Likelihood Linear Regression. *Proc. ICSLP'96*, pp. 1133-1136, Philadelphia.
13. 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.