The 1997 HTK Broadcast News Transcription System

P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk, E.W.D. Whittaker & S.J. Young

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
e-mail: {pcw,th223,sej28,trn,at233,ewdw2,sjy}@eng.cam.ac.uk

ABSTRACT

This paper presents the recent development of the HTK broadcast news transcription system. Previously we have used data type specific modelling based on adapted Wall Street Journal trained HMMs. However, we are now using data for which no manual pre-classification or segmentation is available and therefore automatic techniques are required and compatible acoustic modelling strategies must be adopted. A number of recognition experiments are presented that compare data-type specific and non-specific models; differing amounts of training data; the use of gender-dependent modelling and the effects of automatic data-type classification. Based on these experiments, the HTK system for the 1997 broadcast news evaluation was designed. A detailed description of this system is given which includes a class-based language modelling component. The complete system yields an overall word error rate of 22.0% on the 1996 unpartitioned broadcast news development test data and just 15.8% on the 1997 evaluation test set.

1. Introduction

The transcription of broadcast radio and television news 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 speech recognition systems trained on read speech corpora such as the WSJ corpus have high error rates. A typical news broadcast may include data of different speech styles (read, spontaneous and conversational); native and non-native speakers; high or low bandwidth channels either with or without background music or other background noise. Solving these problems will be of great utility in dealing with both the broadcast news problem and more general transcription of "found" speech.

We have previously investigated [15] the use of specific models for different audio conditions for the somewhat unrealistic situation where the data has been pre-segmented into homogeneous portions (same audio conditions and same speaker) and where the audio conditions associated with each segment are supplied to the system. That system was constructed using HMMs trained on the Wall Street Journal (WSJ) corpus as a base and then adapted to individual data types of broadcast news data using supervised maximum likelihood linear regression (MLLR) [7, 6, 3]. During recognition we used iterative unsupervised MLLR to adapt clusters of segments to the particular audio conditions. This system was shown to give good performance in the 1996 DARPA/NIST broadcast news partitioned evaluation (PE) [15].

Our current research has concentrated on the more general situation where information about data segmentation and type is not supplied to the recogniser (unpartitioned or UE data). To extend our previous approach to the UE case, it is necessary to first segment the data into homogeneous segments of differing data types as well as rejecting segments of data that contain no speech (e.g. background music). Furthermore given an automatic segmentation it is of interest to develop acoustic modelling techniques that do not rely on detailed, manually derived, data classifications.

The rest of the paper is arranged as follows. We first give details of the broadcast news data used in the experiments, and briefly describe our work on segment processing which splits the unpartitioned data stream into moderate length homogeneous segments. This is followed by an overview of the basic recognition architecture and a number of recognition experiments to determine the performance of the system. We compare the performance of acoustic data specific modelling and non-specific models on PE data; the effect of varying the amount of acoustic training data; the use of gender-dependent modelling; and the effects of two automatic segmentation algorithms on recognition performance.

Finally we describe the HTK transcription system used in the the 1997 broadcast news evaluation and give a detailed description of the system's performance on both the 1996 unpartitioned broadcast news development test data and the 1997 evaluation test set.

2. Broadcast News Data

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

For acoustic training a number of US broadcast news shows (both television and radio) transmitted prior to June 30th 1996 were recorded and labelled by the LDC. In total episodes from 11 different shows were present in the training data: ABC Nightline, ABC World News Now, ABC World News Tonight, CNN Early Edition, CNN Early Primetime News, CNN Headline News, CNN Primetime News, CNN The

World Today, CSPAN Washington Journal, NPR All Things Considered and NPR Marketplace. About 35 hours of transcribed data was made available in 1996. We made some corrections to these transcriptions and used them to estimate the HMMs described in [15]. This corpus will be referred to as BNtrain96. A further tranche of data of similar size was released in 1997 to form in total 72 hours of broadcast news training data. We also modified these transcriptions and tried to remove portions of the speech signal where two or more speakers were talking simultaneously. The 72 hour corpus is denoted BNtrain97. Each resulting segment in the training corpora was labelled by speaker and one of the audio "focus" conditions listed in Table 1.

Focus	Proportion of data	
Cond	BNdev96ue	BNeval97
F0	22.3%	45.0%
F1	30.5%	20.0%
F2	16.2%	16.1%
F3	6.2%	5.1%
F4	14.1%	4.9%
F5	2.7%	2.3%
FX	7.8%	6.3%

Table 2: Proportion of test data of different audio type

3. Segment Processing

The goal of the segment processing stages is to convert the continuous input audio stream into clusters of reasonably-sized speech segments. Ideally, each segment should be homogeneous (i.e. same speaker and channel conditions) and the segments should be grouped into clusters such that each cluster is sufficiently similar to share a single set of MLLR adaptation transforms. It is also desirable to remove as much of the non-speech from the input audio stream as possible. Details of these segment processing stages are given in a companion paper [5], but a brief overview is included here for completeness.

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 1: Broadcast news focus conditions.

For development test purposes, data broadcast in July 1996 from six shows (ABC Prime Time, CNN World View, CSPAN Washington Journal, NPR Marketplace, NPR Morning Edition and NPR The World) was used. The hand-partitioned development test data, BNdev96pe, with given segmentation and focus conditions contained extracts from all the shows while the unpartitioned data, BNdev96ue, contained data from the first four shows (about two hours of data). The data from an episode of NPR Marketplace is the only complete show that is common to both the BNdev96pe and BNdev96ue data sets.

Finally the 1997 evaluation data, BNeval97, contained extracts from 9 different shows and totalled about 3 hours of data. As for the development test, the evaluation data contained (different episodes of) shows that also occurred in the training set as well as some shows which were not present. Unlike BNdev96ue for which show boundaries are known, the BNeval97 data is presented to the system as a single 3 hour audio file.

Table 2 gives the proportions of the different audio types present in the BNdev96ue and BNeval97 data sets measured by the number of reference words assigned to each data category. Note that there is a significantly greater proportion of F0 data present in BNeval97 than in BNdev96ue and rather less F1 and F4.

Our approach to segment processing is first to classify the audio data into three broad categories: wide-band speech (S), narrow-band speech (T) and music (M). After rejecting the music, a gender-dependent phone recogniser is used to locate silence portions and gender change points [9] and after applying a number of "smoothing rules" the final segment boundaries are determined.

The initial audio classification uses 4 Gaussian mixture models: one for each of the required classes (S, T and M) plus a model for music and speech. Audio selected by this latter model is also labelled as (S) but its separate inclusion reduces the misclassification of speech as music. Each model was trained on data of the appropriate class extracted from the BNtrain97 data up to a maximum of three hours per model.

After an initial classification of the data, MLLR adaptation transforms were computed for each class and then the decoding was repeated. This adaptation was performed separately for each of the four shows and only for classes with at least 15 seconds of data. This approach gives approximately a 95% frame classification accuracy, and on the BNeval97 set is able to discard 70% of the non-speech material while only erroneously discarding 0.2% of the speech.

Segmentation and gender labelling is applied to both the narrow-band (T) and wide band (S) data using a phone recogniser which has 45 context independent phone models per

gender plus a silence/noise model. The output of the phone recogniser is a sequence of relatively short segments having male, female or silence tags. Silence segments longer than 3 seconds are classified as non-speech and discarded. Sections of male speech with high pitch are frequently mis-classified as female and vice versa. Hence, a number of heuristic smoothing rules are applied. For example, a male segment followed by a short female segment is merged to form a single male segment if the following segment is silence. These smoothing rules also ensure that segments with durations between one second and 30 seconds are created. This basic segmenter is referred to as S1. About 7% (by duration) of the data consists of segments containing more than one speaker when using the S1 segmenter.

Further improvements to the segmentation are effected using a clustering procedure in which all segments are clustered using a top-down covariance-based technique (see below). Segments which appear in the same leaf node and are temporally adjacent (ignoring intervening silences) are merged into a single segment. This process corrects many of the gender mis-classifications but results in long segments. The clustering is then repeated taking account of the inter-segment silences in order to obtain the final segmentation. This approach makes it impossible to distinguish between two consecutive speakers of the same gender unless they are separated by silence. However, since most segment boundaries have at least a short silence segment at the boundary, this does not cause severe degradation in performance. The segmenter integrating segment clustering is denoted S2 and it reduces the proportion of the data represented by multiple speaker segments to 2%. The frame error rate for gender labelling is 3-4%.

Finally the segments are clustered separately for each gender and bandwidth combination for use with MLLR adaptation. Two alternative clustering techniques have been evaluated. The first was a bottom-up method in which each segment is modelled by a single diagonal covariance Gaussian and segments are merged based on a furthest neighbour divergence-like distance measure. Cluster merging stops when the number of frames in the smallest cluster exceeds a threshold. This was the scheme that we used in our 1996 broadcast news evaluation system [15]. The second scheme represents segments by the covariance of the static and delta parameters and uses a hierarchical top-down clustering process in which each node of the hierarchy is split into a maximum of four child nodes. Segments are reassigned to the closest node using an arithmetic harmonic sphericity distance measure [1]. Splitting continues while a minimum occupancy count is exceeded in all clusters. At the end of the process, all segments which were too small to compute a full covariance robustly are assigned to the leaf node with the closest mean. These schemes were found to give similar performance and the bottom-up scheme was used for final segment clustering.

4. Recognition System Overview

This section gives an overview of the basic recognition architecture used for the experiments reported in Section 5. The system is a development of previous HTK large vocabulary recognisers (e.g. [13]).

Each frame of input speech is represented by a 39 dimensional feature vector that consists of 13 (including c_0) MF-PLP cepstral parameters [15] and their first and second differentials. Cepstral mean normalisation (CMN) is applied over a segment.

The system uses the LIMSI 1993 WSJ pronunciation dictionary augmented by pronunciations from a TTS system and hand generated corrections. Cross-word context dependent decision tree state clustered mixture Gaussian HMMs [16] are used with a 65k word vocabulary. The system uses a language model trained on 132 million words of broadcast news texts, the LDC-distributed 1995 newswire texts, and the transcriptions from BNtrain96.

As will be seen in Section 6, the full HTK system can operate in multiple passes and use quinphone HMMs, more complex language models and iterative unsupervised adaptation. However, for the initial experiments reported in Section 5, the decoder was run in a single pass using triphone models, a trigram language model and fairly tight beamwidths. We have found that using the full system with adaptation results in a 20-25% decrease in word error rate on broadcast news data.

5. Single-Pass Recognition Experiments

5.1. Data Specific Models and Extended Training Data

We first compared the performance of models which require knowledge of data type with condition independent models which are more suitable to automatically segmented data since fine classification is not required. Furthermore, it has previously been shown that data condition independent models can give surprisingly good performance [9, 4].

The data type specific models used WSJ secondary channel trained HMMs with 6399 speech states and were subsequently adapted to broadcast news (used in [15]). Two sets of condition independent models were trained: the BNtrain96 HMM-BN1 has 5628 states and the BNtrain97 HMM-BN2 set 6684 states. All models used 12 component mixture Gaussian distributions. In all cases gender independent models were used.

The results given in Table 3 for the BNdev96pe set show that the WSJ models are significantly improved by broadcast news adaptation (4% absolute). Perhaps more surprisingly the HMM-BN1 models give slightly better overall performance than the data specific WSJ adapted models. In par-

ticular, it can be seen that there is a large improvement for the HMM-BN1 models on the spontaneous speech conditions. Furthermore, doubling the amount of training data reduces the error rate by a further 2.1% absolute.

Data	HMM training			
Type	WSJ	WSJ adapt	BNtrain96	BNtrain97
F0	16.3	13.0	12.8	11.6
F1	35.2	31.8	28.5	26.2
F2	51.4	44.8	42.6	38.7
F3	36.4	32.7	35.3	32.8
F4	28.6	25.0	25.4	24.6
F5	28.6	23.8	27.1	24.8
FX	58.5	55.2	56.8	55.4
Avg.	36.0	32.0	31.7	29.6

Table 3: % Word error rates on BNdev96pe for different training conditions. Only the WSJ adapt set is data condition dependent.

Whilst the results shown in Table 3 are encouraging, they mask the separate effects on male and female speakers. Since two thirds of the broadcast news training and test data is from male speakers, there is a significant gender bias which isn't present in the WSJ models. The error rate on the female speakers in the test is 29.8% for the WSJ adapt models but is 33.3% for the HMM-BN1 models (and 31.1% for HMM-BN2). To try to improve the performance for female speakers we investigated gender dependent modelling.

5.2. Gender Dependent Modelling

Gender dependent versions of the HMM-BN2 set were created by splitting the BNtrain97 data according to gender and retraining the Gaussian means and mixture weights on the gender-specific data portions. These gender dependent models were then tested on data only from the corresponding gender (i.e. it is assumed that perfect gender determination is possible). As shown in Table 4, this gave a substantial increase in recognition performance (overall 1.4% absolute and 2.3% for female speakers) and appears to have largely mitigated the gender bias in the training data.

It should be noted that although the automatic gender classification yields 3-4% error, using a forced alignment with the above gender dependent models and making a likelihood based gender choice (based on a first pass recognition with GI models) yields a gender labelling frame error rate of 1-2% [5].

5.3. Automatic Segmentation/Classification

The effect of using the automatically derived segments from both the CMU segmenter described in [12] and the S1 and S2 segmenters described in Sec. 3 was evaluated on the BN-

Data	Model type and data type			
Type	GI / male	GI /fem	GD /male	GD / fem
F0	9.7	13.9	9.9	12.5
F1	25.3	29.0	24.3	28.1
F2	38.3	41.6	35.7	37.3
F3	25.3	39.4	24.6	36.5
F4	24.2	25.1	23.1	21.3
F5	25.7	24.1	25.9	23.0
FX	57.2	53.7	57.1	50.4
Avg.	28.8	31.1	27.8	28.8

Table 4: % Word error rates on BNdev96pe split by gender for gender independent (GI) and gender dependent (GD) versions of HMM-BN2 models.

dev96ue data. It should be noted that some of the data (that identified as pure music) is discarded by the S1/S2 segmenters while the CMU approach retains the entire data stream. As can be seen in Table 5 a), recognition performance improves with the S1 segmenter, particularly on F3 segments due in part to the removal of pure music, and S2 improves overall recognition performance further.

After the 1997 evaluation was complete, additional experiments for the CMU and S2 segmenters were run using the 1997 trigram language model and the HMM-BN2 models. The results of these experiments are shown in Table 5 b), which confirms the advantage of the S2 segmenter for all data types. It is expected that the S2 system will have a further advantage when speaker/environment adaptation is used due to the small amount of data it includes in multiple-speaker segments.

Finally the performance on BNeval97 of the S2 segmenter was compared to that using hand-partitioned segments given in the reference transcriptions. Again, the 1997 trigram LM and the HMM-BN2 acoustic model set were used. It can be seen from Table 6 that the overall degradation caused by the automatic segmentation is very small (on the development data the degradation is approximately 1% absolute). On some types of data it appears that the automatic segmentation actually gives superior recognition results for the BNeval97 data. For instance, for some F3 segments which include a gradual fade in/out of music during a sentence, automatic segmentation improves the results. This is because the automatic segmenter chooses pause points for segmentation, leading to better recognition results. The good recognition performance of the automatically segmented data is also a reflection of the fact that, for this data, only a few errors (less than 0.1%) are introduced by erroneously discarding speech.

Data Type	Segmentation Alg		
	CMU Segs	S1 Segs	S2 Segs
F0	12.3	12.1	12.1
F1	27.1	27.0	25.9
F2	39.7	39.5	38.3
F3	38.8	32.5	33.4
F4	27.9	27.1	26.0
F5	30.4	33.0	30.6
FX	69.5	65.5	66.7
Overall	30.1	29.2	28.6

a)

Data Type	Segmentation Alg	
	CMU Segs	S2 Segs
F0	13.3	13.0
F1	21.6	20.8
F2	35.6	34.9
F3	34.1	32.4
F4	26.2	25.7
F5	29.0	27.5
FX	50.9	46.8
Overall	23.9	23.0

b)

Table 5: % Word error rates for different segmentation algorithms using the gender independent HMM-BN2 model set on a) BNdev96ue and b) BNeval97

Data Type	Segmentation	
	Manual Segments	S2 Segments
F0	12.9	13.0
F1	20.2	20.8
F2	35.5	34.9
F3	34.2	32.4
F4	25.0	25.7
F5	27.5	27.5
FX	45.6	46.8
Overall	22.9	23.0

Table 6: % Word error rates on the BNeval97 for automatic and hand generated segmentations.

5.4. Telephone Bandwidth Models

The S1/S2 segmenters also classify data as narrow-band or wide-band. A narrow-band model version of HMM-BN2 (HMM-BN2T) was trained using single-pass retraining from the HMM-BN2 set and a reduced bandwidth data analysis (125Hz to 3.75kHz) of the BNtrain97 dataset. The performance of the HMM-BN2T models was investigated for the data which had been automatically classified as narrow band.

The use of these models improved performance on F2 data using the S2 segmentation of the BNdev96ue to 35.9% error (from 38.3%) and reduced the overall error rate to 28.3% (from 28.6%). However a much more dramatic effect was observed for the S2 segmentation of the BNeval97 data. Here the word error rate of F2 data was reduced to 26.2% (from 34.9%) and the overall error rate to 21.4% (from 23.0%). It should be noted however that the advantage of using separate telephone bandwidth models decreases significantly when also using MLLR adaptation.

6. 1997 DARPA Evaluation System

This section describes the HTK system used in the 1997 evaluation. The system uses the modelling techniques described in previous sections with the addition of more complex acoustic and language models and multiple passes of unsupervised MLLR adaptation. In this respect the operation is similar to previous HTK evaluation systems [14, 15]. New features, apart from those discussed above, include an interpolated word-based and class-based language model and the combination of different output stages based on confidence annotation.

6.1. Decoding Stages

The overall decoding process is as follows. Firstly the data is segmented using the S2 segmenter described in previous sections. A first pass decoding is performed using the HMM-BN2 and HMM-BN2T gender independent models with a trigram language model. This stage (pass1) provides a putative transcription which is used both to select an appropriate gender dependent model set for subsequent use and also to provide an initial transcription for MLLR adaptation. The MLLR processing at this stage uses one global block-diagonal mean and diagonal variance speech transform per set of clustered segments. The adapted gender dependent model sets are then used with a word bigram language model to generate word lattices and the resultant bigram lattices (pass2/bg) are expanded in stages with a word trigram (pass2/tg), a word 4-gram (pass2/fg) and finally an interpolated language model combining the word 4-gram with a category trigram model that uses 1000 automatically generated classes (pass2/ic).

The best word string, found using an A* search of the interpolated lattices, is then used as supervision for global MLLR with a set of quinphone HMM models (HMM-BN3). These HMMs were trained in a similar manner to HMM-BN2 but the decision tree clustering process takes into account quinphone context and also word boundary locations. The HMM-

BN3 set has 8180 speech states each modelled with a 16 component Gaussian mixture. Gender dependent and bandwidth dependent versions of these models were used for each segment as appropriate. The system includes 3 passes through the data with MLLR-adapted quinphone models. At this stage the search procedure is constrained by the previously generated word lattices incorporating interpolated language model scores. The first quinphone pass (pass 3) uses a global mean and variance MLLR speech transform; the second quinphone pass (pass 4) uses up to two MLLR speech transforms and the final pass (pass 5) uses up to four transforms. Word lattices are also produced by pass 5.

The final lattices can be further processed in two ways. Firstly they can be re-scored using an unsupervised unigram cache model. Alternatively the output from the pass5 lattices can be combined with that from the pass2/ic lattices to form the final recognition output.

The following subsections describe in more detail first the language models used and then describe the hypothesis combination process which uses the NIST ROVER program. Finally complete recognition results are presented using the evaluation set-up for both the BNeval97 data and BNdev96ue.

6.2. Language Models

At various stages bigram, trigram and 4-gram word-based language models were used. These were trained on the 132 million words of the LDC broadcast news training texts, the transcriptions of the BNtrain97 data (added twice), the 1995 newswire texts (both financial and non-financial), and transcriptions for the 1995 train, dev and eval Marketplace transcriptions which were added to the training corpus three times. The text-processing for this data expanded a number of abbreviations and corrected some common spelling errors. Word-based language models using Katz backoff were built and contained 6.9 million bigrams, 8.4 million trigrams and 8.6 million 4-grams.

The category language model used 1000 automatically generated word classes chosen to maximise the training set likelihood based on word bigram statistics [8, 10, 11]. The categories and the trigram category model were built using the broadcast news training texts, the acoustic training data and 1995 Marketplace transcriptions. Bigrams and trigrams were only added to the category model if they improved the training set leave-one-out likelihood. The category trigram model contained 803k bigrams and 7.1 million trigrams.

In use the word 4-gram lattices were rebuilt by interpolating the category trigram language model (weight 0.30) with the word 4-gram (weight 0.70). These weights were found by performing N-best rescoring experiments on a preliminary version of the evaluation system with BNdev96ue data.

Lang Model	Perplexity	
Type	BNdev96ue	BNeval97
bigram	243	240
trigram	172	159
4-gram	161	147
cat-model	246	238
4g+cat	149	137

Table 7: Word level perplexities

The perplexities for the various language models on the BNdev96ue and BNeval97 (filtered) reference transcriptions are given in Table 7. The category model alone has a similar perplexity to the word bigram, however when interpolated with the word 4-gram it reduces the 4-gram perplexity by about 8%. The OOV rate on both these test sets using the filtered transcriptions was about 0.5%.

6.3. Hypothesis Combination

Whilst on average the overall error rate produced by the quinphone models is a little better than the triphone models we have observed that the systems often make rather different errors. Therefore we made an initial attempt at combining the final quinphone output (pass 5) with the best triphone stage (pass 2/ic).

The hypotheses from each stage were first annotated with word confidence scores and then the NIST ROVER program [2] was used to combine the annotated hypotheses. This program uses dynamic programming based string alignment to obtain a word correspondence for all words in the hypotheses and then examines the correspondence pairs and chooses the word with the highest confidence score. The confidence scores were generated using an N-best homogeneity measure found using the 1000-best hypotheses from the lattices generated at the appropriate stage. A decision tree pruned using 10-fold cross-validation was used to convert the N-best homogeneity scores to confidence probabilities. This tree was trained using a preliminary version of the system run on the BNdev96ue data.

6.4. System Performance

Detailed performance for each stage of the evaluation system is shown in Table 8 for both the BNdev96ue set and the BNeval97 data set. The better overall performance of the BNeval97 set seems to be mainly due to the much greater proportion of well-recognised F0 data present (see Table 2). It should also be noted that all results presented in this paper for the BNdev96ue set use the 1996 NIST scoring conventions, while the results for BNeval97 use the 1997 conventions.

Stage	HMMs Type	LM	MLLR	% Word Error							
				Overall	F0	F1	F2	F3	F4	F5	FX
pass1/tg	BN2/gi	tg	N	28.1	11.9	25.8	35.5	33.3	25.8	30.6	66.6
pass2/bg	BN2/gd	bg	1	27.9	13.0	27.9	33.8	32.0	24.4	28.0	61.7
pass2/tg	BN2/gd	tg	1	24.6	10.3	24.0	29.8	26.7	22.1	25.8	59.5
pass2/fg	BN2/gd	fg	1	24.0	9.9	22.9	29.5	26.7	21.5	23.5	59.3
pass2/ic	BN2/gd	ic	1	23.6	9.0	22.6	29.2	26.4	20.8	25.6	60.1
pass3/ic	BN3/gd	ic	1	23.0	9.4	22.0	27.8	26.5	20.0	22.1	58.5
pass4/ic	BN3/gd	ic	2	22.8	9.3	21.8	27.3	26.5	19.8	22.1	58.1
pass5/ic	BN3/gd	ic	4	22.7	9.3	21.8	27.1	26.7	19.7	21.9	57.8
+cache	BN3/gd	ic/cache	4	22.6	9.3	21.9	27.0	26.6	19.6	21.9	57.5
ROVER	BN2/3/gd	ic	1/4	22.0	8.5	21.4	26.2	25.4	19.0	22.3	57.1

a)

Stage	HMMs Type	LM	MLLR	% Word Error							
				Overall	F0	F1	F2	F3	F4	F5	FX
pass1/tg	BN2/gi	tg	N	21.4	13.0	20.8	26.2	32.4	25.7	27.5	43.2
pass2/bg	BN2/gd	bg	1	21.3	13.5	22.0	25.9	30.6	25.4	27.7	38.0
pass2/tg	BN2/gd	tg	1	18.0	10.7	17.7	21.8	29.9	21.4	26.0	34.8
pass2/fg	BN2/gd	fg	1	17.3	10.3	16.9	21.4	28.5	21.1	24.1	32.8
pass2/ic	BN2/gd	ic	1	16.8	9.9	16.3	20.8	28.4	20.7	24.1	32.1
pass3/ic	BN3/gd	ic	1	16.4	10.0	15.4	20.4	27.6	19.9	23.9	30.7
pass4/ic	BN3/gd	ic	2	16.2	9.8	15.5	20.1	27.7	19.6	24.2	30.1
pass5/ic	BN3/gd	ic	4	16.2	9.9	15.4	20.0	27.9	19.3	24.2	29.9
+cache	BN3/gd	ic/cache	4	16.2	9.9	15.4	20.1	27.9	19.4	24.1	29.9
ROVER	BN2/3/gd	ic	1/4	15.8	9.4	15.2	19.5	26.9	19.4	22.1	29.1

b)

Table 8: Word error rates for each stage of the 1997 HTK broadcast news evaluation system on both a) BNdev96ue and b) BNeval97.

The use of global MLLR and gender dependent models reduces the first-pass error by between 12-16% with a word trigram, and it should be noticed that larger reductions occur for the more challenging data types. The word 4-gram gives a 3% reduction in word error and a further reduction of similar size results from the use of the interpolated category language model.

The quinphone models with a global MLLR improve the error rate by 3% over triphone models using global MLLR. It is somewhat surprising to find that while iterative MLLR gives further small gains on the BNdev96ue data (about 1.5%), barely measurable gains were found on BNeval97. Again the use of the unigram cache slightly improved the error rate on BNdev96ue but did not help on BNeval97.

Finally the use of hypothesis combination using ROVER reduced the error rate by a further 3% to give a 15.8% overall word error rate on BNeval97 and 22.0% on BNdev96ue. It was noted that the confidence scores associated with the combined hypotheses had a normalised cross entropy of 0.173 for BNdev96ue and 0.179 for BNeval97. The reduction in error rate for the combined quinphone and triphone output results in more than doubling the overall gain from the use of quinphone modelling.

7. Conclusion

This paper has described the development and performance of the 1997 HTK broadcast news transcription system. The system uses a data segmentation and classification scheme which incorporates clustering. The use of HMMs that are independent of detailed data type fits well with automatic data segmentation and classification and yields at least as good performance as data type specific models. The system includes an interpolated language model and we have performed some preliminary investigations on hypothesis combination. The final system yielded the lowest overall word error rate in the

1997 DARPA broadcast news evaluation by a statistically significant margin.

8. Acknowledgements

References

1. Bimbot F. & Mathan L. (1993). Text-Free Speaker Recognition using an Arithmetic Harmonic Sphericity Measure. *Proc. Eurospeech'93*, pp. 169-172, Berlin.

2. Fiscus, J.G. (1997) A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). *Proc. IEEE ASRU Workshop*, pp. 347-352, Santa Barbara.

3. Gales M.J.F. & Woodland P.C. (1996). Mean and Variance Adaptation Within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249-264.

4. Gauvain J.L., Lamel L., Adda G.& Adda-Decker M. (1997). Transcription of Broadcast News. *Proc. Eurospeech'97*, pp. 907-910, Rhodes.

5. 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*, Virginia.

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. Kneser R. & Ney H. (1993). Improved Clustering Techniques for Class-Based Statistical Language Modelling. *Proc. Eurospeech'93*, pp. 973-976, Berlin.

9. Kubala F., Hubert J., Matsoukas S., Nguyen L., Schwartz R. & Makhoul J. (1997). Advances in Transcription of Broadcast News. *Proc. Eurospeech'97*, pp. 927-930, Rhodes.

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. To appear *Proc. ICASSP'98*, Seattle.

12. Siegler M.A., Jain U., Raj B. & Stern R.M. (1997) Automatic Segmentation, Classification and Clustering of Broadcast News Data. *Proc. DARPA Speech Recognition Workshop*, pp. 97-99, Chantilly, Virginia.

13. 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.

14. 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. ARPA Speech Recognition Workshop*, Harriman, New York.

15. Woodland P.C., Gales M.J.F., Pye D. & Young S.J. (1997) The Development of the 1996 Broadcast News Transcription System. *Proc. DARPA Speech Recognition Workshop*, pp. 73-78, Chantilly, Virginia.

16. 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, March 1994*, pp. 307-312. Morgan Kaufmann.