EXPERIMENTS IN BROADCAST NEWS TRANSCRIPTION

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

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK
e-mail: {pcw,th223,sej28,trn,at233,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 experimenting with data for which no manual pre-classification or segmentation is available and therefore automatic techniques are required and compatible acoustic modelling strategies adopted. An approach for automatic audio segmentation and classification is described and evaluated as well as extensions to our previous work on segment clustering. 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. It is shown that robust segmentation into a small number of audio types is possible and that models trained on a wide variety of data types can yield good performance.

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 [8] 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 the audio conditions associated with each segment is 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) [4, 2]. 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) [8].

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, and then describe our work on segment processing (segmentation, classification and clustering) which splits the unpartitioned data stream into moderate length homogeneous segments. This is followed by an overview of the 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.

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. About 35 hours of transcribed data was made available in 1996. Some corrections to these transcriptions were made by us and used to estimate the HMMs described in [8]. 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	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 was used. The PE data, BNdev96pe, (given segmentation and focus conditions) contained extracts from all the shows while the unpartitioned data, BNdev96ue, contained data from four of the shows. The data from an episode of NPR Marketplace is the only complete show that is common to both the BNdev96pe and BNdev96ue data sets.

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.

Our approach to segment processing is to first 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 [5] and segment boundaries are determined. Finally, the resulting segments are clustered in preparation for adaptation.

3.1. Audio Classification

The initial audio classification uses Gaussian mixture models with 1024 mixture components and diagonal covariance matrices. Four models are used, 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. For the speech models, data was selected to ensure that each training segment contained at least 90% speech. Since the training (and test) data doesn't explicitly identify narrow bandwidth data, a simple classifier based on the ratio of energy above 4kHz to that from 300Hz to 4kHz was used to label all data segments as either wideband or narrowband.

For classification, each frame of data was labelled using a conventional Viterbi decoder with each of the four models in parallel. An additional insertion penalty was applied to the music model in order to control the misclassification of speech into music.

After an initial classification of the data, MLLR mean 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.

	Baseline	Adapted
Frame Accuracy	95.72 %	95.98 %
Frames Lost	0.72 %	0.48 %

Table 2: Overall audio classification accuracy and percentage loss of speech to discarded music class on the BNdev96ue data set.

Table 2 shows the overall audio classification accuracy on the four shows from BNdev96ue data measured as the percentage of audio frames correctly labelled and the percentage of audio frames which were incorrectly labelled as music and therefore erroneously discarded.

Table 3 shows a confusion matrix for the adapted models. Notice that although some of the data is labelled as noise (N), the classifier does not attempt to explicitly identify noise. Thus, noise is distributed amongst the recognition classes. This table shows that around 80% of the pure music and 15% of the noise is classified as music and discarded. Overall 63% of the non-speech is discarded with only 0.5% loss of speech data.

3.2. Segmentation and Gender Detection

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 when the following segment is silence. These smoothing rules also ensure that segments with durations between three seconds and 30 seconds are created.

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 then merged into a single segment. This process corrects many of the gender misclassifications 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 approximately 85 % of segments boundaries have at least a short silence segment at the boundary, this does not cause severe degradation in performance.

Table 4 summarises the performance of the segmentation and gender detection on the BNdev96ue set. As can be seen, the use of segment clustering improves both the speaker segmentation and the gender detection. For comparison, the segmentation given by the CMU software [6] distributed by NIST is also included in Table 4 along with the reference hand derived segmentation. It can be seen the segmentation algorithms described here give a substantial reduction in the number of multiple speaker segments compared to the CMU approach. This should lead to better recognition performance since subsequent cepstral mean normalisation and speaker adaptation stages assume that individual segments are homogeneous.

	M	S	T
M	82.11	17.89	0.00
N	15.27	84.22	0.51
S	0.56	98.24	1.20
T	0.00	1.19	98.81

Table 3: Confusion matrix for audio classification (%) for the BNdev96ue data set.

	#seg	#MSseg	# Spkr/seg	Gen Detect
Ref	439	0	1.000	100 %
CMU Seg	491	144	1.318	-
S1	539	100	1.189	95.13 %
S2	553	64	1.108	97.07 %

Table 4: Segmentation results showing number of segments with multiple speakers (#MSseg), average speakers per segment and gender detection accuracy for the basic system with heuristic smoothing (S1) and the improved system which combines smoothing with segment clustering (S2).

3.3. Segment Clustering

The goal of segment clustering is to group segments in order to optimise subsequent adaptation. This requires a compromise between the desire for homogeneity within clusters and the need for clusters of sufficient size for robust unsupervised adaptation.

Two speaker clustering schemes have been studied using the CMU clustering software distributed by NIST [6] as a baseline for comparison. The first scheme was used in our 1996 BN system [8]. This is 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.

The second method 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 until a minimum occupancy count is reached 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.

	F-WB	M-WB	M-NB
CMU	2.183 (45)	2.500 (53)	4.593 (13)
BDIV	2.337 (46)	2.442 (66)	4.183 (14)
TCOV	2.297(44)	2.363 (53)	4.189 (13)

Table 5: Percentage improvement in log likelihood after MLLR adaptation using the CMU segment clustering (CMU), bottom-up divergence-based clustering (BDIV) and top-down covariance-based clustering (TCOV). Numbers in brackets are the actual numbers of clusters formed. The three conditions tested are female wide-band (F-WB), male wide-band (M-WB) and male narrow-band (M-NB).

Table 5 compares the three speaker clustering methods in terms of the percentage increase in log likelihood achieved by the subsequent MLLR-based mean adaptation with a global MLLR transform for each clustered group. The HMM-BN2 set (see Sec. 5.1) was used and the likelihoods are calculated on automatically segmented BNdev96ue data. In each case, the clustering thresholds have been adjusted to give similar numbers of clusters so that measuring the increase in log likelihood provides a reasonably valid comparison. As can be seen, all of the methods give fairly similar performance.

4. RECOGNITION SYSTEM OVERVIEW

The recognition system is a development of previous HTK large vocabulary recognisers (e.g. [7]).

Each frame of input speech is represented by a 39 dimensional feature vector that consists of 13 (including c_0) MF-PLP cepstral parameters [8] and their first and second differentials. Cepstral mean normalisation (CMN) is applied over a segment.

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

In the full HTK system the decoder can operate in multiple passes and use quinphone HMMs, 4-gram language models and iterative unsupervised adaptation. However for all experiments reported here 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% decrease in word error rate on broadcast news.

5. 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 [5, 3].

The data type specific models used WSJ secondary channel HMMs with 6399 speech states and were subsequently adapted to broadcast news (used in [8]). Two sets of condition independent models were trained: the BNtrain96 HMM-BN1 has 5628 states and the BNtrain97 HMM-BN2 set 6944 states. All models used 12 component mixture Gaussian distributions. In all cases gender independent models were used.

The results given in Table 6 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. Furthermore, doubling the amount of training data reduces the error rate by a further 1.5% 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	27.3
F2	51.4	44.8	42.6	40.1
F3	36.4	32.7	35.3	33.9
F4	28.6	25.0	25.4	24.4
F5	28.6	23.8	27.1	26.5
FX	58.5	55.2	56.8	55.0
Avg.	36.0	32.0	31.7	30.2

Table 6: % Word error rates on BNdev96pe for different training conditions. Only the WSJ adapt set is data condition dependent.

Whilst the results shown in Table 6 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. Hence 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 31.3% for HMM-BN2). To try to improve the performance on 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 only on data of the corresponding gender (i.e. it is assumed that perfect gender determination is possible). As shown in Table 7 this gave a substantial increase in recognition performance (overall 1.2% absolute and 1.9% for female speakers) and appears to have largely mitigated the gender bias in the training data.

Data Type	Model type and data type			
	GI / male	GI /fem	GD /male	GD / fem
F0	9.2	14.7	8.7	13.8
F1	26.2	30.8	25.7	28.5
F2	40.2	39.0	38.4	36.4
F3	27.4	39.6	25.1	37.2
F4	24.1	24.7	24.5	22.6
F5	27.6	25.8	25.9	23.2
FX	57.8	52.3	57.0	50.3
Avg.	29.6	31.3	28.7	29.4

Table 7: % Word error rates on BNdev96pe split by gender for gender independent (GI) and gender dependent (GD) models.

It should be noted that although the automatic gender classification in Table 4 yields 3-5% 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 an gender detection error rate of about 2%.

5.3. Automatic Segmentation/Classification

The effect of using the automatically derived segments from both the CMU segmenter and the S1 segmenter described in Sec. 3 was evaluated on the BNdev96ue data. It should be noted that some of the data (that identified as pure music) is discarded by the S1 segmenter while the CMU approach retains the entire data stream. As can be seen in Table 8 recognition performance improves with the S1 segmenter, particularly on F3 segments due in part to the removal of pure music. It is expected that S2 segmenter (which uses clustering) would improve recognition performance further.

Data Type	HMM training	
	CMU Segments	S1 Segments
F0	12.3	11.7
F1	28.1	27.1
F2	41.1	41.5
F3	40.2	32.4
F4	28.3	27.8
F5	31.6	31.4
FX	67.6	67.3
Overall	30.7	29.7

Table 8: % Word error rates on BNdev96ue for different segmentations algorithms using the gender independent HMM-BN2 model set.

The S1 segmenter also classifies data as narrow-band or wide-band. A narrow band model version of HMM-BN2 was trained using a reduced bandwidth data analysis (125Hz to 3.8kHz) and use of these models improved performance on F2 data to 38.3% error and reduced the overall error rate to 29.2%.

Finally the performance of the S1 segmenter was compared to the hand-partitioned PE segments for the Marketplace show of the development data using the GI HMM-BN2 model set. The results are given in Table 9. It can be seen that automatic segmentation has added 2% absolute to the PE error rate. However nearly half of this increase is due to recognition of a Bosnian speaker who is excluded from the PE segments but causes insertion errors in the automatically segmented data, causing a substantial increase in the error rate for FX data. Hence we expect that in general the automatic segmentation will only add about 1% absolute to the error rate from the PE segments.

Data Type	Segmentation	
	PE Segments	S1 Segments
F0	11.1	11.9
F1	22.0	21.2
F2	18.8	22.2
F3	19.5	20.5
F4	27.3	32.7
F5	29.5	31.0
FX	56.0	77.8
Overall	19.0	21.0

Table 9: % Word error rates on the Marketplace dev-test show for automatic and hand generated segmentations.

6. CONCLUSION

This paper has described a number of experiments in broadcast news transcription. It has been shown that a segmentation scheme which integrates clustering is particularly effective and that automatically segmented broadcast news data can yield close to the performance of hand partitioned data using data condition independent modelling.

7. ACKNOWLEDGEMENTS

This work is in part supported by an EPSRC grant on “Multimedia Document Retrieval” reference GR/L49611.

8. 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] 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., Lamel L., Adda G.& Adda-Decker M. (1997). Transcription of Broadcast News. *Proc. Eurospeech’97*, pp. 907-910, Rhodes.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.