Segment Generation and Clustering in the HTK Broadcast News Transcription System

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

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

ABSTRACT

This paper describes the segmentation, gender detection and segment clustering scheme used in the 1997 HTK broadcast news evaluation system and presents results on both the unpartitioned 1996 development and the 1997 evaluation sets. The stages of our approach are presented, namely classification, segmentation and gender detection, gender relabelling, and clustering of speech segments. The evaluation audio stream has been segmented according to audio type with a frame accuracy up to 95%. Further segmentation and gender labelling gave up to 99% frame accuracy with 127 multiple speaker segments. Experiments using two different segmentation approaches and three clustering schemes are presented.

1. Introduction

The transcription of broadcast news requires techniques to deal with the large variety of data types present. Of particular importance is the presence of varying channel types (wide-band and telephone); data portions containing speech and/or music often simultaneously and a wide variety of background noises from, for example, live outside broadcasts. Furthermore, if a transcription system is to deal with complete broadcasts, it must be able to deal with a continuous audio stream containing a mixture of any of the above data types.

To deal with this type of data, transcription systems generally use a segmentation stage that splits the audio stream into discrete portions of the same audio type for further processing. Ideally, segments should be homogeneous (i.e. same speaker and channel conditions), and should contain the complete utterance by the particular speaker. Because of the large variety of audio types present, the data segments should be tagged with additional information so that subsequent stages can perform suitable processing. If possible, non-speech segments should be completely removed from the audio stream but it is important not to delete segments that in fact contain speech to be transcribed. It is also desirable, particularly for implementing speaker adaptation schemes, that data segments from a particular speaker under particular audio conditions can be grouped with other data of the same type.

This paper deals with the segment processing strategies employed by the 1997 HTK broadcast news transcription system. The following section gives a brief system overview which is followed by a detailed description and evaluation. Results are shown on the 1996 DARPA unpartitioned broadcast news development set (BNdev96UE) and the November 1997 evaluation set (BNeval97).

2. System Overview

The overall segment processing can be subdivided into audio type classification, segmentation and finally clustering. The operation of the audio type classification and segmentation steps are shown in Figure 1. The classification stage labels audio frames according to bandwidth and discards non-speech segments, while the segmentation step generates homogeneous segments labelled with gender and bandwidth. The clustering stage groups the segments into a reasonable number of clusters, for use in subsequent adaptation stages. Due to our specific recognition system setup [6], we are able to improve gender labelling using the results of our first pass recognition.

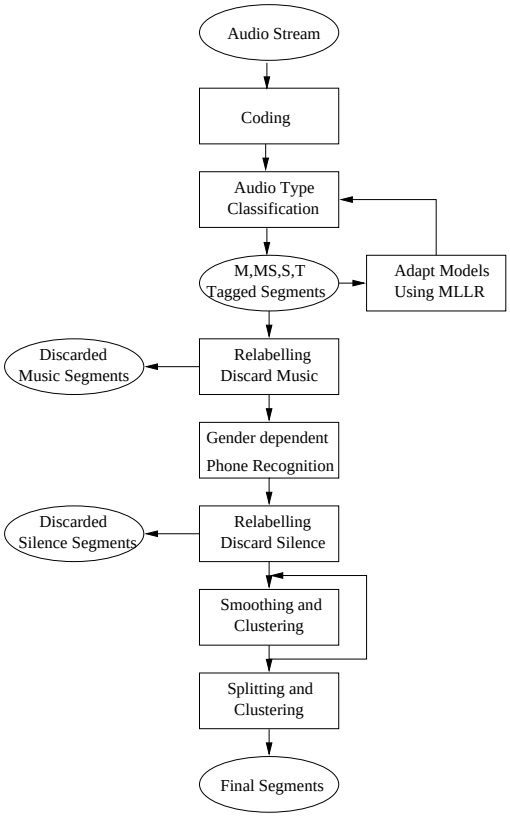

Figure 1: Audio Classification and Segmentation Overview.

In reality the classification process makes errors. Since misclassification of speech as non-speech is more detrimental than keeping undetected non-speech segments, the design goal for segmentation should be minimising loss of speech. Secondly, short segments are not easy to classify or to cluster (e.g. short interjections or confirmation by other speakers during a monologue). Thus segments should be confined to a duration between 0.5 seconds and 30 sec-

onds, where the lower durational range below three seconds should be sparse. Nevertheless this implies that the system will generate some multiple speaker segments.

The following sections present the audio type classification, segmentation and clusterings stages in more detail.

3. Audio Type Classification

The aim of this stage is to classify each frame of a continuous audio stream into three groups : S for wide-band speech, T for narrow-band speech and M for music or other background not relevant for recognition. Because the M-labelled segments are discarded, the major design goal for this stage is not only minimum frame classification error rate, but minimal misclassification of speech as music, i.e. loss of speech.

The audio classification uses Gaussian mixture models (GMM) with 1024 mixture components and diagonal covariance matrices. Four models are trained with approximately 3 hours of audio each. The models used are S for pure wide-band speech, T for pure narrow-band speech, MS for music and speech, and M for Music. The use of a separate model for music and speech has been beneficial to decrease the loss of speech data. Using an additional model for various other background noises present in the data (e.g. helicopter or battlefield noise) turned out to be impossible due to lack of training data and the large diversity in the nature of the data. Moreover some of the material contains background speakers or speech in different languages, which adds to confusion with speech classes.

	background			speech		
	M	BGS	BGO	MS	T	S
BNtrain97	206	13	71	213	270	4016
BNdev96UE	6	1	14	7	9	85
BNeval97	6	< 1	< 1	9	26	142

Table 1: Training and test material available in broadcast news (minutes) for music (M) background speaker (BGS), other background (BGO), music and speech (MS), narrow-band (T) and wide-band (S) speech.

The distribution of broadcast news data suitable for GMM training can be seen in Table 1. The training data contains information about the various speech data types (tagged F0 to FX) and various background noise conditions. The F2 labelled segments are nominally from telephone channels but they have been found to not necessarily have narrow bandwidth and therefore a separate deterministic classifier was used to label training segments as being narrow or wide-band. The classifier is based on the ratio of energy above 4kHz to that between 300Hz and 4kHz.

Pure wide-band speech has been chosen for GMM training from all conditions except narrow-band and F3 (speech with music) labelled segments. A subset of appropriately-sized data was selected to train the GMMs for S and T. The data selection criterion was based on maximising the speech content measured as the ratio of the number of frames aligned to speech phones (not silence) to the total number of frames in a segment. For training the MS model all segments labelled as F3 have been used. The music model data was selected using gaps between speech segments, where the background condition music was tagged. This is problematic, since signature tunes are the major type of music present. The same tune occurs repeatedly in each show, thus decreasing the generalisability of the model. Secondly, in some cases various different background conditions are labelled simultaneously, a fact which has not been taken into account in our selection mechanism.

	%BG corr	%M corr	%Correct	%Loss
train/test	66.26	97.04	97.54	0.03
test only	33.41	39.71	83.91	1.05

Table 2: Table showing frame accuracies on arbitrary non-speech detection (%BG corr), music detection (%M corr), overall and loss of speech accuracy using unadapted GMMs on BNdev96UE plus two additional shows. The test set is split into shows available both in training and test and test only.

The acoustic feature vectors consisted of 12 MFCCs, normalised log energy and the first and second differential coefficients of these. We found that this representation was more effective than the PLP-based features used in word recognition for data-type classification.

Each frame of data was labelled using a conventional Viterbi decoder with each of the four models in parallel and finally MS labelled frames are relabelled as S. An inter-class transition penalty is used which forces decoding to produce longer segments. An additional penalty on leaving the M model should result in fewer misclassifications of speech, however the influence of this parameter on loss reduction is rather limited.

Due to the problem concerning training data for background models mentioned above, classification of music seems to give relatively poor generalisation capability so that shows not available in training seem to produce worse results for classifying music (see Table 2).

To reduce this effect, after an initial classification the models are adapted to the current show using maximum likelihood linear regression (MLLR) [3, 2] for adapting both means and variances using first stage classification as supervision. MLLR transforms (block-diagonal for means, diagonal for variances) for each model were computed when more than 15 seconds of adaptation material was available. For BNdev96UE this was done per show, but on the evaluation set this had to be done on all shows simultaneously, since show boundaries were unknown. 15 iterations of MLLR were performed using first stage classification transcription. This relatively high number of matrix reestimations is required due to the high number of mixture components used. The results of adaptation (Table 3) show an increase in classification accuracy as well as a decrease in loss of speech frames.

Table 4 shows confusion matrices for the adapted models. Note 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. Overall 65% of the non-speech is discarded with only 0.5% loss of speech data. On BNeval97 the behaviour is similar with slightly poorer overall performance. 70.4 % of non-speech has been discarded with only 0.18% speech being lost.

	BNdev96UE		BNeval97	
	Baseline	Adapted	Baseline	Adapted
Frame Accuracy	96.10%	96.23%	93.67%	94.73%
Frames Lost	0.72%	0.48%	0.25%	0.18%
BG correct	62.72%	65.65%	59.20%	70.40%

Table 3: Overall audio classification accuracy and percentage loss of speech on the BNdev96UE and BNeval97 data sets. On the BNeval97 only 0.18% of speech frames were lost, which is equivalent to 20.18 seconds.

	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

a)

	M	S	T
M	78.40	21.55	0.05
N	41.74	54.60	3.66
S	0.22	95.60	4.17
T	0.00	3.54	96.46

b)

Table 4: Confusion matrices for audio classification (%) using adapted GMMs on a) the BNdev96UE data set and b) the evaluation set BNeval97 .

4. Segmentation

The result of the audio type classification stage is a preliminary set of segments labelled as narrow-band or wide-band speech. In this segmentation stage both classes are treated separately, although the same processing is used. The target is to produce homogeneous segments containing a single speaker and data type.

Segmentation and gender labelling is performed using a phone recogniser which has 45 context independent phone models per gender plus a silence/noise model with a null language model. The output of the phone recogniser is a sequence of phones with male, female or silence tags. The phone tags are ignored and phone sequences with the same gender are merged.

Silence segments longer than 3 seconds are classified as non-speech and discarded. Sections of male speech with high pitch are frequently misclassified as female and vice versa. This results in short misclassified segments usually at the beginning or the end of sentences. Even though long silence segments are relatively reliable, short silence segments often cut into words. Hence, a number of heuristic smoothing rules are applied to relabel short segments and merge them into their neighbours.

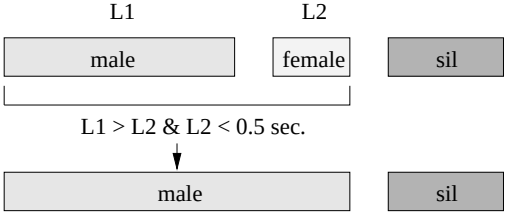

Figure 2: Smoothing rule example with span 3.

The smoothing rules use the label types as well as the absolute duration and the segment duration relative to the neighbouring segments to decide which segments are to be merged and how. Figure 2 shows an example of a rule to merge a long male-labelled segment with a short female labelled segments before silence. The maximum span of rules, i.e. the number of segments considered at once, is five. It was found to be beneficial to use long span rules initially.

SegType	#seg	#MSseg	# Spkr/seg	%Dmult	%GD
Ref	439	0	1.000	0.0	100
CMU	491	144	1.318	8.9	-
S1	539	100	1.189	8.2	95.13
S2	553	64	1.108	2.8	97.07

a)

SegType	#seg	#MSseg	# Spkr/seg	%Dmult	%GD
Ref	634	0	1.000	0.0	100
CMU	769	172	1.239	6.4	-
S2	749	127	1.173	1.6	96.32

b)

Table 5: Segmentation results on BNdev96UE (a) and BNeval97 (b) using various segmentation schemes. The number of segments with multiple speakers (#MSseg), the average speakers per segment, gender detection accuracy (%GD) and the percentage of multiple speaker frames (%Dmult) are shown.

In total 42 rules have been used. Application of a rule is repeated as long as the segmentation changes, then the next rule is used. After a complete loop over all rules, the loop is repeated itself, until the segmentation remains unchanged. An additional constraint is applied to ensure the duration of segments between one and 30 seconds. These conditions cause errors, since short utterances in a discussion or confirmations can be shorter than a second.

The smoothing scheme based only on rules is referred to as S1 in the tables.

This purely heuristic method has a number of disadvantages

- Erroneous grouping of segments not only results in incorrect boundaries, but also wrong gender labels
- Many short silence tags are unreliable and hence have to be merged with neighbouring segments
- Neighbouring speakers with the same gender could be indistinguishable, since short silences between may have been merged.
- Durational constraints might produce suboptimal splits

A possible solution to this problem is the use of segment clustering in the smoothing process. At certain stages in the smoothing process the locally available segments are clustered using a top-down covariance based technique (see below). Segments which appear in the same leaf node and are temporally adjacent are merged into a single segment. The allocation of the gender label of the merged segment is made according to the number of frames per gender label. Clustering is repeated, until segmentation does not change.

Silence segments are not involved in the clustering process, since they are usually too short to give a good estimate of the covariance matrix. The use of segment clustering improves gender labelling

and segmentation, but since it is embedded in the smoothing rules, the problem of adjacent segments having the same gender label still remains. It was observed that a high percentage of the correct segment boundaries are labelled as silence by phone recognition (over 85% on BNdev96UE). Taking advantage of this effect, all segments are then split up again in the middle of silence segments clustering is repeated. This final clustering stage also gives more control on splitting long segments into parts of approximate equal duration. This smoothing and clustering scheme is referred to as S2.

Table 5 shows segmentation results and the improvements on the number of speakers per cluster on the BNdev96UE and BNeval97 sets using both methods S1 and S2. Results are compared with the segmentation given by the CMU software [4] distributed by NIST. Table 6 shows the overall class confusion matrices incorporating classification and segmentation stages.

	M/sil	S male	T male	S female	T female
M/N	67.43	20.65	0.00	11.77	0.15
S male	0.52	96.34	0.47	2.51	0.16
T male	0.00	0.40	99.36	0.18	0.06
S female	0.25	4.19	0.09	94.05	1.42
T female	0.00	0.00	0.00	0.00	100.00

a)

	M/sil	S male	T male	S female	T female
M/N	78.50	13.94	0.55	6.96	0.04
S male	0.62	91.31	5.86	1.67	0.54
T male	0.00	1.88	84.55	1.01	12.56
S female	0.22	1.35	0.44	97.63	0.35
T female	0.00	5.06	5.62	0.50	88.82

b)

Table 6: Overall Confusion Matrix on BNdev96UE (a) and BNeval97 (b) using method S2(%).

A general disadvantage of this method is that it is impossible to detect speaker transitions between two speakers of the same gender, if there is no intervening silence. However, as the results in Table 5 imply, that this is rarely a problem.

5. Gender Relabelling

The evaluation system setup [6] enables us to further improve gender labelling. Segments of the above stage are decoded using gender labels and bandwidth labels to select the appropriate model set for further decoding stages. After decoding forced alignments on both gender models are carried out. The gender label for each segment is selected by the model set with highest log-likelihood. This improves gender labelling accuracy by about 2% absolute(Tables 7 & 5).

	#Seg Changed	% gender correct
BNdev96UE S2	28	99.01 %
BNeval97 S2	31	98.61 %

Table 7: Table showing the number of segments changed and resulting accuracy on gender relabelling.

Recognition experiments using segments based on the S2 scheme show for both BNdev96UE and BNeval97 show a significant improvement over CMU segmentation and the S1 scheme (Table 8). Segments have been decoded using gender independent wide-band triphone HMM models and a trigram language model. For the BNeval97 set, when narrow-band data coding and appropriate HMMs are used, a further reduction in word error rate of 1.6% absolute is achieved. Note that we also expect to get benefits in adaptation due to the fact that the segments lie closer to our goal of homogeneity. Detailed recognition results for the complete HTK transcription system can be found in [6].

	BNdev96UE	BNeval97
CMU	30.1	23.9
S1	29.2	-
S2	28.6	23.0

Table 8: Error rates using different segmentation schemes. Segments have been decoded using gender independent wide-band models. Error rates have been computed using the 1997 scoring protocols.

6. Segment Clustering

The goal of segment clustering to group segments in order to optimise subsequent adaptation performance. 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 [4] as a baseline for comparison. The first scheme was used in our 1996 BN system [5]. 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 no node can be split to produce nodes satisfying a minimum occupancy criterion. 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.

Table 9 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. A gender and condition independent model set 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, and the final transcription system used the same bottom-up clustering scheme used the the 1996 HTK broadcast news transcription system.

	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 9: 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).

7. Conclusions

The segment generation and clustering stages of the 1997 HTK broadcast news transcription system have been described. It is shown that the techniques used are reasonably successful in producing homogeneous segments and that the segments produced yield improved word recognition performance compared to the standard segmentation software distributed by NIST.

8. Acknowledgements

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

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. 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.
4. 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.
5. 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.
6. Woodland P.C., Hain T. , Johnson S.E., Tuerk A., Niesler T.R. & Young S.J. (1998) The 1997 HTK Broadcast News Transcription System. to appear in *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, Virginia.