SEGMENTATION AND CLASSIFICATION OF BROADCAST NEWS AUDIO

T. Hain *P.C. Woodland*

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.
email : {th223, pcw}@eng.cam.ac.uk

ABSTRACT

Broadcast news audio data contains a wide variety of different speakers and audio conditions (channel and background noise). This paper describes a segmentation, gender detection and audio classification scheme for such data which aims to provide a speech recogniser with a stream of reasonably-sized segments, each from a single speaker and audio type while discarding non-speech data. Each segment is labelled as either narrow or wide band and from either a female or male speaker. The segmentation system has been evaluated on the DARPA 1997 broadcast news data set and detailed segmentation accuracy results are presented. It is shown that the speech recognition accuracy for these automatically derived segments is very nearly the same as that for manually segmented data.

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.

This paper deals with the segmentation strategy developed for the HTK broadcast news transcription system [6]. The following sections briefly describe the broadacst news data we are using and then we give a system overview which is followed by a more detailed description and experimental evaluation. Finally speech recognition experiments using the 1997 HTK broadcast news transcription system are presented on the 1997 DARPA broadcast news Hub4 evaluation set (BNeval97).

2. Broadcast News Data

The Hub4 English broadcast news data that is distributed by the Linguistic Data Consortium (LDC) contains the complete audio track for a number of US radio and television shows. The various types of audio present in a broadcast are denoted by the focus conditions (Table 1). While F0, F1 and F5 represent clean speech, F3,F4 and FX contain background noise or music. F2 most commonly labels segments containing telephone interviews. Since different recognition models may be used for different types of data the segmenter aims to label each segment as to bandwidth and gender.

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.

3. System Overview

The overall segment processing can be subdivided into audio type classification and segmentation. The segment processing 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 and adds gender labels.

In reality the classification process makes errors. Since misclassification of speech as non-speech is more detrimen-

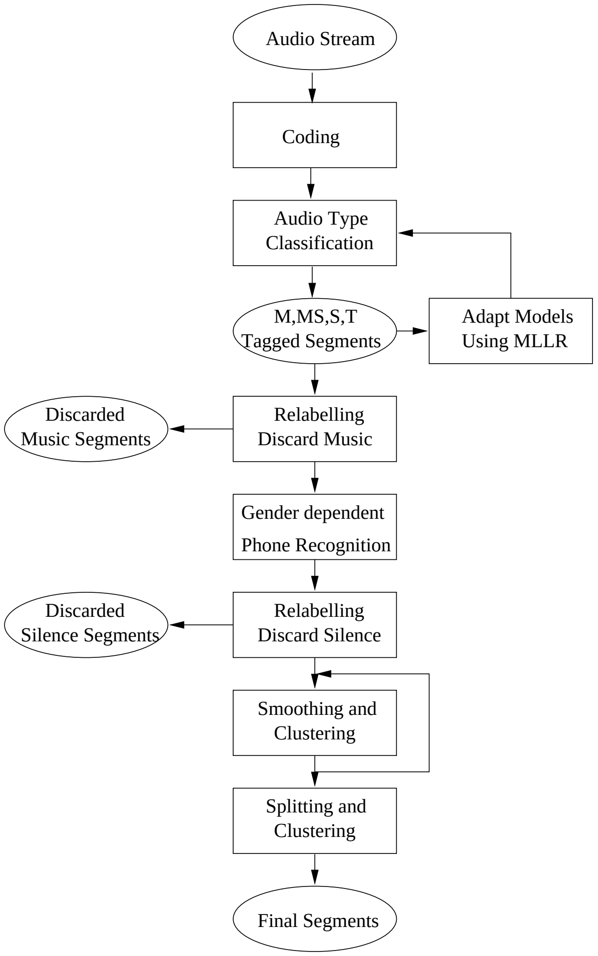

Figure 1: Audio Classification and Segmentation Overview.

tal than keeping undetected non-speech segments, the design goal for segmentation is to minimise "loss" of speech. Secondly, short segments are not easy to classify or to recognise (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 seconds. Nevertheless this implies that the system will generate some segments with containing data from multiple speakers.

4. 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 of the data. Moreover some of the material contains background speakers or speech in different languages, which adds to confusion with speech classes.

Data Set	background			speech		
	M	BGS	BGO	MS	T	S
BNtrain97	206	13	71	213	270	4016
BNeval97	6	< 1	< 1	9	26	142

Table 2: 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 is shown in Table 2. 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 segmental 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. Since data for training the music model has not been tagged in the reference transcripts, an automatic procedure extracting gaps between speech segments has been applied. The selection of data for the music model is problematic, since signature tunes are the main type of music present. The same tune occurs repeatedly in each show, thus decreasing the model's generalisation ability.

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. For classification each frame of test 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 interclass transition penalty is used which forces decoding to produce longer segments.

Due to the problem concerning training data for background models mentioned above, the effects of shows appearing in test data only have been investigated on the DARPA 1996 broadcast news development set (BN-dev96ue). In Table 3 the degradation of accuracy on both music and general background conditions is clearly visible for shows not observed in training and this is accompanied by an increased loss of speech.

Show type	%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 3: 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.

To reduce this effect, after an initial classification the models are adapted to the current show using maximum likelihood linear regression (MLLR) for adapting both means and variances [1] using the 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 has been available. 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 4) show an increase in classification accuracy as well as a decrease in loss of speech frames. Overall, adaptation increased the percentage of frames correctly discarded to 70.4 % together with descreasing the percentage of frames lost to 0.18%.

Measure	Baseline	Adapted
Frame Accuracy	93.67%	94.73%
Frames Lost	0.25%	0.18%
BG correct	59.20%	70.40%

Table 4: Overall audio classification accuracy and percentage loss of speech on the BNeval97 set. Only 0.18% of speech frames were lost, which is equivalent to 20.18 seconds.

Table 5 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.

	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

Table 5: Confusion matrices for audio classification (%) using adapted GMMs on BNeval97

5. 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 bandwidths 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 (cf. [4]) 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, relabelling certain configurations of segments. These rules both take into account the length and the label of each segment considered, with further durational constraints on the final duration of segments. Each rule is applied until segmentation is unchanged. More information on these smoothing rules is given in [2].

This purely heuristic method has a number of disadvantages

- Erroneous grouping of segments results both in incorrect boundaries and incorrect 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 [3].

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

Table 6: Segment purity using various schemes. The number of segments with multiple speakers (#MSseg), gender detection accuracy (%GD) and the percentage of multiple speaker frames (%Dmult) are shown on BNeval97.

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 again is repeated, until segmentation does not change. This smoothing and clustering scheme is referred to as the S2 segmenter.

Table 6 compares S2 segmentation results with the segmentation given by the CMU software [5] distributed by NIST. The percentage of frames associated with multiple segments is very much lower for the S2 segmenter and the remaining multiple speaker segments are quite short (average length 1.3s). Table 7 shows the overall class confusion matrices incorporating classification and segmentation stages.

	sil	male		female	
	M	S	T	S	T
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

Table 7: Overall Confusion Matrix using method S2(%) on BNeval97

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 6 imply, this is rarely a problem.

6. Recognition Experiments

The effect of using the automatically derived segments from both the CMU segmenter and the S2 segmenter described above was evaluated using the HTK Broadcast News Transcription System [6]. This experiment used bandwidth independent gender-independent cross-word state clustered triphones and a triphone language model and decoding operated in a single pass with fairly tight beamwidths. It should be noted that some of the data (that identified as pure music) is discarded by the S2 segmenter while the CMU approach retains the entire data stream. As can be seen in Table 8, recognition performance improves the overall performance, but particularly F3 performance due to removal of non-speech segments. The overall word error is just 0.1% higher than if a perfect manual segmentation is used.

Data	Segmentation Alg		
Type	CMU	S2	Manual
F0	13.3	13.0	12.9
F1	21.6	20.8	20.2
F2	35.6	34.9	35.5
F3	34.1	32.4	34.2
F4	26.2	25.7	25.0
F5	29.0	27.5	27.5
FX	50.9	46.8	45.6
Overall	23.9	23.0	22.9

Table 8: % Word error rates for using the first pass HTK Broadcast News transcription system on manually and automatically generated segments for BNeval97

7. Conclusions

A segment generation scheme for broadcast news audio data has been described. It has been shown that the techniques used are reasonably successful in producing homogeneous segments. Speech recognition experiments on the resulting segments yield word error rates very close to those from manual segmentation.

8. Acknowledgements

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

9. REFERENCES

1. Gales M.J.F. & Woodland P.C. (1996). Mean and Variance Adaptation Within the MLLR Framework. *Computer Speech & Language*, Vol. 10, pp. 249-264.
2. 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. DARPA BNTUW*, pp. 133-137.
3. Johnson S.E. & Woodland P.C. (1998). Speaker Clustering Using Direct Maximisation of the MLLR-Adapted Likelihood. To appear in *Proc. ICSLP'98*, Sidney
4. 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.
5. 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.
6. Woodland P.C., Hain T. , Johnson S.E., Niesler T.R., Tuerk A.,Whittaker E.W.D. & Young S.J. (1998) The 1997 HTK Broadcast News Transcription System. *Proc. DARPA BNTUW*, pp. 41-48.