# SEGMENTATION AND CLASSIFICATION OF BROADCAST NEWS AUDIO

*T. Hain* *P.C. Woodland*

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

## ABSTRACT

Broadcast news 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 and presents experimental results on the DARPA 1997 broadcast news evaluation set.

The goal of the segment processing algorithm is to convert the continuous input audio stream into reasonably-sized speech segments, which are labelled as either being narrow or wide-band speech and belonging either to a female or male speaker. Ideally, each segment should be homogeneous (i.e. same speaker and channel conditions) and the removal of non-speech segments should be designed to minimise incorrectly discarded speech.

Since the reason for developing the algorithm has been to enable recognition of broadcast news data, the recognition performance using various segmentation sources has been tested. On the evaluation data, the first pass of the HTK broadcast news transcription system using gender independent HMMs gave 23.0% word error using this segmentation scheme compared to 22.9% using manual segmentation and 23.9% based on CMU segmentation software distributed as reference algorithm by NIST.

## 1. Introduction

Broadcast news data as distributed by the Linguistic Data Consortium (LDC) is a set of complete television or radio shows like CNN Headline News or MPR Marketplace. The various types of speech present in a typical broadcast are denoted by the focus conditions (Table 1). Opposed to F0, F1 and F5 the conditions F3,F4 and FX are sometimes severly distorted by non-speech sounds. F2 most commonly labels segments containing telephone interviews. Since using different approaches for various conditions has shown to be effective, the segmentation stage also has to label segments according to bandwidth and speaker gender.

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.

| 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.

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.

The following section gives a brief system overview which is followed by a more detailed description and evaluation. Finally recognition experiments using the 1997 HTK broadcast news transcription system are presented on the November 1997 broadcast news evaluation set ( BNeval97 ).

## 2. 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

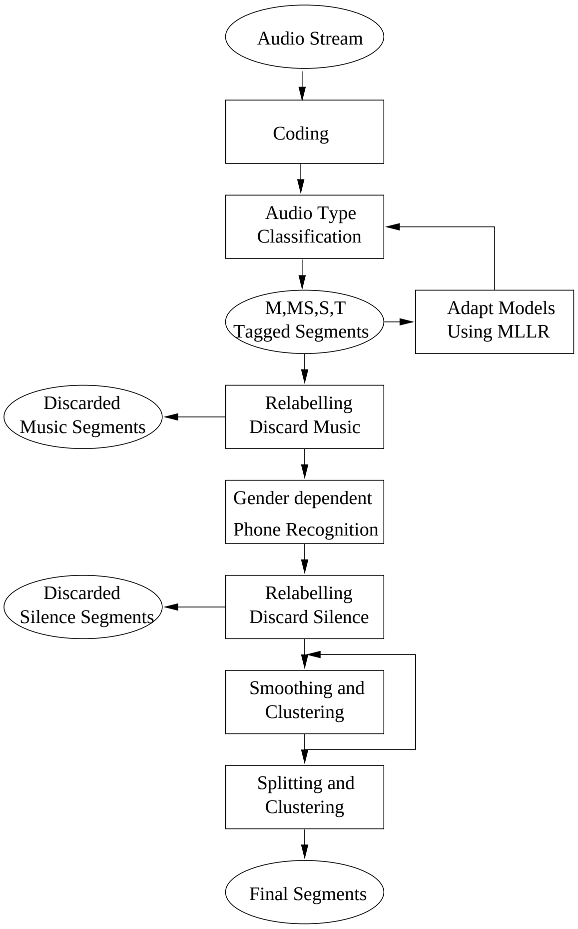


**Figure 1:** Audio Classification and Segmentation Overview.

generates homogeneous segments and adds gender labels.

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 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 multiple speaker segments.

## 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 |
| 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 can be seen 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. This selection of data for the music model is generally 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.

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

| | %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.

| | 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.

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.

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 (BNdev96ue). In Table 3 the degradation of accuracy on both music and general background conditions is clearly visible, paired with an increased loss of speech.

To reduce this effect, after an initial classification the models are adapted to the current show using maximum likelihood linear regression (MLLR) [1] 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 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.

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. On BNeval97 adaptation increased the percentage of frames correctly discarded to 70.4 % together with descreasing the percentage of frames lost to 0.18%.

| | 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

## 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, relabelling certain configurations of segments. These rules both take into account the length and the label of each segment considered, with furhter durational constraints on the final duration of segments. The maximum number of segments considered at once is five, a total of 42 rules have been used. Each rule is applied until segmentation is unchanged.

A more detailed investigation of segmentation based on heuristic rules only can be found in [2].

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

| 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.

the smoothing process the locally available segments are clustered using a top-down covariance based technique [3]. 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 S2.

Table 6 compares S2 segmentation results with the segmentation given by the CMU software [4]. The percentage of frames associated with multiple segments is significantly lower. This is not only caused by a decrease of the number of segments with multiple speakers, but also very short overlaps with neighbouring segments, which also might be caused by reference transcript timing inaccuracies. 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(%)

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, that this is rarely a problem.

## 5. Recognition Experiments

Transcription systems for broadcast new data as used in the yearly DARPA Boradcast news evaluations usually involve multiple stages of processing the input audio stream. Stages for first computing low accuracy estimates of the data are followed by unsupervised per-speaker model adaptation schemes. The quality of speech segments for such a system has both an influence on initial decoding as well as further adaptation stages, where the purity of segments becomes more important.

The effect of using the automatically derived segments from both the CMU segmenter and the S2 segmenter described above was evaluated using the first pass HTK Broadcast News Transcription System [5]. 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 performance loss to manual segmentation is just 0.1%.

| 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

## 6. Acknowledgements

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

## 7. 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 Broadcast News Transcription and Understanding Workshop*, pp. 133-137, Lansdowne, Virginia
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. 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., 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 Broadcast News Transcription and Understanding Workshop*, pp. 41-48, Lansdowne, Virginia