A STUDY OF BROADCAST NEWS AUDIO STREAM SEGMENTATION AND SEGMENT CLUSTERING

Matthew Harris, Xavier Aubert, Reinhold Haeb-Umbach and Peter Beyerlein.
Philips Research Laboratories, Weisshausstrasse 2, D-52066 Aachen, Germany
{harris,aubert,haeb,beyerlei}@pfa.research.philips.com

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ABSTRACT

In transcription of broadcast news, dividing the signal into homogeneous segments, and clustering together similar segments is important. Decoding a complete broadcast news program in one chunk is technically difficult. Also, through creation of homogeneous clusters of segments, improvement from adaptation can be increased.

Two systems of segmentation and clustering are compared. The best system used the BIC algorithm to produce long, homogeneous segments, and a nearest neighbour bottom-up agglomerative clustering algorithm to produce homogeneous clusters. Adaptation brought a word error rate (WER) improvement from 23.4% to 21.0% using the automatic segmentation and clustering, compared to an improvement from 21.8% to 20.0% using a handmade "correct" segmentation and clustering.

1. INTRODUCTION

The automatic transcription of broadcast news is a task that contains many challenging, real world problems. One encounters, for example, telephone speech, speech in noisy "real life" surroundings, spontaneous speech (as opposed to planned, or read speech) and non-speech (such as music, traffic noise etc.). Model adaptation (as in many other settings) is an important ingredient in broadcast news transcription systems. Models should be adapted to the speaker and acoustic conditions at hand. In many previously studied transcription scenarios, like the Wall Street Journal corpus, the speaker and background conditions are predefined, and the utterance boundaries are clear. Decoding the complete broadcast news program in one chunk is technically difficult. The aim of the segmenter in the broadcast news setting is to divide the signal into "segments" with one speaker and constant acoustic background conditions. The input stream can then be decoded segmentwise, compared to utterancewise decoding in the Wall Street Journal systems. One speaker often occurs in several segments (during an interview, for example). The aim of the clusterer is to group these segments together for use in adaptation, which leads to better, more effective models.

In this paper we compare two systems for segmentation and clustering. These were evaluated on the 1997 Hub4 evaluation data. This data consists of three hours of broadcast news programs. This broadcast data also has an "official" segmentation and clustering generated by human listeners. The data is divided into segments, the borders of which are changes in speaker or background conditions. Each segment contains utterances by just one speaker, speaking in uniform acoustic background conditions. Each segment is given various attributes such as speaker name and background conditions. This official segmentation is used as a basis for the assessment of the quality of the automatically generated segmentations.

1.1. System Overview

The two segmentation and clustering systems (NOV97 and NOV98) were used in the Hub4 97 and 98 evaluations respectively[1] [2]. The architecture of the two segmentation and clustering systems is given in Figure 1. The three common stages are (1) initial chopping, (2) segmentation and classification, and (3) clustering - partitioned in the figure. The input signal is first chopped into chunks required in the further stages. The signal is then segmented at points of speaker change. Finally the same speaker segments are clustered together. The set up of the two systems is as follows:

NOV97 The CMU segmenter [3] was used to produce segments with bandwidth classification. A gender dependent phoneme decoder (see Section 3) was then run, producing the final segmentation, with gender classification. Non-speech passages were also discarded at this stage. The (m/f)+(telephone/non-telephone) segments were then clustered separately (see Section 5).

NOV98 A simple silence detector was run to chop the data at reliable silences. Long non-speech passages were removed using a GMM decoder (see Section 2). The speech passages were subsequently segmented using the BIC algorithm (see Section 3.1). The resulting segments were classified as telephone/non-telephone, and male/female (see Section 4). The (m/f)+(telephone/non-telephone) segments were then clustered separately (see Section 5), and finally further divided at times of silence to produce segments no longer than 20 seconds.

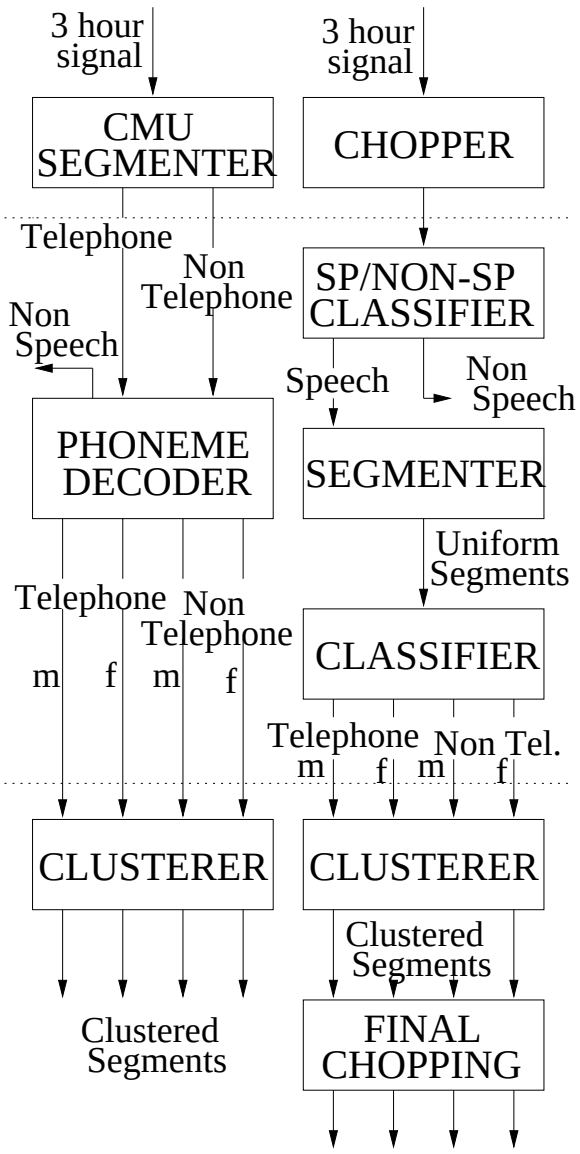

Figure 1: Segmenter and clusterer architecture

2. REMOVAL OF NON-SPEECH

If segments of non-speech are decoded, (spurious) transcriptions with bad scores are produced. Segments with bad scores can be discarded later, but, if the segments also contain speech, it is lost.

In both the NOV97 and NOV98 systems, *reliably* detected non-speech passages were discarded. This way only very little speech was lost (an unrecoverable error), and the resulting segments had less non-speech passages.

NOV97 200 reliable passages of non-speech were discarded (made up of 227 segments) during the phoneme decoder pass (see Section 3).

NOV98 HMM models of speech, speech with music, music, noise and silence were trained using the broadcast news training data. A decoding with these models was done, where jumps between different HMM models were penalised to inhibit noisy transitions. After the decoding, all speech with music frames were relabeled as speech frames, and then a simple smoothing was carried out. 227 long passages of non-speech were eliminated. This approach is similar to that in [4] [5] and [6].

non-sp detector	speech to noise (%)	noise to speech (%)	words cut (%)
NOV97	0.35	78.21	0.049
NOV98	0.26	73.69	0.055
official	0	73.4	0

Table 1: Misclassification of speech and non-speech.

2.1. Comparison for the removal of non-speech

There are two types of classification errors possible. Firstly, non-speech can be classified as speech. Resulting segments with mostly non-speech can be discarded later as they are decoded with a bad score. This error is therefore not serious. Secondly, speech can be classified as non-speech, and discarded. The discarded words cannot be recognised, resulting in deletion errors. This is an unrecoverable error.

Another minor error that can occur is that parts of words are lost due to misplaced non-speech passage boundaries. A word may be cut at such a boundary, resulting in part of the word lost.

These error figures are given in Table 1. True speech was considered as being passages where a word is uttered. Between word pauses and longer (music or noise) pauses were all considered as being true non-speech. As only larger blocks were classified as speech, also in the official segmentation, much "true" non-speech (short pauses between words) was "misclassified" as speech. We see that more of the non-speech passages were removed in the NOV98 system than in the NOV97 system. At the same time, NOV98 discarded less speech than NOV97. The number of words cut in the two systems is comparable, and is negligable.

We conclude that the NOV98 setup is better at removing non-speech passages. Thus, in our setup, using few GMM models for speech and non-speech is more effective than using the phoneme models.

3. SEGMENTATION

NOV97 A decoding was carried out using male and female context independent phonemes together with a non-speech model. The models were trained on the broadcast news (BN) training data. A Viterbi one-pass decoding was carried out, guided by a bigram language model. The output was then smoothed. The segmentation was achieved by creating segments with male-female or speech-nonspeech transitions as the segment boundaries. This resulted in 483 female speech segments and 948 male speech segments, and

227 nonspeech segments. The phoneme decoder segmentation approach is similar to that of BBN [7].

NOV98 The key step was the use of the Bayesian Information Criterion (BIC)[8]. This criterion was used to determine the positions of speaker and background change. The BIC algorithm produced 552 segments - 378 were later classified as male and 174 as female. These segments were further divided at times of nonspeech after the clustering algorithm using a GMM decoding run similar to that described in Section 2. The final NOV98 segmentation consisted of 594 male and 331 female speech segments.

3.1. The BIC algorithm

Given a passage of BN data, the BIC method is able to find the most likely position of speaker or background condition change. (BIC actually looks for positions where the signal characteristics change.) It also gives a criterion to determine whether the change at this point is significant, or not. We give a brief description of the BIC method, and refer to [8] for more details.

We only describe here how BIC is used to find *one* speaker change in a stream of cepstral data $\{x_i \in \mathbb{R}\}^d, i = 1 \ldots N\}$. (We used 16 component MFCC feature vectors (so $d = 16$).) We suppose that cepstral vectors generated from *one* speaker speaking in *one* set of background conditions can be modelled by *one* multivariate Gaussian distribution. The mean and covariance matrix of this Gaussian should clearly be the sample mean and sample covariance. We now compare two scenarios: 1: only one speaker speaks in the passage $x_1 \ldots x_N$. 2: One speaker speaks for the first t frames $x_1 \ldots x_t$, and another for the remaining frames $x_{t+1} \ldots x_N$. There are now two hypotheses.

1. $x_i \sim \mathcal{N}(\mu, \Sigma)$, $i = 1 \ldots N$,
   where μ and Σ are estimated on $x_1 \ldots x_N$.
2. $x_i \sim \mathcal{N}(\mu_1, \Sigma_1)$, $i = 1 \ldots t$,
   $x_i \sim \mathcal{N}(\mu_2, \Sigma_2)$, $i = t+1 \ldots N$
   where (μ_1, Σ_1) and (μ_2, Σ_2) are estimated on $x_1, \ldots, x_t$ and $x_{t+1}, \ldots, x_N$ respectively.

In the first hypothesis, all N vectors are modelled by one Gaussian, and in the second, the first t are modelled by one Gaussian, and the last $N - t$ are modelled by another. The log likelihood ratio of these two hypotheses is

$$R(t) = N \log|\Sigma| - t \log|\Sigma_1| - (N - t) \log|\Sigma_2|,$$

and the time of most probable speaker change is the value $\hat{t}$ of t for which $R(t)$ is maximum.

Modelling N vectors by two Gaussians rather than one can always give a better fit, as one uses twice as many parameters. The improvement gained by this increase in parameters can be offset. $BIC(t)$ is defined to be

$$BIC(t) = R(t) - \lambda \alpha(d) \log N,$$

segmenter	segment purity	avg seg length	words cut (%)
NOV97	97.6	7.33	0.241
BIC	97.7	18.86	0.067
NOV98	97.7	11.26	0.165
official	100	15.87	0

Table 2: Segmentation evaluation.

where $\alpha(d) = \frac{1}{2}(d + \frac{1}{2}d(d+1)) = \frac{1}{2}\#Parameters$, and $\lambda \in \mathbb{R}$. We say there was a speaker change at $\hat{t}$ if $BIC(\hat{t}) > 0$. There are some arguments for the choice of $\lambda = 1$, however, varying λ can change the sensitivity to speaker change (and noise).

The NOV98 segmentation used the BIC method above, extended to find multiple speaker changes. (See [8]).

3.2. Comparison of segmentations

We judge the quality of a segmentation by the *speaker purity* of its segments. Later we will need the definition of the *cluster purity*, and will define it here. Consider a segment (cluster) S with feature frames $x_1, \ldots, x_n$. Suppose speakers $1 \ldots m$ speak in the segment (cluster), with n_i frames spoken by speaker i, $n_1 \geq n_2 \geq \ldots \geq n_m$, and thus $\sum_{i=1}^{m} n_i = n$. Speaker 1 is called the main speaker in the segment (cluster), speaking for n_1 frames. Then, the speaker (cluster) *purity*

$$\mathcal{P}(S) := \frac{100 n_1}{n} \tag{1}$$

is the percentage of time that the main speaker is speaking in the segment (cluster).

Also, the percentage of words split across two segments (word cuts) is given. The NOV98 segmenter is better than NOV97 although extra word cuts are introduced in NOV98 in the last segment refinement. The purity of the segments generated by the NOV97 segmentation is comparable to that of the BIC and NOV98 segmentation, but the average BIC segment length is much greater than that in the NOV97 segmentation. It is easy to reach a high segment purity with short segments, and thus the BIC algorithm is a more effective algorithm in detecting speaker changes than that used in NOV97.

4. CLASSIFICATION

In the NOV97 system, the bandwidth classification was taken from the CMU segmentation, and the gender classification was done by the phoneme decoder. In the NOV98 system, a segment is deemed to be a telephone segment if there is little energy outside of the 300-3500 Hz range, and non-telephone otherwise. To determine the gender of a segment, monophone male and female phoneme decoding runs were carried out. The gender was decided to be that of the run with the best likelihood.

5. CLUSTERING

The clustering algorithms in both the NOV97 and the NOV98 system were based on the Kullback-Leibler

system	cluster purity	gender accuracy
NOV97	73.6	97.49
NOV98	89.2	97.87
NOV98 contrast	89.1	96.02
official	100	100

Table 3: Cluster purity & framewise gender accuracy.

system	WER % before adap	WER % after adap
NOV97	23.7	22.6
NOV98	23.4	21.0
NOV98 contrast	23.4	21.3
official	21.8	20.0

Table 4: Improvement in WER from adaption.

distance, augmented with a term for favouring the merging of neighbouring segments (see [1], [3]). Both systems used bottom-up agglomerative clustering algorithms. In the NOV97 system, segments were clustered using a greedy algorithm: a segment was clustered to the first existing segment within a certain distance of it. In the NOV98 system, at each stage, the two nearest clusters were merged to form a new cluster. After the initial clustering, all small clusters were merged to form larger ones.

We measure the effectiveness of the clustering algorithm by the cluster purity, defined in (1). The results are given in Table 3. We see that the NOV98 clustering algorithm produced a superior cluster purity to that of NOV97.

At this point, we note the importance of carrying out a gender classification before clustering. A contrast clustering was carried out, clustering the telephone and non-telephone segments separately, and then determining the gender for each of the resulting clusters. There was a slight degradation in the cluster purity, and a significant degradation in the framewise gender classification accuracy (see Table 3). We note that due to the impurity of the clusters, a perfect gender classification on the segment or cluster level is impossible.

6. IMPROVEMENT IN WORD ERROR RATE

A gender dependent one-pass trigram decoding with word internal triphones was carried out using the NOV97, NOV98 and the official segmentations and clusterings. A first recognition was done without adaption, and subsequently, one using VTN and MLLR adapted models.

We see that a high cluster purity is an important factor in maximising gains from adaptation techniques. Adaptation brought a 10.3% improvement in the NOV98 system, compared to a 4.6% improvement in the NOV97 system. Adaptation using the official clustering and segmentation reduced the error from 21.8% to 20.0% (an 8.3% relative improvement). The baseline error rate before adaption using the official segmentation was, however, lower.

Also, we see the importance of optimal gender classification. The contrast clustering with inferior gender classification accuracy described in Section 5 resulted in only an 8.9% relative improvement from model adaption.

7. CONCLUSION

In this paper two methods of segmentation and clustering of the Hub4 broadcast news data were examined. Music and other non-speech can be effectively identified using GMM models, trained on 'speech' and 'noise'. The BIC algorithm was accurately able to find speaker changes, and thus produces pure, long segments. A nearest neighbour bottom-up agglomerative clustering scheme was superior to a greedy clustering scheme. The relative improvement from adaptation was 10.3% for our best automatic segmentation and clustering system, compared with 4.6% for the alternative system.

References

[1] P. Beyerlein, X. Aubert, R. Haeb-Umbach, D. Klakow, M Ullrich, A. Wendemuth and P. Wilcox, "Automatic Transcription of English Broadcast News.", Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Virginia, February 1998.

[2] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A Wendemuth, S. Molau, M. Pitz and A. Sixtus, "The Philips/RWTH system for transcription of Broadcast News", Eurospeech 1999.

[3] M.A. Siegler, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio", Proc. of the DARPA Speech Recognition Workshop, Virginia, February 1997.

[4] T. Hain, S. Johnson, A. Tuerk, P. Woodland and S. Young, "Segment Generatation and Clustering in the HTK Broadcast News Transcription System". Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Virginia, February 1998.

[5] J. Gauvain, L. Lamel and G. Adda, "The LIMSI 1997 Hub-4E Transcription System", Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Virginia, February 1998.

[6] S. Chen, M. Gales, P. Gopalakrishnan, R. Gopinath, D. Kanevsky, P. Olsen and L. Polymenakos, " IBM's LVCSR System for Transcription of Broadcast News Used in the 1997 Hub4 English Evaluation", Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Virginia, February 1998.

[7] H. Jin, F. Kubala and R. Schwartz, "Automatic Speaker Clustering", Proc. of the DARPA Speech Recognition Workshop, Virginia, February 1997.

[8] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion". Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Virginia, February 1998.