SPEAKER DETECTION IN BROADCAST SPEECH DATABASES

Aaron E. Rosenberg *Ivan Magrin-Chagnolleau* *S. Parthasarathy* *Qian Huang*
Speech and Image Processing Services Research Lab
AT&T Labs
Florham Park, NJ 07932 - USA

ABSTRACT

Experiments have been carried out to assess the feasibility of detecting target speaker segments in multi-speaker broadcast databases. The experimental database consists of NBC Nightly News broadcasts. The target speaker is the news anchor, Tom Brokaw. Gaussian mixture models are constructed from labelled training data for the target speaker as well as background models for other speakers, commercials, and music. Four labelled 30-min. broadcasts are used for testing. Mel-frequency cepstral features, augmented by delta cepstral features are calculated over 20 msec. windows shifted every 10 msec. through a broadcast. Likelihood ratio scores are calculated for each test frame averaged over blocks of frames with a specified duration. The block scores are input to a detection routine which returns estimates of target segment boundaries. The range of best results obtained over the test broadcasts is 82% to 100% detection of target segments with segment frame accuracy ranging from 86% to 95%. 0 to 2 false alarm segments are detected over each 30 min. broadcast.

1. INTRODUCTION

As more and more multimedia databases are digitized and stored in accessible archival files, the need grows for effective ways to search and retrieve useful information from them. For the speech portions of such databases, a useful search key, in combination with such keys as words and topics, is speaker identity. This paper reports on experiments carried out to assess the feasibility of detecting target speaker segments in multi-speaker broadcast databases. The experimental database consists of NBC Nightly News broadcasts. The target speaker is the news anchor person, Tom Brokaw. The underlying application is to segment a news broadcast into individual news stories for further segmentation and/or browsing. The premise is that each news story is introduced by the anchor person. Thus detecting anchor person segments in the broadcast helps to locate the beginning of each news story.

Speaker detection of a single target speaker can be considered a generalization of the speaker verification task. In speaker verification, a speech sample is provided which is claimed to be generated by the target speaker. The task is to decide whether or not the claim is valid. In speaker detection the sample to be processed is a sequence of audio segments. Each segment may contain speech from the target speaker, speech from another speaker, audio data from a variety of sources such as music, mixed speech and music, noise, and also silence. The target and other speaker segments may be contaminated by the presence of noise and music. The task is to locate target speaker segments in the data by providing estimates of the start and end times of each such segment.

A number of other studies have been reported, e.g. [4, 5, 6, 3, 1], on various speaker segmentation and identification tasks in multi-speaker databases, with both unsupervised and supervised training conditions. For the current application we assume that labelled training data is available for the target speaker, for other speakers, and for other types of audio data present in the broadcasts such as commercials, music, and noise. Our approach is to construct Gaussian mixture models to represent the speech of the target speaker as well as background models. Background models are constructed for speakers other than the target, for commercials (mixed speech and music), and for music only. A detection routine provides estimates of target speaker segment start and end times making use of a likelihood ratio score calculated frame by frame through a test sample.

2. DATABASE

The experimental database consists of 17 half-hour broadcasts of NBC Nightly News recorded off the air from January to March 1998. 13 broadcasts are reserved for extracting data for training while the last 4 are used for testing. The broadcasts were recorded digitally, digitized at a 16 kHz sampling rate into 16-bit PCM samples. The digitized audio data is manually labelled and segmented according to the following categories: target speaker, other speaker, commercial, music, noise, and silence. Additional descriptions are provided for each segment including the gender and identity of the speaker (where possible), and assessments of recording source (such as studio, on site, telephone) and quality.

Table 1 summarizes the segment statistics for the four test broadcasts. Note that test broadcast number 3 contains no target speaker segments. A substitute anchor person (female) was used in place of Tom Brokaw for this broadcast. Roughly speaking, in the other three broadcasts, target speaker segments account for 18% of the total, other speakers for 50%, and commercials for 25%. The remaining 7% of the segments are labelled either music, noise, or silence, with music predominating among these minor categories. There are 15 or 17 target speaker segments in each broadcast ranging in duration from 3 to 67 secs with an average duration of 20 secs.

3. AUDIO PROCESSING

The digitized audio data files are converted to 12th order cepstral coefficients by carrying out a DCT on the output of 31 mel frequency spaced filters. The analysis windows are 20 msecs in duration spaced every 10 msec through each file. The cepstral features are augmented by 12 delta cepstral features calculated over 5-frame windows. Included in the analysis is a measurement of energy which is converted to a peak normalized log energy where the peak energy is calculated over the duration of the file. All data frames falling below a specified energy threshold are omitted in

	test broadcast number			
	1	2	3	4
	DURATIONS			
target	339.4	315.9	—	282.6
other speakers	945.2	856.6	1234.4	787.7
commercials	424.4	418.8	484.1	442.6
music	63.5	122.9	58.6	50.2
noise	6.1	58.9	1.0	45.8
silence	17.7	24.5	19.4	19.7
TOTAL	1796.3	1797.6	1797.6	1628.7
	TARGET SEGMENTS			
number	15	17	—	15
min duration	6.7	3.1	—	4.9
max duration	66.8	53.3	—	44.3
avg duration	22.6	18.6	—	18.8

Table 1. Test broadcast file segment statistics (in secs)

subsequent processing. The energy threshold is set at 30 dB below peak.

4. MODELLING

The target speaker and background speakers and other background audio categories are represented by Gaussian mixture models (GMM's) with diagonal covariance matrices. Gaussian mixture models are commonly and effectively used in text independent speaker recognition [2]. All the models for these experiments are constructed with 64 mixture components.

Descriptions of the experimental models are shown in Table 2. The table entries show the number of training segments used to construct the model, the total and average duration of the segments, and brief notes. The number of training vectors represents approximately a 10% to 20% reduction from the actual duration due to energy thresholding for the speaker models. The following should be noted about the content of the speaker models. Brokaw2 contains all the training segments used in Brokaw1. The Brokaw1 segments are generally labelled as mostly "clean", whereas the additional segments included in Brokaw2 contain some interfering speech, noise, and/or music background. About half of the speakers in Back1 are correspondents and the remainder, interviewees. Almost all of the speakers in Back3 are interviewees. The quality of the speech for interviewees is generally poorer than for correspondents. Four of the correspondent speakers in Back1 are also found in two of the test broadcasts. As far as can be determined, none of the speakers in Back3 are contained in Back1, and other than the correspondents noted, none of the Back1 or Back3 speakers are contained in test broadcasts.

model name	no. of segs.	total/avg. seg. dur. (secs)	notes
TARGET:			
Brokaw1	7	133/19.0	4 broadcasts
Brokaw2	13	267/20.5	8 broadcasts
BACKGROUND:			
Back1	20	275/13.8	9m, 11f speakers
Back3	23	270/11.7	13m, 10f speakers
Back1+3	43	545/12.7	22m, 21f speakers, sum of Back1 and Back3
BackComm1	20	468/23.4	commercials
BackMusic1	5	63/12.6	music

Table 2. Experimental model descriptions

5. SCORING AND TARGET SEGMENT DETECTION

Scoring a test sample proceeds as follows. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ be a sequence of feature vectors representing an audio test sample. Let λ_T be the target speaker GMM and $\lambda_{B_1}, \lambda_{B_2}, \ldots, \lambda_{B_K}$ be a set of background GMM's. Log likelihood scores are computed for each frame t of the feature vector sequence with respect to the target model and the background models as follows:

$$s_t(\lambda_T) = \log p(\mathbf{x}_t|\lambda_T) \tag{1}$$

$$s_t(\lambda_{B_k}) = \log p(\mathbf{x}_t|\lambda_{B_k}), k = 1, 2, \ldots, K \tag{2}$$

where $p()$ is the Gaussian mixture probability density function. Successive frame scores are averaged over blocks of M frames shifted every L frames through the sample. Thus, the j-th block score for the target model is given by

$$S_j(\lambda_T) = \frac{1}{M}\sum_{m=1}^{M} s_{t_j+m}(\lambda_T) \tag{3}$$

A likelihood ratio calculation between the target and background block scores produces a normalized score:

$$S_j(\lambda_T; \lambda_{B_1}, \ldots, \lambda_{B_K}) = S_j(\lambda_T) - \max_k S_j(\lambda_{B_k}) \tag{4}$$

Normalized scores are input to the detector to obtain estimates of the starts and ends of target speaker segments. The detector operation is based on the normalized score passing a double threshold test (in effect, a sequential decision test) to mark both the start and end of a target segment. Following is a fragment of program code (written in C) to describe the operation.

```
cand=0;
seg=0;
block=0;
while (block<Nblocks) {
  if (cand==0 && score[block]>th0) {
    tentstart=block;
    cand=1;
  }
  if (cand==1 && score[block]>th1) cand=2;
  if (cand==1 && score[block]<th0) cand=0;
  if (cand==2 && score[block]<th0)
    tentend=block;
  if (cand==2 && score[block]<th2) {
    start[seg]=tentstart;
    end[seg]=tentend;
    seg++;
  }
block++;
}
```

The thresholds are `th0`, `th1` and `th2`. In these experiments `th2` is set equal to `-th1` and `th0` < `th1`. `score[block]` is the normalized score for the current block. `seg` is the target segment counter. `cand` is a flag indicating the current status of a proposed target speaker segment.

An example of the output of the scoring and detection processes is shown in Fig. 1. The normalized score is shown as a function of time with each point plotted indicating a block score. The dashed vertical lines show actual segment boundaries while estimated target segment boundaries are indicated by the solid vertical lines. All the segments shown in this example are speaker segments. The "T" label indicates a target speaker segment; "C" and "A" indicate correspondent and announcer segments; "P" indicates interviewee segments.

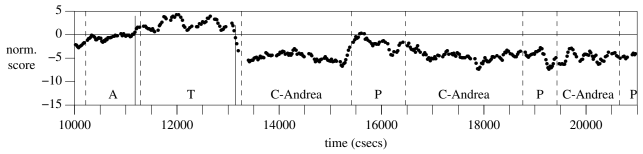

Figure 1. Normalized score in a portion of a test broadcast showing actual segment boundaries (dashed vertical lines) and estimated target segment boundaries (solid vertical lines). The "T" labels indicate target segments; other labels indicate other speakers.

6. PERFORMANCE MEASUREMENTS

Two types of performance measurements are used in the experiments reported here. The first is based on the number of test frames correctly segmented. (Although segmentations are specified in terms of frames they are actually calculated by blocks so that segmentation resolution is equal to the block shift.) The second is based on the number of segments correctly labelled. For frame segmentation performance, two error rates are calculated. The first, the Frame level MIss Rate (FMIR), is the fraction of actual target frames not included in estimated target segments. The second, the Frame level False Alarm Rate (FFAR), is the fraction of non-target frames included in estimated target segments.

Segment level performance is calculated as follows. A threshold fraction of frames correctly detected in a target segment is specified, denoted FCD. If the estimated fraction of frames correctly detected exceeds FCD, the segment is considered to be a "hit". If an estimated segment contains no target frames, the segment is considered a false alarm. Also, if an estimated segment contains more than one target segment, the excess number counts as false alarms. For example, if an estimated segment includes 2 target segments, it seems reasonable to count the non-target interval between the target segments as a false alarm.

Although these performance measurements are not completely satisfactory, particularly with respect to defining segment false alarms, they are consistent and logical and lead to useful measures of performance.

7. EXPERIMENTAL RESULTS

Experimental performance is evaluated by scoring and estimating target speaker segments in four 30-minute test broadcasts. The test broadcast segment statistics, based on manual labelling, are shown in Table 1. Only non-target performance is measured for test broadcast 3 since, as noted earlier, it contains no target speaker segments.

Performance is measured as a function of the following experimental variables, choice of target speaker model, choice of speaker background model, the number and content of background models, and the upper detection threshold, $th1$. The following experimental variables are held fixed: the number of mixture components (64), the content of the commercial and music background models, the energy threshold (30 dB below peak), the block shift (20 csecs) and block window size (120 csecs), and the lower detection threshold $th0$ (0).

Representative frame level error rates are shown in Fig. 2 as a function of $th1$ for the four test broadcasts. The target model used is Brokaw1 and two background models, Back3 and BackComm1 are used. The frame error measurements, FMIR (target frame miss rate) and FFAR (non-target detection rate) are plotted. As $th1$ increases detection accuracy can be expected to increase while the number of detected segments decreases. Both these trends imply that FFAR should decrease monotonically, which is more or less

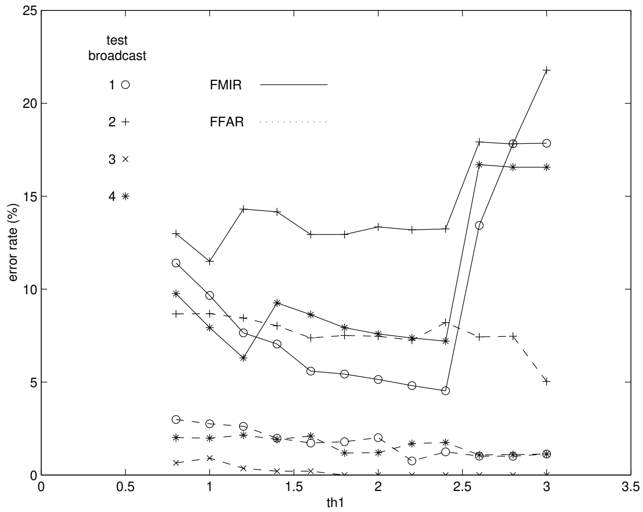

Figure 2. Frame based error rates as a function of $th1$ using target model Brokaw1 and background models Back3 and BackComm1

the case for each test broadcast. The behavior of FMIR is more complex. FMIR should decrease with increasing $th1$ as detection accuracy increases. However, sharp jumps in FMIR may occur when $th1$ increases enough to miss an entire segment. This is the behavior seen in the figure where the typical pattern of falling FMIR followed by sharp jumps repeats until $th1$ reaches a level sufficient to miss all target segments when FMIR increases to 100%.

For a fairly broad range of values of $th1$ between 1.8 and 2.2 the sum of FMIR and FFAR is roughly minimized. These values of FMIR and FFAR can be used to compare the overall performance for each test broadcast. It can be seen that test broadcast 1 has the best performance with FMIR approximately 5% and FFAR 1%, test broadcast 2, the worst with FMIR approximately 14% and FFAR 7%, and test broadcast 4, FMIR about 8% and FFAR about 1%. A FMIR of 10% translates into a segmentation error of 2 secs for a target segment with the average duration of 20 secs. There are no target segments in test broadcast 3, but it has the lowest FFAR, essentially 0 in this threshold region. The overall balance between FMIR and FFAR can changed somewhat by adjusting $th0$.

target	spkr back	FMIR (%)	FFAR (%)	sum (%)	mi's (15)	fa's
speaker background only						
Brokaw1	Back1	4.3	7.8	12.1	0	4
Brokaw1	Back3	4.3	2.5	6.8	0	0
Brokaw1	Back1+3	7.8	4.2	12.0	1	2
Brokaw2	Back3	4.1	7.0	11.1	0	6
speaker background + BackComm1						
Brokaw1	Back1	5.2	1.2	6.4	0	1
Brokaw1	Back3	4.8	0.8	5.6	0	0
Brokaw1	Back1+3	5.6	1.3	6.9	0	0
Brokaw2	Back3	4.0	4.7	8.7	0	3
speaker background + BackComm1 + BackMusic1						
Brokaw1	Back3	4.6	1.2	5.8	0	0

Table 3. Performance comparisons for different selections of target and background speaker models for test broadcast 1. The target seg hit criterion, FCT, is set at 80%. There are 15 target segments.

Performance variations associated with selections of target and speaker background models and the number of background models are illustrated in Table 3 for test broadcast 1. The performances shown are obtained by adjusting

$th1$ to obtain the minimum sum of FMIR and FFAR error. In some cases $th0$ was also varied in an attempt (not always successful) to make the FMIR values as consistent as possible from one condition to another. The number of background models is varied from 1 to 3. The commercials and music background models are not varied.

In addition to FMIR and FFAR, segment level performance figures are shown. These are the number of target segment hits with FCD set to 80% and the number of segment false alarms.

Consider first the effect of adding background models in addition to the speaker background model. For each combination of target and speaker background model, there is a sizeable reduction in overall error rate from the speaker background only condition to the condition in which a commercials background model is added. However, the further addition of the music background model produces no additional improvement. The performance improvement obtained with the addition of the commercials background model is mainly associated with a reduction in FFAR and segment false alarms.

The selection of a speaker background model is compared using Back1, Back3, and Back1+3 models together with the Brokaw1 target model. As noted earlier, the speakers in Back1 are both correspondents and interviewees, while the speakers in Back3 are almost all interviewees. Back1+3 is made up of all the training segments in Back1 and Back3. Back3 is seen to perform better than Back1. Only for test broadcast 2 does Back1 perform better than Back3. This may be because 2 of the correspondents in test broadcast 2 are also included in Back1. Using Back1+3, containing all training segments contained in Back1 and Back3 does not provide any improvement over using Back1 or Back3 alone. In fact, it performs only marginally better than the worse performing model, Back1.

Target model selection is compared using Brokaw1 and Brokaw2 target models in combination with Back3 as the speaker background model. Brokaw2 contains the 7 training segments found in Brokaw1 plus 6 additional segments. Since Brokaw2 contains about twice as much training material as Brokaw1 it might be predicted to provide some improvement. In fact, Brokaw2 performs worse than Brokaw1. The degradation might be attributed to the fact that many of the additional segments contained in Brokaw2 are not "clean". They contain contaminants from the addition of speech or noise to the target speaker speech. It seems likely that a "clean" target model is necessary to accurately detect target speaker segments against such contaminants.

8. CONCLUSION

The experimental results are shown summarized in Table 4 for the best experimental selections of thresholds and models. The selected models are Brokaw1 and Back3 except for test broadcast 2 which uses Brokaw1 and Back1. At the segment level of performance, the fractions of target segments misssed, using an 80% FCD threshold, are 0/15, 3/17, and 1/15 for test broadcasts 1, 2, and 4 which target speaker segments. The number of false alarm segments for each half-hour test broadcast is 0 except for test broadcast 2 which has 1.

We conclude that although the precision with which target segments are detected is not especially high (measured by FMIR), the number of target segments detected is high and the number of false alarms is low. Thus overall performance is quite satisfactory for the intended application which is locating anchor person segments in news broadcasts for browsing and indexing.

The variations in performance from one test broadcast to another are somewhat surprising. In particular, the relatively high FMIR rate for test broadcast 2 is unexplained.

test broadcast	FMIR (%)	FFAR (%)	sum (%)	misses	fa's
1	4.8	0.8	5.6	0	0
2	13.8	2.3	16.1	3	1
3	–	0.0	–	–	0
4	7.6	1.2	8.8	1	0

Table 4. Performance for each test broadcast for the best combination of target and speaker background model and the best threshold condition.

There does not seem to be any obvious reason for the 2 or 3 persistently missed target segments in this broadcast.

The requirements for good target and background models are not completely understood. In most speaker recognition applications including more training data in the models improves performance. Our experiments suggest that the content of the training data is at least as important. We have seen that detection performance is sensitive to the presence of "contaminated" training data in the target speaker model. Also, no improvement in performance is observed using a speaker background model with twice as much data as the best performing model which includes the data used in that model.

Further studies in speaker detection will include more than one target speaker in broadcast databases and detecting speakers in a teleconference database.

REFERENCES

[1] C. Montacie and M-J. Caraty, Sound channel video indexing, *Proc. Eurospeech 97,* Fifth European Conference on Speech Communication and Technology, Rhodes, 2359-2362, 1997.

[2] D. A. Reynolds and R. C. Rose, Robust text-independent speaker identification using Gaussian mixture models, *IEEE Trans. on Speech and Audio Processing*, **3**, 72-83, 1995.

[3] D. Roy and C. Malamud, Speaker identification based text to audio alignment for an audio retrieval system, *Proc. ICASSP 97*, IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Munich, 1099-1102, 1997.

[4] M-H. Siu, G. Yu, and H. Gish, An unsupervised, sequential learning algorithm for the segmentation of speech waveforms with multiple speakers, *Proc. ICASSP 92*, IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, San Francisco, vol. II, 189-192.

[5] M. Sugiyama, J. Murakami, and H. Watanabe, Speech segmentation and clustering based on speaker features, *Proc. ICASSP 93* IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, 395-398, 1993.

[6] L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian, Segmentation of speech using speaker identification, *Proc. ICASSP 94*, IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Adelaide, 161-164, 1994.