STRATEGIES FOR AUTOMATIC SEGMENTATION OF AUDIO DATA

Thomas Kemp *Michael Schmidt* *Martin Westphal* *Alex Waibel*

Interactive Systems Laboratories, ILKD
University of Karlsruhe
76128 Karlsruhe, Germany

ABSTRACT

In many applications, like indexing of broadcast news or surveillance applications, the input data consists of a continuous, unsegmented audio stream. Speech recognition technology, however, usually requires segments of relatively short length as input. For such applications, effective methods to segment continuous audio streams into homogeneous segments are required.
In this paper, three different segmenting strategies (model-based, metric-based and energy-based) are compared on the same broadcast news test data. It is shown that model-based and metric-based techniques outperform the simpler energy-based algorithms. While model-based segmenters achieve very high level of segment boundary precision, the metric-based segmenter performes better in terms of segment boundary recall (RCL). To combine the advantages of both strategies, a new hybrid algorithm is introduced. For this, the results of a preliminary metric-based segmentation are used to construct the models for the final model-based segmenter run. The new hybrid approach is shown to outperform the other segmenting strategies.

1. INTRODUCTION

The segmentation of audio data is of interest for a broad class of applications, like surveillance applications, meeting summarization or indexing of broadcast news. In this work, we mainly focus on the indexing of broadcast news in the context of our multimedia information indexing and retrieval system 'View4You'. The goal of the View4You project is to provide content based natural language access to information in TV news. For this, news shows are regularly collected from public TV and transcribed using an ASR system. The transcription is used as index for the information retrieval system.

There are at least three reasons why segmentation is required in a broadcast news indexing system like View4You. First, speech recognition technology usually requires segments of relatively short length as input. *Any* segmentation for which the maximum segment length does not exceed the capability of the speech recognizer, and that avoids cutting within words, satisfies this requirement.

Second, as speakers tend to repeat within a given news show, speaker adaptation schemes can be used to improve ASR performance. This is usually done by an initial segmenter run, followed by a clustering step that tries to group segments from one speaker together. The speaker adaptation for a given segment is carried out using all segments of the corresponding cluster. A segmentation that is used for speaker adaptation needs to have a high *segment purity*, e.g. one segment should contain only one single speaker and acoustic condition. A speaker turn can be segmented into two or more segments without harm, since over-segmentation is unproblematic due to the clustering step.

The third reason for segmentation in a BN retrieval system is user friendliness. For a given query topic, it is rarely an appropriate answer to return one complete, unsegmented news show and let the user decide which part of it is of interest. Ideally, the information system presents only the relevant parts. For this, however, the starting point and length of each story must be determined by the segmenter. Since it is disturbing to have either too short segments, where part of the information remains hidden, or too long segments where irrelevant information is displayed, the segment boundary must match the story boundary as exactly as possible.

Recently, several groups have investigated the problem of segmenting broadcast news in the context of ARPA's hub-4 broadcast news transcription and understanding evaluations ([1] [2][3][4][5][6][7][8]). The goal of the segmentation in the ARPA-supported experiments was mainly to provide a basis for speech recognition and speaker adaptation, not to find the true story boundaries as required by a retrieval system. For the View4You segmentation, however, it is necessary to find the true story boundaries, and therefore the evaluation in this paper measures how well different segmenting approaches can find the story boundaries. The main difference to the more common segment purity or word error rate based segmenter evaluations is, that oversegmentation matter in our case, but does not matter (much) if segment purity or WER is measured.

The different approaches which have been used in the ARPA evaluations can be categorized into three classes [3]:

- Model-based segmentation. Different models, e.g. Gaussian mixture models, are constructed for a fixed set of acoustic classes, such as anchor speaker, music etc, a training corpus. The incoming audio stream can be classified by maximum likelihood selection. Segment boundaries are assumed where a change in the acoustic class occurs.
- Metric-based segmentation. The audio stream is segmented at maxima of the distances between neighbouring windows placed in evenly spaced time intervals.
- Energy-based segmentation. Silence in the input audio stream is detected either by a decoder or directly by measuring and thresholding the audio energy. The segments are then generated by cutting the input at silence locations.

All segmenting approaches are reported to work reasonably well for speech recognition and speaker adaptation. However, no evaluation has yet been carried out to examine how well the different algorithms work when applied to the problem of finding the true

story boundaries.

In this paper we present the results of a comparative evaluation on a common test set. Then, a new hybrid segmentation strategy is proposed, and the results of the new segmenter are compared to those of the three standard approaches mentioned earlier.

2. EVALUATION SETUP

All experiments have been carried out in the framework of the View4You video indexing and retrieval system.

A detailed description of the View4You system is given in [10].

2.1. Database

The TV news shows are received from a television satellite and stored as MPEG-compressed files. The audio data is compressed with MPEG audio layer 2 compression at a data rate of 192 kbit/s and a sampling rate of 44.1 kHz. The recorded audio signal is sampled down to a 16 kHz/16 bit PCM format which was used throughout all our experiments. The video part of the signal was not used for the experiments described in this paper.

To evaluate the performance of the different segmenting approaches, four German news shows (dated 03/30/97, 04/13/97, 05/28/97 and 06/30/97) were used. Each of the news shows is approximately 15 minutes long. The four shows were manually segmented into topic stories. If there was a speaker change or a change from anchor speaker to field speech *within* one topic, an additional segment boundary was introduced at the location of the change. The resulting manual segmentation, which can be used both for acoustic adaptation and information retrieval, is taken as the reference for the evaluation.

The reference contained 141 segment boundaries, which corresponds to an average segment length of approximately 25 seconds.

2.2. Evaluation metrics

The result of a segmentation can contain two possible types of error. Type-I-errors occur if a true segment boundary has not been spotted by the segmenter (deletion). Type-II-errors occur if a found segment boundary does not correspond to a segment boundary in the reference (false alarm, or segment insertion). The information retrieval community uses two closely related numbers, precision (PRC) and recall (RCL). Precision and recall can be expressed by Type-I-error rate and Type-II-error rate, and vice versa. They are defined as

$$\text{RCL} = \frac{\text{number of correctly found boundaries}}{\text{total number of boundaries}}$$

$$\text{PRC} = \frac{\text{number of correctly found boundaries}}{\text{number of hypothesized boundaries}}$$

Most segmentation algorithms can be made to work at different operating points. Each operating point corresponds to a (PRC,RCL) pair. As the relative cost of a missed boundary versus the cost of a false alarm depends on the application, a segmenter is fully characterized by a plot of Precision over Recall for all possible operating points. Such a plot is referred to as 'receiver operator characteristic'.

Sometimes it is desirable to have one single number for the performance of an algorithm instead of two. In such cases, the F-measure F is frequently used [13]. It can be parameterized to put higher weight to either PRC or RCL. The neutral parametrization, where Precision and Recall are weighted equally, is used throughout this work. F is defined as

$$F = \frac{2 * PRC * RCL}{PRC + RCL} \quad (1)$$

Like RCL and PRC, it is bounded between 0 and 1.

The correct position of a segment boundary is not exactly defined. In most cases, two segments are separated by a short period of silence. Any segment boundary within the silence period should be regarded as correct. Therefore, a tolerance Δt is defined. If a segment boundary is hypothesized within the time interval $t_0 - \Delta t < t < t_0 + \Delta t$ of the reference boundary t_0, it is judged correct. For our experiments, we chose Δt = 1.5 s.

3. EXPERIMENTS

3.1. Energy-based segmentation

Energy-based approaches have been widely used (e.g. see [8], [4]) and are particularly easy to implement. Basically, silence periods in the input signal are detected, and segment boundaries are hypothesized in such silence periods if some additional constraints are satisfied, like minimum length of the silence period.

In our energy-based segmenter, the power of the input signal is measured every 10 ms and smoothed using a 9-frame FIR filter. The smoothing implicitely imposes a minimum length constraint on the silence period. A threshold is applied, and the regions of the signal that have their energy below the threshold are categorized as silence. A segment boundary is assumed in the center of each silence region.

In this segmenter type, the operating point can be easily adjusted by changing the energy threshold.

The best result in terms of F-measure on the four-show test set is shown in table 1.

Algorithm	RCL	PRC	F-measure
energy-based	0.62	0.54	0.58

Table 1. Performance of the energy-based segmenter

3.2. Model-based segmentation

In model-based segmentation [1][6], a set of models for different acoustic classes is defined and trained prior to segmentation. The incoming audio stream is classified using the models, usually imposing additional minimum class length constraints. Boundaries between the classes are used as segment boundaries. Model-based segmentation assumes knowledge about the type of the audio that is to be segmented.

In our model-based segmenter, a speech recognizer was used with a four-word dictionary ('Anchor', 'Field', 'Music' and 'Silence'). The corresponding HMM states used diagonal variance gaussian mixture models (GMMs) as emission probabilities. The GMMs were trained on two hours of manually labelled audio. The audio data used for training was disjunct from the four newscasts used for testing. The number of mixture components per class was chosen according to table 2. By duplicating HMM states, a minimum word duration as shown in table 2 was enforced. No state transition probabilities and no language model were used.

In the acoustic preprocessing, 16 mel-spectral parameters were computed every 50 msec, using a 16 msec window. Although this parameterization does not make use of two thirds of the signal, it performed equally well as compared to a frameshift of 10 msec, but requires only one third of the computing time. Mel-cepstral parameters led to performance degradation and were therefore replaced by the mel-spectral parameters. [1] proposed to perform MLLR adaptation on the segments resulting from the first run, and

class	number of mixtures	minimum length
anchor speaker	128	5 sec
field speech	128	5 sec
music	32	2.5 sec
silence	2	0.2 sec

Table 2. Parameters for the model-based segmenter

then re-run the segmenter with the adapted models. However, using this method did not improve the performance on our data.

The 'word' boundaries in the hypothesis of the recognizer were used as segment boundaries. Different operating points could be achieved by changing the value of the word insertion penalty during the search: a high word insertion penalty led to fewer words in the hypothesis and hence fewer segment boundaries.

The results for the operating point yielding the highest F-measure are shown in table 3.

Algorithm	RCL	PRC	F-measure
model-based	0.56	0.70	0.62

Table 3. Performance of the model-based segmenter

3.3. Metric-based segmentation

In metric-based approaches, two neighbouring windows of relatively small size are moved over the audio signal. The similarity between the contents of the two windows is computed using a distance function. Local maxima exceeding a threshold indicate segment boundaries.

Various metric-based algorithms differ in the kind of the distance function they employ, the size of the two windows, the time increments of the shifting of the two windows, and the way the resulting similarity values are evaluated and thresholded.

[7] proposed to use a symmetrized Kullback-Leibler distance metric. For two PDFs P_A and P_B describing the two neighbouring windows, the symmetrized Kullback-Leibler distance is defined as

$$KL(A,B) := \frac{1}{2}\int_x (P_A(x)\log\frac{P_A(x)}{P_B(x)} + P_B(x)\log\frac{P_B(x)}{P_A(x)})dx \tag{2}$$

[7] used single gaussian distributions to represent each of the two windows. For gaussian distributions P_A and P_B, there is a closed-form solution to (2).

[3] proposed a distance measure introduced by Gish [9]. This metric is basically a likelihood ratio hypothesis test. The null hypothesis assumes that the data in both neighbouring windows has been produced by the same stochastic source. The alternate hypothesis assumes that the data has been produced by different sources. The likelihood ratio LR

$$LR := \frac{(N_A+N_B)\quad L(A+B|\lambda_{A+B})}{N_A L(A|\lambda_A)\quad N_B L(B|\lambda_B)} \tag{3}$$

is used to decide whether the null hypothesis is true (no segment boundary is assumed) or not (a segment boundary is introduced). For the actual computation of the likelihoods in (3), one multivariate full covariance gaussian model is estimated on each of the windows and on the union of the two windows. The threshold, which is used to decide whether a segment boundary is introduced or not, can be determined by an information theoretical measure (minimum description length MDL [12], Bayesian information criterion [3]). This approach has the advantage that no thresholds need to be defined and hence the algorithm is robust against changes of the acoustic conditions. In our experiments, however, the thresholds determined by MDL and BIC yielded significantly lower F-scores than the best possible threshold selection. Therefore, we did not use MDL or BIC in our experiments.

The *entropy loss* when coding the two windows separately instead of coding them with one global probability distribution is defined as

$$EL := N_{A+B}H_{A+B} - N_A H_A - N_B H_B \tag{4}$$

where H_x is the entropy of the probability distribution x. Entropy loss has been extensively used as a distance function, e.g. for triphone clustering [11]. In this work, we investigated its use for segmentation. In our implementation, a generic codebook of 32 gaussians is trained offline on all training data. The test data in each window is vector quantized using this codebook. The resulting discrete probability distribution is used to compute the entropy H_x.

The results achieved with the three different distance functions are summarized in table 4. The entropy loss does not seem to be effective.

distance metric	RCL	PRC	F-measure
Kullback-Leibler	0.81	0.60	0.69
Gish-distance	0.80	0.63	0.70
Entropy loss	0.61	0.26	0.37

Table 4. Performance of the metric-based segmenters

3.4. Results

In the previous paragraph, the best results achieved over all operating points of the various algorithms were shown. It is interesting to see, which range of operating points can be reached by each of the algorithms. Therefore, each of the segmenting algorithms was run at several operating points and the resulting pairs of precision and recall were computed on the 4 news show test set. The results are shown in figure 1.

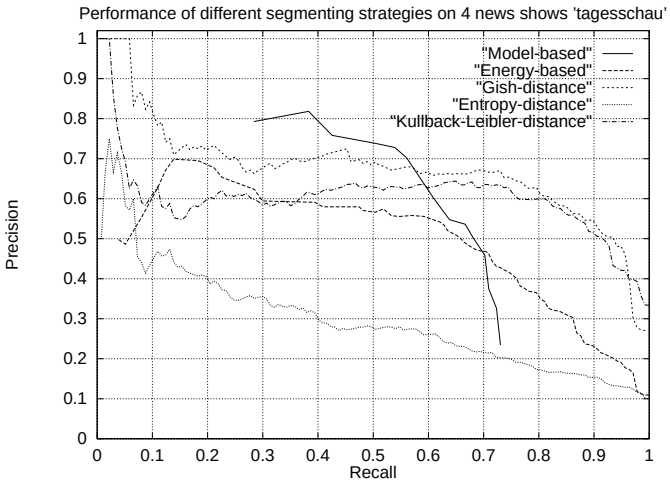

Figure 1. Result of different segmenting algorithms

The results can be summarized as follows.

- Model-based and metric-based algorithms outperform the simple energy-based approaches

- Model-based segmenters achieve high precision at moderate recall
- Metric-based segmenters are capable to achieve high recall at moderate precision

4. THE HYBRID SEGMENTER

From the previous paragraph it can be concluded, that a combination of model-based and metric-based algorithms could provide both the high recall of metric-based algorithms and the high precision of model-based approaches. The Gish function yielded the best performance and is therefore used as the distance function. The following new hybrid segmenting algorithm was examined:

1. Chop the input signal into many chunks of equal length
2. Perform a bottom-up clustering of the chunks using the Gish distance function until N clusters remain
3. Train a M-mixture GMM model for each of the N clusters
4. Run a model-based segmenter that makes use of the GMM cluster models

We chose the chunk length as 1 second, $N = 6$ and $M = 64$. The model-based segmenter was constructed as described in the previous paragraph. As there is no *a-priori* knowledge about the type of data contained in each cluster, all clusters had the same minimum duration of 1.5 seconds and the same number of gaussians (64). The results of the final segmenter are shown in table 5 and figure 2.

Algorithm	RCL	PRC	F-measure
new hybrid	0.67	0.93	0.78

Table 5. Performance of the new hybrid segmenter

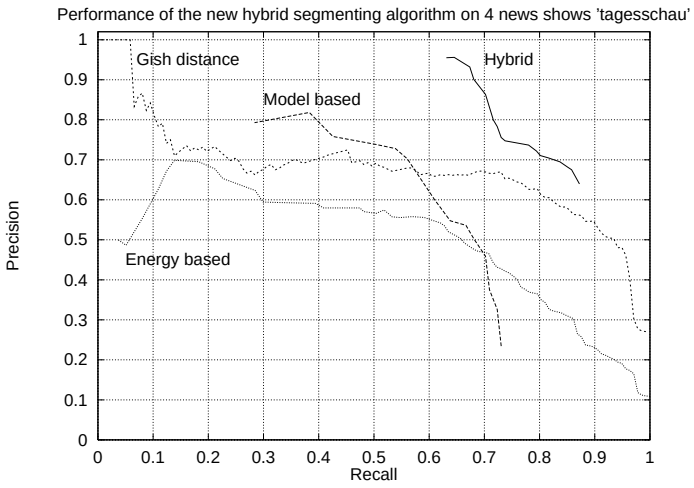

Figure 2. Result of different segmenting algorithms

Except at very high levels of recall, the new segmentation algorithm works significantly better than all other approaches. Since it does not require any precomputed models, the algorithm should robustly generalize to unseen types of audio input.

5. CONCLUSIONS

Different segmenting strategies for audio data have been compared on a common broadcast news data test set. While simple energy-based worked quite well, they were outperformed by both model-based and metric-based algorithms. Model-based algorithms were shown to yield high precision at moderate recall, whereas metric-based algorithms result in high recall at moderate precision. A new hybrid algorithm is proposed that combines metric-based and model-based techniques. The new algorithm successfully combines the advantages of both approaches into one system.

6. ACKNOWLEDGEMENTS

The views and conclusions contained in this document are those of the authors.
The authors wish to thank all members of the Interactive Systems Labs, especially Florian Metze and Manfred Weber, for useful discussions and active support.

REFERENCES

[1] P.C.Woodland, T. Hain, S. Johnson, T. Niesler, A. Tuerk, S. Young, *Experiments in broadcast news transcription*, Proc. ICASSP 1998, pp. 909 ff, Seattle, Washington, May 1998

[2] L. Polymenakos, P. Olsen, D. Kanvesky, R. Gopinath, P. Gopalakrishnan, S. Chen, *Transcription of broadcast news - some recent improvements to IBM's LVCSR system*, Proc. ICASSP 1998, pp. 901 ff, Seattle, Washington, May 1998

[3] S. Chen, P. Gopalakrishnan, *Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion*, DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, Feb 8-11, 1998

[4] S. Wegmann, F. Scattone, I. Carp, L. Gillick, R. Roth, J. Yamron, *Dragon System's 1997 broadcast news transcription system*, DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, Feb 8-11, 1998

[5] J.-L. Gauvain, L. Lamel, G. Adda, *The LIMSI 1997 Hub-4E Transcription System*, DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, Feb 8-11, 1998

[6] A. Sankar, F. Weng, Z. Rivlin, A. Stolcke, R. Gadde, *The development of SRI's 1997 broadcast news transcription system*, DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, Feb 8-11, 1998

[7] M. Siegler, U. Jain, B. Raj, R. Stern, *Automatic segmentation, classification and clustering of broadcast news audio*, Proc. of the DARPA Speech Recognition workshop, The Westfields Conference Center, Chantilly VA, Feb 2-5, 1997

[8] H. Wactlar, A. Hauptmann, M. Witbrock, *Informedia: news-on-demand experiments in speech recognition*, Proc. of the ARPA SLT workshop, 1996.

[9] H. Gish, M.H. Siu, R. Rohlicek, *Segregation of Speakers for Speech Recognition and Speaker Identification*, in Proc. ICASSP-91, S. 873 ff

[10] T. Kemp, P. Geutner, M. Schmidt, B. Tomaz, M. Weber, M. Westphal, A. Waibel, *The Interactive Systems Labs View4You video indexing system*, Proc. ICSLP 98, Vol 4, pp. 1639 ff, Sydney, Australia, December 1998

[11] Kai-Fu Lee, *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System*, Ph. D. thesis, CMU-CS-88-148, Carnegie Mellon University, Pittsburgh, PA 15213, April 1988

[12] J. Rissanen, *Universal coding, information, prediction, and estimation*, IEEE transactions on information theory, Vol. IT-30, No. 4, July 1984

[13] C. J. van Rijsbergen, *Information Retrieval*, London, Butterworth, 1979, pp. 174 ff.