AUDIO SEGMENTATION FOR SPEECH RECOGNITION USING SEGMENT FEATURES

David Rybach, Christian Gollan, Ralf Schlüter, Hermann Ney

Human Language Technology and Pattern Recognition
Computer Science Department, RWTH Aachen University, Germany
{rybach,gollan,schluter,ney}@cs.rwth-aachen.de

ABSTRACT

Audio segmentation is an essential preprocessing step in several audio processing applications with a significant impact e.g. on speech recognition performance. We introduce a novel framework which combines the advantages of different well known segmentation methods. An automatically estimated log-linear segment model is used to determine the segmentation of an audio stream in a holistic way by a maximum a posteriori decoding strategy, instead of classifying change points locally. A comparison to other segmentation techniques in terms of speech recognition performance is presented, showing a promising segmentation quality of our approach.

Index Terms— speech recognition, audio segmentation, broadcast news transcription

1. INTRODUCTION

Audio segmentation is a vitally important task in several audio processing applications like speaker diarization, speaker tracking, and automatic speech recognition (ASR). The requirements on the segmentation differ depending on the application. This paper concentrates on speech recognition applications, in particular broadcast news transcription.

The quality of the segmentation affects the recognition performance in several ways: Speaker adaptation and speaker clustering methods assume that a segment is spoken by a single speaker. The language model performs better if segment boundaries correspond to boundaries of sentence-like units [1]. Furthermore, non-speech regions, like music and other sounds often occurring in broadcast news shows, may cause insertion errors and should be detected and removed. Regions with overlapping speech are not recognized correctly in most cases and should be separated to limit the impact on the recognition of surrounding speech. Obviously, segment boundaries positioned inside a spoken word deteriorate the recognition quality as well.

Various segmentation methods have been investigated in the literature, which can be categorized into 3 classes [2]. *Decoder-guided* segmentations use the output of a speech recognition system, e.g. silence regions recognized. *Model-based* approaches classify acoustic regions and divide the audio stream at changes in the acoustic class. A distance between adjacent regions in the audio stream is used in *metric-based* methods, e.g. the KL distance or the Bayesian information criterion. Techniques relying solely on acoustic features tend to produce longer segments, because their focus is often primary to detect changes in acoustic conditions. In [1, 3] it was shown that the usage of ASR output improves the segmentation w.r.t. recognition performance.

The method proposed in this paper combines the advantages of ASR-, model-, and metric-based approaches by incorporating several features of a segment. Furthermore, the decision of positioning a segment boundary is not made locally, but by considering the segmentation of the complete audio stream. This global optimization of the segmentation is done by a form of maximum a posteriori (MAP) decoding. The model used for segment classification has been estimated on a manual segmentation. A preliminary version of our approach was presented in [4]. Moreover, we trained a classifier to label the automatically generated segments and to remove non-speech segments.

Since the objective of this task is to improve the recognition performance, we measured the segmentation quality in terms of word error rate obtained on the segmented audio data.

In the remainder of this paper, the broadcast news transcription system used is described in Section 2, and the proposed segmentation method in Section 3. Experimental results are presented in Section 4, and conclusions are drawn in Section 5.

2. RECOGNITION SYSTEM

In the acoustic front end, consisting of MFCC features, vocal tract length normalization (VTLN) is applied to the filterbank. The VTLN warping factors are estimated online by a Gaussian mixture classifier. For the recognition of unsegmented data, the warping factors are estimated every second on a sliding window (7s wide). On segmented data, the warping factor is estimated segment-wise.

In an initial recognition pass, the unsegmented data is processed with a speaker independent acoustic model. During decoding, transitions from "sentence-end" to "sentence-start" are hypothesized, too (as proposed by [1, 5]). The results of this recognition pass, including confidence scores, are used for the segmentation process. The resulting segments are clustered using a generalized likelihood ratio clustering with Bayesian information criterion based stopping condition. The segment clusters act as speaker labels as required by the adaptation techniques applied.

The recognition of the segmented data is performed in two passes. The output of a speaker independent recognition pass is used as input for the text dependent speaker adaptation. CMLLR feature transformations and MLLR mean transformations are estimated using the segment clusters. The transformed features and models (the models were trained speaker adaptively) are used during the second speaker dependent recognition pass.

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation. This material is partly based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023.

3. AUDIO SEGMENTATION

3.1. Segmentation Framework

The segmentation task can be formulated as an optimization problem for a given set of possible boundaries. Each time frame t corresponding to a possible boundary is assigned to the class "boundary" (class 1) or to the class "no boundary" (class 0): $b_1^T = (b_1, \ldots, b_T)$ with $b_t \in \{0, 1\}$. Instead of classifying a boundary independent of all other boundaries, we optimize the complete segmentation:

$$\hat{b}_1^T = \underset{b_1^T:\, b_T=1}{\operatorname{argmax}} \Big\{ p(b_1^T \mid X_1^T) \Big\}$$

where X_1^T is a sequence of arbitrary features which will be specified in Section 3.2. Using this formulation we can consider *segments* in place of boundaries, which allows us to use context dependent features of the segmentation. A sequence b_1^T implies the segments

$$[t_i, t_j] \text{ with } b_{t_i} = b_{t_j} = 1 \text{ and } b_t = 0 \; \forall t_i < t < t_j \; .$$

For the first segment we define $b_0 := 1$.

The probability of a sequence b_1^T for given features X_1^T is

$$p(b_1^T | X_1^T) = \prod_{t=1}^{T} p(b_t \mid b_1^{t-1}, X_1^T) \; .$$

We assume that the probability of a time frame t being in class b_t depends on the time frames t', t of the segment $[t', t]$ it terminates and on the features for this segment. By introducing the index sequence

$$\tau_1^N := \{\tau \mid b_\tau = 1,\; 1 \le \tau \le T\}$$

of the boundaries in b_1^T and by using the dependency model assumption, $p(b_1^T | X_1^T)$ can be modeled by

$$\begin{aligned} p(b_1^T | X_1^T) \quad = \quad & \prod_{n=1}^{N} \Big(p(b_{\tau_n} \mid \tau_{n-1}, \tau_n, X_1^T) \\ & \cdot \prod_{t=\tau_{n-1}+1}^{\tau_n - 1} p(b_t \mid \tau_{n-1}, t, X_1^T) \Big) \end{aligned}$$

where $p(b_{\tau_n} \mid \tau_{n-1}, \tau_n, X_1^T)$ is the probability for the right boundary of segment $[\tau_{n-1}, \tau_n]$ ($b_{\tau_n} = 1$ for all n). The second product gives the probability that all time frames inside this segment belong to class "no boundary" ($b_t = 0 \; \forall t \notin \tau_1^N$).

We model the class posterior probability of a segment boundary directly using a log-linear model:

$$p(b | t, t', X_1^T) = \frac{\exp\left(\sum_i \lambda_{i,b} \, f_i(t, t', X_1^T)\right)}{\sum_{c=0}^{1} \exp\left(\sum_i \lambda_{i,c} \, f_i(t, t', X_1^T)\right)}$$

where the $f_i(t, t', X_1^T)$ are feature functions, which will be defined in the next section, and the $\lambda_{i,c}$ are class specific feature weights.

The optimization problem over the 2^{T-1} possible segmentations can be solved using dynamic programming, utilizing the first order dependency of a segment boundary on the previous boundary. The auxiliary function $Q(t)$ gives the probability of the best segmentation ending at time t:

$$\begin{aligned} Q(t) \quad & = \max_{t' < t} \Big\{ p(b_1^{t'} \mid X_1^T) \cdot p(1 \mid t', t, X_1^T) \cdot Z(t', t-1) \Big\} \\ & = \max_{t' < t} \Big\{ Q(t') \cdot p(1 \mid t', t, X_1^T) \cdot Z(t', t-1) \Big\} \\ Q(0) \quad & := 1 \end{aligned}$$

Where

$$Z(t_b, t_e) = \prod_{t=t_b+1}^{t_e} p(0 \mid t_b, t, X_1^T)$$

is the probability that no boundary occurs inside the segment $[t_b, t_e + 1]$.

In our implementation we constrain the search space by allowing only segments with a length between 1 and 30 seconds. Furthermore, our implementations uses sums of negative logarithms of the probabilities for reasons of efficiency and numerical stability.

Boundaries are hypothesized at time frames, where either silence, sentence end, or noise has been recognized. In principle, other change point detection methods can be applied to generate a set of boundary hypotheses. However, in this paper we consider only ASR-based boundaries.

3.2. Segment Features

The feature functions $f_i(t', t, X_1^T)$ used in the previous section represent specific features of a segment $[t', t]$. We analyzed the following features in the experiments presented:

segment length The length of the segment.

words The number of words recognized inside the segment.

boundary length The length of the boundary token (i.e. the silence or non-speech region) recognized at the end of the segment.

boundary confidence The confidence score of the boundary token recognized.

warping factor variance The variance of the VTLN warping factors classified inside the segment (see Section 2). This feature should give an estimate of the speaker homogeneity in a segment.

BIC score We apply the Bayesian Information Criterion (BIC) to the acoustic feature vectors (unnormalized MFCCs) in a window around the segment end and compute the BIC "distance" [6, 7]. This distance yields a feature for changes in the acoustic condition. If there is silence recognized at the segment end, we remove the corresponding feature vectors and enlarge the window accordingly. Hereby we can detect for example a speaker change with a longer pause between the two speakers.

signal type From the results of the speech/non-speech detection (see Section 3.4) we compute the degree of signal type homogeneity in the segment as the proportion of the maximum over the length of time frames labeled as non-speech, music, or speech relative to the segment length.

sentence end A binary feature which is 1 if the boundary token recognized at the end of the segment is a sentence end symbol.

Furthermore, a zero-th feature $f_0(t', t, X_1^T) := 1$ is added to integrate an offset value. Second-order features can be augmented or used instead of the first-order features:

$$f_{i,j}(t', t, X_1^T) := f_i(t', t, X_1^T) \cdot f_j(t', t, X_1^T) \; , \; i \ge j$$

The framework proposed is not limited to ASR-based features, but can incorporate any segment or boundary dependent feature.

3.3. Segment Model Training

In the training process the feature weights $\lambda_{i,c}$ are optimized in a discriminative way according to the maximum entropy criterion using the generalized iterative scaling (GIS) algorithm.

First, we automatically transcribed the recordings of the training corpus using the boundary-enabled recognition system (Section 2). Then feature vectors for class "boundary" were computed for all segments corresponding to a segment in the reference transcription (Section 4.2). Feature vectors for class "no boundary" were calculated for segments with the reference segment beginning as start time frame and an end time frame at any possible boundary either before the reference segment end or inside the next reference segment. Thus, we assign both too long and too short segments to class "no boundary".

3.4. Speech vs. Non-speech Detection

The speech/non-speech detection system consists of 3-state HMMs for the signal types speech, non-speech, pure music, and a 1-state model for silence. As suggested in [8], a separate model for speech in the presence of background noise was added. Furthermore, we used gender dependent speech models. Gaussian mixtures are used as emission models. In total we used 576 densities for the 19 HMM states with mixture specific diagonal covariance matrices.

The mixture models were trained on 9.5h audio material (training corpus, cf. 4.2) labeled with the 5 classes. Silence is not labeled in the reference transcriptions, but was hypothesized during the forced Viterbi alignment of pure speech segments. Thus, the silence model should not contain background noise.

The acoustic front end consists of MFCC features (13 coefficients, mean and variance normalized) with first and second derivatives. Similar to [8], the zero-th coefficient is not included, but its derivatives.

To incorporate a-priori probabilities for the individual signal types, we estimated a bigram "language model". This model consists of relative frequencies of the signal type bigrams in the training corpus. To include silence in the model, we used the results of the forced alignment of the training corpus. Using this model, we incorporate also transition probabilities automatically estimated for intra signal type HMM state transitions

The non-speech detection is applied to both unsegmented and segmented data. The results of the detection on unsegmented data are used to generate signal type features for the segmentation.

3.5. Segment Rejection and Post-processing

Insertion errors produced by the recognizer can be reduced by removing non-speech data. Therefore, segments almost completely labeled as non-speech or music were removed. Furthermore, we shrunk segments which had more than 0.6s of silence at the segment end (and hence at the begin of the next segment). The silence frames were not completely removed, but 0.2s were kept. A similar procedure was used in [3].

4. EXPERIMENTAL RESULTS

Since the purpose of the segmentation is to improve the recognition performance, we measure the segmentation quality in terms of word error rate achieved on the audio data segmented. However, the computational effort for a 2-pass recognition is high, even for a relatively small development corpus. Therefore, not every manually adjusted parameter of the system (e.g. the size of sliding windows) has been analyzed yet.

4.1. Recognition System

The acoustic front end consists of MFCC features derived from a bank of 20 filters. VTLN is applied to the filterbank as described in Section 2. We use 16 cepstral coefficients (including the zeroth coefficient) which are normalized using cepstral mean and variance normalization. These MFCC features are augmented with a voicedness feature. 9 consecutive feature vectors in a sliding window are concatenated and projected to 45 components by applying an LDA.

We used 45 phonemes in across word triphone context, which are modeled by 3-state HMMs. A phonetic decision tree tied the triphone states to 4500 generalized triphone states. The pronunciation dictionary contains 58k words. The acoustic model consists of 1.1M densities with a globally pooled diagonal covariance matrix. The 4-gram language model consists of 61M multi-grams.

Table 1. Data sets used.

	training	dev	eval-1	eval-2
audio data [h]	9.5	1.0	2.6	3.0
running words	99.3K	9K	22K	31K
reference segments	3288	173	418	938
avg. ref. seg. length [s]	10.3	19.4	23.3	11.5
OOV rate [%]	0.0	0.6	0.8	0.8

Table 2. Recognition results (WER [%]) obtained on the dev set using segmentations produced with first- and second-order features.

feature order		pass	
1st	2nd	1	2
X		16.4	14.5
	X	16.3	14.3
X	X	16.3	14.2

4.2. Data Sets

All experiments were carried out on American English broadcast news data. Table 1 lists the corpus statistics.

We used parts of the Hub-4 training corpus to estimate the segmentation models. The rich transcriptions were split at speaker change, background change, overlapping speech, and at annotated sentence boundaries to produce a fine grained segmentation.

The development set used for parameter tuning and feature selection consists of two recordings from the NIST RT-03S Evaluation corpus (STT, English broadcast news). The original corpus was reduced to keep the parameter tuning feasible. The evaluation set eval-1 is the English broadcast news part from the EARS RT-04F STT development set and eval-2 is the 1998 Hub4 Evaluation corpus.

The manual segmentation used for development and evaluation sets is derived from the reference transcription provided with the corpus and thus contains mainly speaker changes. Therefore, the number of segments (relative to the recording length) in the training set is higher than in the two other sets.

4.3. Feature Evaluation

In a first step, the impact of first- and second-order features (using all segment features) is analyzed. From Table 2 can be seen, that using both first- and second-order features yields the best results. Proceeding from this result, we studied the impact of the individual segment features. The results of these experiments are shown in Figure 1. For the first pass recognition, the signal type feature is the most important one. Due to the speaker adaptation applied in the second pass, the BIC feature yields the biggest performance gain for the final result.

4.4. Recognition Results

A comparison of different segmentation approaches is shown in Table 3. We compared the error rates obtained using unsegmented data, a manual segmentation, a segmentation produced by the publicly available NIST segmenter contributed by CMU [9], and a fixed length segmentation, where each segment is 15s long. In addition, we applied a simple ASR-based method which split the recordings at silence regions longer than some threshold. This approach makes local decisions, disregarding context, properties of surrounding segments, and speaker changes. The results labeled as "MAP segmentation" were obtained using a segmentation produced with our new approach (all segment features, 1st and 2nd order features). As expected, the unsegmented and the fixed length segmentation produce

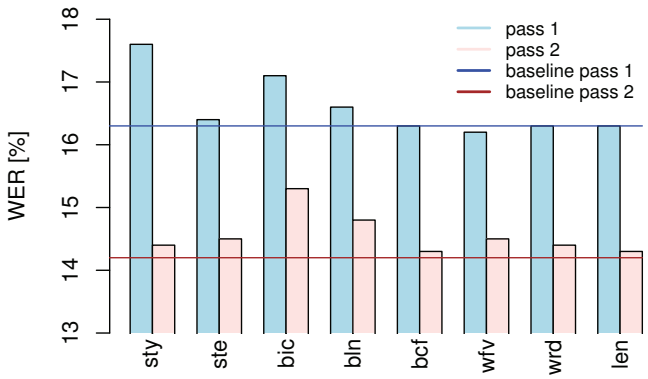

Fig. 1. Impact of discarding individual features on the recognition performance. (sty: signal type, ste: sentence end, bic: BIC score, bln: boundary length, bcf: boundary confidence, wfv: warping factor variance, wrd: words, len: segment length)

Table 3. Comparison of several segmentation methods in terms of WER [%] for pass 1 (p.1) and pass 2 (p.2).

segmentation method	dev		eval-1		eval-2	
	p.1	p.2	p.1	p.2	p.1	p.2
none	19.6	-	30.3	-	21.3	-
15s segments	19.3	17.0	29.9	27.2	22.2	20.6
NIST	16.4	14.5	28.6	25.6	19.7	17.7
ASR-based	17.3	15.3	27.3	24.8	19.5	17.7
MAP	16.3	14.2	27.3	24.9	18.9	17.0
manual	15.4	13.4	26.6	23.9	17.7	15.8

the worst results. The segmentation method introduced yields better results than the segmentation produced by the NIST segmenter. The segmentation of the dev- and the eval-2 corpus using our new approach yields better results compared to the other automatic segmentations. On the eval-1 corpus, the simple ASR-based method works slightly better, albeit the overall recognition performance is quite low on this set.

The effectiveness of the non-speech rejection is shown in Table 4. The number of deletion errors increases only slightly, while the number of insertion errors drops noticeable. The total number of errors was reduced for all corpora, using both the manual and the automatically generated segmentation.

5. CONCLUSIONS

We presented a novel MAP decoder framework for audio segmentation incorporating several segment features using log-linear models. This framework is applicable for various segment or boundary features and for different change point detection methods. Furthermore, constraints, like the maximum segment length or the maximum number of words per segment, are easy to integrate. It was shown that the segmentation has a significant impact on the quality of a subsequent automatic transcription.

The improvements achieved in recognition performance are promising, but there is still room for further improvements before reaching the quality of a manual segmentation. In principle it should be possible to generate a segmentation, which yields better recognition performance than the reference segmentation, because e.g. not all sentence boundaries are annotated. Nevertheless, we achieved a better recognition performance compared to both the acoustic- and the ASR-based segmentation.

Table 4. Effect of automatic non-speech rejection on the deletion-, insertion-, and total word error rate [%]. Results given for manual and MAP segmentation.

corpus	seg. method	w/ non-speech			w/o non-speech		
		del	ins	WER	del	ins	WER
dev	MAP	2.9	3.4	17.7	2.9	2.0	16.3
	manual	2.5	3.3	16.6	2.5	2.1	15.4
eval-1	MAP	5.3	4.5	27.9	5.3	3.9	27.3
	manual	4.9	4.4	27.0	5.0	4.1	26.6
eval-2	MAP	3.9	2.6	19.0	3.9	2.5	18.9
	manual	3.4	3.0	18.2	3.6	2.4	17.7

Future research subjects are the usage of multiple classes for different types of segment boundaries and the addition of more boundary hypotheses by e.g. acoustic change point detection. By adding boundary hypotheses at acoustic change points, e.g. speaker changes with overlapping speech could be detected. Furthermore, the appropriateness of other segment features could be analyzed.

6. ACKNOWLEDGMENTS

We thank Björn Hoffmeister for providing the English broadcast news ASR system and Pavel Golik for performing many preliminary experiments.

7. REFERENCES

[1] F. Kubala, T. Anastasakos, H. Jin, L. Nguyen, and R. Schwartz, "Transcribing radio news," in *Proc. ICSLP*, Philadelphia, PA, USA, Oct. 1996, pp. 598–601.

[2] S.S. Chen and P.S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *DARPA Broadcast News Transcription and Understanding Workshop*, Feb. 1998, pp. 127–132.

[3] S. E. Tranter, K. Yu, G. Evermann, and P. C. Woodland, "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," in *Proc. ICASSP*, Montreal, Canada, May 2004, vol. 1, pp. 753–756.

[4] D. Rybach, S. Hahn, C. Gollan, R. Schlüter, and H. Ney, "Advances in Arabic broadcast news transcription at RWTH," in *Proc. ASRU*, Kyoto, Japan, Dec. 2007, pp. 449–454.

[5] P.C Woodland, M.J.F. Gales, D. Pye, and S.J. Young, "The development of the 1996 HTK broadcast news transcription system," in *Proc. DARPA Speech Recognition Workshop*, Arden House, NY, USA, Feb. 1996.

[6] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[7] M. Cettolo, M. Vescovi, and R. Rizzi, "Evaluation of BIC-based algorithms for audio segmentation," *Computer Speech and Language*, vol. 19, no. 2, pp. 147–170, Apr. 2005.

[8] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, May 2002.

[9] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.