UNSUPERVISED AUDIO SEGMENTATION USING EXTENDED BAUM-WELCH TRANSFORMATIONS

Tara N. Sainath

MIT Computer Science and
Artificial Intelligence Laboratory
32 Vassar St. Cambridge, MA 02139
tsainath@mit.edu

Dimitri Kanevsky and Giridharan Iyengar

IBM T. J. Watson Research Center
Yorktown, NY 10598, U.S.A.
{kanevsky, giyengar}@us.ibm.com

ABSTRACT

Audio segmentation has applications in a variety of contexts, such as audio information retrieval, automatic sound analysis, and as a pre-processing step in speech recognition. Extended Baum-Welch (EBW) transformations are most commonly used as a discriminative technique for estimating parameters of Gaussian mixtures. In this paper, we derive an unsupervised audio segmentation approach using these transformations. We find that our algorithm outperforms both the Bayesian Information Criterion (BIC) and Cumulative Sum (CUSUM) segmentation methods. In particular, our EBW segmentation algorithm provides improvements over the baseline approaches in detecting landmarks of short duration and minimizing landmark oversegmentation. In addition, we show that the EBW approach provides faster computation compared to the baseline methods.

Index Terms— Acoustic signal detection, gradient methods, unsupervised learning.

1. INTRODUCTION

Many audio streams, such as television shows or radio broadcasts, contain audio from a wide variety of sources, including speech, music and laughter. Since each source varies in acoustic nature, a single method cannot be used to process the audio stream. Thus, audio segmentation has become an important pre-processing step to divide an audio stream into homogenous segments, where each segment can be handled in a different manner. Furthermore, the rapid increase in the amount of audio data in recent years has increased the need for segmentation algorithms that are computationally efficient.

Current approaches for audio segmentation include both supervised and unsupervised techniques. Supervised segmentation methods can be categorized as model-based or decoder-based. In model-based segmentation [1], Gaussian mixture models are constructed for a fixed set of acoustic classes. A maximum a-posteriori approach can be used to classify the input audio stream and segmental boundaries are delineated at changes in the audio class. In decoder-guided segmentation [2], speech and silence models are used to decode the input audio stream and the input is divided into segments at these silence locations. Model-based methods perform classification over a small number of frames in the audio stream, and are able to detect short duration segments. However, both of these methods require training models for each audio class to be used in segmentation. Thus they are limited to applications where acoustic classes are known a priori and a large amount of training data is available.

Unsupervised audio segmentation approaches are generally derived as a likelihood ratio test between two hypotheses of change and no change for a given observation sequence. Examples of this approach include both BIC [3] and CUSUM [4]. These methods do not require prior knowledge of audio classes, as models are estimated directly from the observation sequence, thus allowing them to serve a wider range of applications. However, BIC estimates models for every candidate change point within an observation sequence [3] and is computationally expensive. In addition, both methods detect changes over a large window and tend to miss many short-duration segments. Furthermore, they often suffer from over-segmentation in regions of very rapid acoustic change, such as music.

Extended Baum-Welch (EBW) transformations have been used extensively in the speech recognition community as a discriminative training technique to estimate model parameters of Gaussian mixtures. Given an initial model and input data, [5], [6] derive an explicit formula to measure the gradient steepness required to estimate a new model via the EBW transformations. This gradient measurement is an alternative to likelihood to describe how well the initial model explains the data.

In this paper, we present a novel segmentation approach using the EBW transformations. Specifically, we redefine the likelihood ratio test used in unsupervised segmentation with a measure of gradient steepness. We show that our segmentation algorithm is able to outperform both BIC and CUSUM, and specifically improves upon the short-duration missed landmark and oversegmentation problems. Finally, we demonstrate that the EBW method provides faster computation time compared to the baseline methods.

In the following sections, we provide background on the EBW transformations. Our implementation of the segmentation algorithms is described in Section 3. Section 4 presents the experiments performed, followed by a discussion of these results in Section 5. Finally, Section 6 concludes the paper and discusses future work.

2. EXTENDED BAUM-WELCH TRANSFORMATIONS

2.1. Motivation of using EBW Transformations

Given some input data, there are many different approaches used to calculate how well a model represents this data. One common approach is to calculate the likelihood, that is $p(data|model)$. Another method is to calculate the gradient, as shown in Figure 1. Given an initial model for our data and an objective function, we can estimate a new model for our data by finding the best step along the gradient of the objective function. We can think of the gradient slope as measuring how much we have to adapt an initial model to fit the data. More specifically, a steep slope indicates the initial model does not fit the data well, while a flat slope indicates the initial model is a

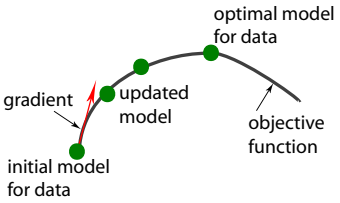

Fig. 1. EBW Model Update Graph

good fit for the data. The EBW transformations provide solutions to estimate this new model, and also provide a measure of the gradient steepness to explain the quality of the initial model to fit the data.

2.2. Derivation of EBW Transformations

The EBW procedure involves continuous transformations that can be described as follows. Assume that data $y_k^n = (y_k, ...y_n)$, from frames k to n, is drawn from a multivariate Gaussian, with each component described by the following mean and variance parameters $\epsilon_j = \{\epsilon_j, \epsilon_j\}$. Let us define the probability of frame $y_i \in y_k^n$ given model ϵ_j as:

$$z_i^j = p(y_i|\epsilon_j) = \frac{|\Delta_j|^{-1/2}}{(2\epsilon)^{n/2}} e^{-1/2(y_i-\delta_j)^T \Sigma_j^{-1}(y_i-\delta_j)} \quad (1)$$

Let $F_{k,n}(z_i^j)$ be some objective function over z_i^j and $c_i^j = z_i^j \frac{\delta}{\delta z_i^j} F_{k,n}(z_i^j)$. Given this function, the EBW transformations provide a way to estimate parameters $\hat{\epsilon}_j(C) = \{\hat{\epsilon}_j(C), \hat{\Delta}_j(C)\}$ as:

$$\hat{\epsilon}_j = \hat{\epsilon}_j(C) = \frac{\sum_{i\in I} c_{ij} y_i + C\epsilon_j}{\sum_{i\in I} c_{ij} + C} \quad (2)$$

$$\hat{\Delta}_j = \hat{\Delta}_j(C) = \frac{\sum_{i\in I} c_{ij} y_i y_i^T + C(\epsilon_j \epsilon_j^T + \Delta_j)}{\sum_{i\in I} c_{ij} + C} \geq \hat{\epsilon}_j \hat{\epsilon}_j^T \quad (3)$$

Here C is a large constant chosen such that value of the objective function increases with each iteration, that is $F_{k,n}(\hat{z}_i^j) > F_{k,n}(z_i^j)$.

Using EBW transformations (2) and (3) such that $\epsilon_j \rightarrow \hat{\epsilon}_j(C)$ and $\{z_i^j\} \rightarrow \{\hat{z}_i^j\}$, [5] derives a linearization formula between $F_{k,n}(\{\hat{z}_i^j\})$ and $F_{k,n}(\{z_i^j\})$ as:

$$F_{k,n}(\{\hat{z}_i^j\}) \geq F_{k,n}(\{z_i^j\}) = T_{k,n}^j/C + o(1/C) \quad (4)$$

Here $T_{k,n}^j$ measures the gradient required to adapt initial model ϵ_j to data y_i, or equivalently how well the data is explained by the initial model ϵ_j. A large value in T means the gradient to adapt the initial model to the data is steep and $F_{k,n}(\{\hat{z}_i^j\})$ is much larger than $F_{k,n}(\{z_i^j\})$. Thus the data is much better explained by the updated model $\hat{\epsilon}_j(C)$ compared to the initial model ϵ_j. However a small value in T indicates that the gradient is relatively flat and $F_{k,n}(\{\hat{z}_i^j\})$ is close to $F_{k,n}(\{z_i^j\})$. Therefore, the initial model ϵ_j is a good model for the data. In the next section, we derive our EBW segmentation approach using the gradient steepness value T.

3. UNSUPERVISED SEGMENTATION APPROACHES

The goal of a segmentation algorithm is to divide an audio segment into homogeneous regions. Given an observation sequence in a fixed window length, unsupervised segmentation techniques are generally derived as hypothesis testing problems. In hypothesis H_0, no change has occurred in the sequence, whereas in hypothesis H_1 a change has occurred. In this section, we describe the actual implementation of the BIC, CUSUM and EBW segmentation techniques. A typical approach for change detection can be described as follows:

1. Initialize start of sequence as $f = 0$ and end as $l = wsize$;
2. Look for a change in data from y_f to y_l
3. If change is found at point r, set $f = r$ and $l = r + wsize$, go to step 2
4. If no change is found, $l = l + winres$
5. If $wsize > wmax$, $f = f + winres$ and $l = r + wsize$, go to step 2

Below we discuss the specific details of the BIC, CUSUM and EBW segmentation methods.

3.1. Bayesian Information Criterion

BIC is a model selection problem which can be formulated as a log-likelihood ratio between two hypotheses representing a change or no change in a given observation sequence. It is penalized by a model complexity term denoting the difference in number of model parameters in each hypothesis. Given observations $(y_1, \ldots, y_n)$ drawn from a multi-dimensional Gaussian process, the two hypotheses for a possible change at k can be formulated as follows:

$$\begin{cases} H_0, & y_1, \ldots, y_n \geq N(\epsilon_0, \Delta_0) \\ H_1, & y_1, \ldots, y_k \geq N(\epsilon_1, \Delta_1), y_{k+1}, \ldots, y_n \geq N(\epsilon_2, \Delta_2) \end{cases}$$

The likelihood ratio can be reduced to the following decision rule for a change detection at point k [3]:

$$\begin{aligned} \Delta BIC_k &= \frac{1}{2}(N \geq k) \log|\Delta_2| + k \log|\Delta_1| \geq N \log|\Delta_0|) \\ &\geq \frac{1}{2}\epsilon\,(d + \frac{1}{2}d(d+1)) \log N > 0 \end{aligned} \quad (5)$$

Here ϵ is a penalty factor which weights the model complexity term, and d is the dimension of data y_i. This test is applied to all possible change points k within the observation and we choose the change point which has the most positive $\Delta BIC_k > 0$ value. The model parameters under each test depend on the location of change point k and are re-estimated via maximum likelihood for each new hypothesized boundary within the observation sequence.

3.2. Cumulative Sum

CUSUM is another segmentation algorithm used in a variety of change detection problems [4]. Under the assumption that each y_i is drawn from an independent, identically distributed process, the CUSUM test has been shown to be optimal in minimizing the detection time for a given false alarm rate [7]. Given observations $(y_1, \ldots, y_n)$, the CUSUM test for a change at point k is as follows:

$$\begin{cases} H_0, & y_1, \ldots, y_n \geq N(\epsilon_0, \Delta_0) \\ H_1, & y_1, \ldots, y_k \geq N(\epsilon_0, \Delta_0)\,, y_{k+1}, \ldots, y_n \geq N(\epsilon_1, \Delta_1) \end{cases}$$

Here parameters $\epsilon_0 = \{\epsilon_0, \Delta_0\}$ are estimated from a few samples in the beginning of the observation sequence, and $\epsilon_1 = \{\epsilon_1, \Delta_1\}$ from the end of the observation sequence [4]. Thus the models are only estimated once for a given observation sequence and are independent of the change point. If we define the log-likelihood as $\log p(y_1^k|\epsilon_1) = \sum_{i=1}^k \log p(y_i|\epsilon_1)$, then the decision rule for the

best change point within the observation sequence is derived as a log-likelihood ratio between the two above hypotheses as follows:

$$\hat{k} = \arg\max_{k} \{\log p(y_{k+1}^{n}|\epsilon_1) + \log p(y_1^{k}|\epsilon_0) \geq \log p(y_1^{n}|\epsilon_0)\} \geq \epsilon \quad (6)$$

The likelihood ratio is compared to an empirically determined threshold ϵ to hypothesize a change point.

3.3. Segmentation via EBW Transformations

While CUSUM estimates models once in a data sequence and is faster than BIC, estimated models can sometimes be poor when inferred from few samples. Since the gradient steepness value T takes into account model re-estimation using the entire data sequence, it corrects for initial estimation error [8]. Thus our motivation for using EBW for segmentation is to provide benefits in segmentation performance and computational efficiency over the baseline methods.

First let us define the objective function $F_{k,n}$ as the log-likelihood over the data y_k^n given model ϵ_j as $F_{k,n}(z_i^j) = \sum_{i=k}^{n} \log p(y_i|\epsilon_j)$. Using Equation 4 and the above objective function, [6] derives a closed-form expression for T. Given initial model parameters $\epsilon_j = \{\epsilon_j, \epsilon_j\}$ which are estimated similar to CUSUM and data y_k^n , the gradient steepness measure $T_{k,n}^{j}$ is written as:

$$T_{k,n}^{j} = \frac{1}{2} \sum_{r,l \in \{1,\ldots,d\}, r \neq l} \left(\frac{1}{\epsilon_{jr}^{4}} + \frac{1}{\epsilon_{jl}^{4}}\right) \left[\sum_{i=k}^{n} (y_{ir} \geq \epsilon_{jr})(y_{il} \geq \epsilon_{jl})\right]^2 + \sum_{r=1}^{d} \frac{1}{\epsilon_{jr}^{2}} \left\{ \frac{\{\sum_{i=k}^{n} [(y_{ir} \geq \epsilon_{jr})^2 \geq \epsilon_{jr}^{2}]\}^2}{2\epsilon_{jr}^{2}} + [\sum_{i=k}^{n} (y_{ir} \geq \epsilon_{jr})]^2 \right\} \quad (7)$$

With this formula for T, our segmentation criterion becomes:

$$\hat{k} = \arg\min_{k} \{T(y_{k+1}^{n}|\epsilon_1) + T(y_1^{k}|\epsilon_0) \geq T(y_1^{n}|\epsilon_0)\} \geq \epsilon \quad (8)$$

Intuitively this means that we detect a change at point $\hat{k}$ if $(y_1, \ldots, y_k)$ is explained better by ϵ_0 (i.e. $T(y_1^k|\epsilon_0)$) and $(y_{k+1}, \ldots, y_n)$ is explained better by ϵ_1 (i.e. $T(y_{k+1}^n|\epsilon_1)$), rather than the entire sequence being explained by ϵ_0 (i.e. $T(y_1^n|\epsilon_0)$). The better the data is explained by a specific set of models, the smaller T is, so we look for the change point k which produces the minimum T and again compare this to an empirically found threshold ϵ.

To improve upon missed landmark and oversegmentation problems inherent in unsupervised segmentation, we describe below added refinement and merge stages using the EBW transformations.

3.3.1. Refinement Stage

Unsupervised segmentation methods tend to miss short duration segments since they look for changes over a large window length. To alleviate this problem, we introduce a boundary refinement stage. First, approximate boundaries are found via the EBW algorithm described above. Let us denote these boundaries as $\mathbf{B} = (b_1, \ldots, b_n)$. Between two neighboring boundaries b_{i-1} and b_i, we estimate a model M_i from the data in this segment. If every observation in the segment belongs to the same model, then if we compute the gradient steepnesss measure T from a subset of data given the model M_i, we would expect it to remain small. However, a large value in T indicates that the data is not well explained by M_i. Thus during refinement, for each segment we compute T for subsets of data in the segment given model M_i. If T exceeds an empirically found threshold in a given subset, we hypothesize a new boundary here.

3.3.2. Merge Stage

Furthermore, we introduce a merge stage to improve upon over-hypothesizing boundaries in regions of fast acoustic change. Given a set of segments from the first two stages, $\mathbf{S} = (s_1, \ldots, s_k)$, we now check how similar data in two segments are. We estimate a model M_i for data in s_i and compute a T value for data in s_{i+1} using model M_i. This is a measure of how well M_i explains data in s_{i+1}. If the T value is below an empirically found threshold, the data in the two segments are similar so we merge the two segments together.

4. EXPERIMENTS

4.1. Corpus

We perform segmentation experiments using the Computers in the Human Interaction Loop (CHIL) Isolated Acoustic Event data set. This database has been collected by the University Polytechnic of Catalonia (UPC) for their Acoustic Event Detection and Classification tasks [9]. Our motivation for using CHIL stems from a future goal of using EBW for a unified segmentation and classification system. The acoustic variety in CHIL provides a good corpus for testing these ideas. The set is divided into 3 sessions, with 10 participants per session. Sounds are recorded in a closed room using 16 different microphone types. At each session, each participant takes a different place in the room and records isolated acoustic events from 15 different classes, including knocks, doors opening/closing, applause, laughter, etc. In total, there are over 6000 change points per session, which are annotated manually by UPC. In our experiments, the data is sampled at 16kHz, and then windowed to 20ms frames with a 10ms overlap. 19 dimension MFCCs are calculated for each frame.

4.2. Evaluation Metrics

Two common errors can occur in segmentation algorithms. Type-I errors occur if a true boundary is not detected. Here we define a true boundary as detected if it is within 1 second of the hypothesized boundary. Type-II errors occur if a hypothesized boundary does not correspond to a true boundary. Type I and II errors can be measured by precision and recall respectively, defined as:

$$\text{Precision} = \frac{\text{\# detections}}{\text{total \# true bndries}}, \quad \text{Recall} = \frac{\text{\# detections}}{\text{\# hyp. bndries}}$$

Segmentation can also be measured by the F-measure as:

$$\text{F-measure} = \frac{2 \geq \text{Precision} \geq \text{Recall}}{\text{Precision} + \text{Recall}}$$

For our performance experiments, we train on the first 40 waveforms ($\geq$ 2 hours) from session 1. Our thresholds for each algorithm are chosen to minimize the F-measure. We then test on 100 recordings ($\geq$ 5 hours) from each session. In addition, we compute the execution time of the three algorithms, defined as the time the CPU spends executing the segmentation code. We compute this total time for 5.5 hours of data in session 1, and average the results over 5 trials.

5. RESULTS

Table 1 shows the evaluation metric scores for the CUSUM, BIC and EBW algorithms, the latter with and without the refinement and merge stages. Note that the BIC implementation discussed here also includes a similar refinement and merge stage [10]. The EBW algorithm outperforms CUSUM for all sessions. Each term in CUSUM

CHIL Data Segmentation Results			
	Session 1		
	Precision	Recall	F-measure
EBWSeg, ref. & merge	0.87	0.83	0.85
EBWSeg	0.84	0.80	0.82
BIC	0.84	0.79	0.81
CUSUM	0.76	0.76	0.76
	Session 2		
EBWSeg, ref. & merge	0.88	0.82	0.85
EBWSeg	0.86	0.80	0.83
BIC	0.85	0.81	0.83
CUSUM	0.78	0.76	0.77
	Session 3		
EBWSeg, ref. & merge	0.88	0.83	0.85
EBWSeg	0.86	0.82	0.83
BIC	0.87	0.81	0.83
CUSUM	0.79	0.77	0.78

Table 1. Statistics for EBW, CUSUM and BIC Segmentation

calculates a likelihood of the data when models are estimated from a subset of the data. However, each term T in EBWSeg captures the difference between the likelihood of a data given the initial model and the likelihood with a model estimated from the entire data sequence. CUSUM does not always provide the best estimate of data with the initial model, so our model re-estimation via EBW using the entire sequence is able to correct for this initial model error [8].

The performance of BIC and EBW are comparable, even though BIC obtains a better estimate of model parameters at each hypothesized boundary compared to EBW. One explanation is that the EBW objective more closely matches the goal of segmentation. BIC is a model selection problem, where the objective is to chose the best set of models to explain data while limiting model complexity [3]. However, the objective of a CUSUM and EBW-hypothesized test is to minimize detection time for a given false alarm rate [7].

Lastly, adding refinement and merge, EBW offers on average a 2% absolute improvement over BIC and 7% over CUSUM, and a t-test verifies these results to be statistically significant. Figure 2a shows a histogram of segment duration of missed boundaries for each method. EBW has a 10.2% relative improvement in missed boundaries over BIC and 17.2% over CUSUM. In particular, the figure shows that EBW offers significant improvement for small segments. Figure 2b shows the normalized frequency of oversegmented boundaries for each method. EBW offers a 19.7% relative improvement in oversegmentation over BIC and 34.6% over CUSUM.

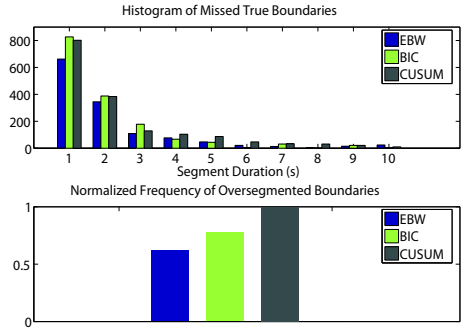

Fig. 2. Histograms of missed and oversegmented boundaries

5.1. Computation Time

Table 2 shows the average total execution time on 5.5 hours of data for the three methods. EBW performs more than 5 times faster than BIC and 3 times faster than CUSUM. At each new hypothesized boundary within an observation sequence, BIC re-estimates model parameters. However, EBW and CUSUM only estimate model parameters once within an observation sequence. Yet, CUSUM computes the inverse covariance and determinant in the likelihood formulation, and is thus slower than EBW which computes variances.

	Average Total Execution Time (hrs)
EBWSeg, ref. & merge	0.42
BIC	2.25
CUSUM	1.43

Table 2. Average Total Execution Time of Segmentation Algorithms

6. CONCLUSIONS AND FUTURE WORK

In this paper, we explored a specific gradient steepness measure derived from the EBW Transformations. We presented a novel segmentation approach using this gradient measurement and found that our segmentation method was able outperform both BIC and CUSUM and provide faster computation time. The EBW transformations appear to be a general technique to explain the quality of a model used to represent the data. We would like to explore using EBW in a variety of other contexts, including for other unsupervised segmentation tasks, clustering and classification. In addition, we would like to explore other approaches to measuring gradient steepness.

7. REFERENCES

[1] M. Spina and V. Zue, "Automatic Transcription of General Audio Data: Preliminary Analyses," in *Proc. ICSLP*, 1996.

[2] B. Ramabhadran, J. Huang, U. Chaudhari, G. Iyengar, and H. J. Nock, "Impact of Audio Segmentation and Segment Clustering on Automated Transcription Accuracy of Large Spoken Archives," in *Proc. EuroSpeech*, 2003, pp. 2589 – 2593.

[3] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via The Bayesian Information Criterion," in *Proc. Broadcast News Trans. and Under. Workshop*, February 1998, pp. 127–132.

[4] M. Omar, U. Chauduri, and G. Ramaswamy, "Blind Change Detection for Audio Segmentation," in *Proc. ICASSP*, 2005.

[5] D. Kanevsky, "Extended Baum Transformations for General Functions," in *Proc. ICASSP*, 2004.

[6] D. Kanevsky, "Extended Baum Transformations For General Functions, II," Tech. Rep. RC23645(W0506-120), Human Language Technologies, IBM, 2005.

[7] M. Basseville and I. Nikiforov, *Detection of Abrupt Changes: Theory and Application*, Prentice-Hall, 1993.

[8] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, Wiley Interscience, 2nd edition, 2001.

[9] A. Waibel et al., "CHIL: Computers in the Human Interaction Loop," in *WIAMIS*, 2004.

[10] A. Tritschler, "A Segmentation-enabled Speech Recognition Application using the BIC criterion," M.S. thesis, Institut EURECOM, France, 1998.