AN ONLINE EM ALGORITHM IN HIDDEN (SEMI-)MARKOV MODELS FOR AUDIO SEGMENTATION AND CLUSTERING

Alberto Bietti[⋆] *Francis Bach*[†] *Arshia Cont*[⋆]

[⋆] MuTant Project-team, Ircam, Inria, UPMC & CNRS
[†] Sierra Project-team, Inria, ENS & CNRS

ABSTRACT

Audio segmentation is an essential problem in many audio signal processing tasks, which tries to segment an audio signal into homogeneous chunks. Rather than separately finding change points and computing similarities between segments, we focus on joint segmentation and clustering, using the framework of hidden Markov and semi-Markov models. We introduce a new incremental EM algorithm for hidden Markov models (HMMs) and show that it compares favorably to existing online EM algorithms for HMMs. We present results for real-time segmentation of musical notes and acoustic scenes.

Index Terms**—** Hidden Markov models, online learning, audio segmentation, EM algorithm

1. INTRODUCTION

The task of audio segmentation aims to discover regions in the audio stream that present steady statistical properties over time. Audio segmentation is a key front-end to many applications such as audio surveillance systems [1], computational auditory scene analysis [2], and music information retrieval systems [3] such as automatic indexing and music summarization [4]. Despite this relevance, most existing approaches focus on specific acoustic features [5, 4] thus reducing application to wider audio, or focus on variants of change detection [6, 7], making detections intractable. Requirements for such front-ends often boils down to detecting significant changes in the signal over time (segmentation) and associating similar segments in the signal (clustering). These two subtasks are usually undertaken separately [4, 6, 7, 5] making the system immune to error propagation between each stage.

Our motivation in this paper is to provide *online* algorithms that perform segmentation and clustering in one pass. Rather than separately detecting changes and finding similarities, we perform online unsupervised joint segmentation and clustering. Our motivation in providing online algorithms is to enable real-time applications as well as scalability of such systems to very large databases and signals (such as music). A natural modelling framework for our task is that of hidden Markov and semi-Markov models [8, 9, 10, 11]. Cappé [12] proposed an online EM algorithm for HMMs with finite state space based on a forward smoothing recursion, but this recursion is quite expensive computationally. We develop a new online algorithm for learning parameters in HMMs based on incremental optimization of lower bounds on the log-likelihood, thus extending the ideas of the incremental EM algorithm of Neal and Hinton [13] or other online algorithms for independent observations (e.g., [14]) to Markovian observations. We then apply our algorithm to real-time audio segmentation tasks, both on musical notes and on acoustic scenes, using examples from the Office Live Dataset [15].

2. RELATED WORK

While online learning algorithms have been studied extensively in the context of independent observations, including the case of multiple observed sequences of an HMM, little work has been done on online learning of HMM parameters from a single observation sequence, where the model is updated after each new observation. Some approaches consider small blocks of the sequence and relies on algorithms for independent observations, others incrementally approximate the full posterior. See [16] for a survey.

Perhaps the approach most closely related to our work is the online EM algorithm for HMMs of Cappé [12]. The algorithm uses a stochastic approximation procedure in the space of sufficient statistics in order to try to reach a *limiting EM* recursion (corresponding to a batch EM algorithm with infinite data). The updates are inspired by a forward-only smoothing recursion, in which the expected sufficient statistics needed for parameter updates are computed recursively. Our algorithm relies instead on an incremental minorization-maximization algorithm by updating and improving lower bounds on the likelihood after each observation. The bounds are based on an auxiliary distribution on the hidden chain, which can be updated incrementally, and under which expected sufficient statistics can be computed exactly and updated incrementally. We obtain a complexity per observation of $O(K^2 + Kp)$ compared to $O(K^4 + K^3p)$ by Cappé [12], where K is the number of states and p the dimensionality

Research project partly funded by ANR-11-JS03-005-01.

of observations, and the empirical prediction results are just as good, if not better. The runtime on an entire sequence is similar to a single batch EM iteration, but the incremental updates lead to faster improvements in the parameters, and are suitable for a real-time or streaming setting where one cannot access the entire sequence.

Existing frameworks for audio segmentation can be studied within two categories: those engineered for specific application domains such as music timbre similarity [10], western music tonality [4, 11], music onset [17] and more; and blind systems focusing on change detection such [6, 7]. Most existing systems are offline (requiring the entire sequence), and online methods are confined to the single task of change detection (e.g., [17, 6]).

3. HIDDEN MARKOV MODELS FOR AUDIO SEGMENTATION AND CLUSTERING

Hidden Markov models (HMMs) are a powerful tool for modeling sequential data, and have been very successful in audio and speech applications [8, 9]. We use them for joint segmentation and clustering of an audio signal by letting the hidden chain $z_{1:T} = (z_1, \dots, z_T) \in \{1, \dots, K\}^T$ represent the sequence of cluster identities of the segments. The distribution of z_1 will be denoted by $\pi \in \mathbb{R}^K$ ($\pi_i = p(z_1 = i)$), and the (homogeneous) transition matrix by $A \in \mathbb{R}^{K \times K}$ ($A_{ij} = p(z_t = j | z_{t-1} = i)$).

The audio signal representation is given by a sequence $x_{1:T} = (x_1, \dots, x_T)$ of observations, where each $x_t \in \mathbb{R}^p$ is a normalized magnitude short-time Fourier spectrum given by $x_t = |\hat{x}_t| / \|\hat{x}_t\|_1$, with $\hat{x}_t$ the FFT coefficients of frame t. The normalization is used to have some invariance to volume changes. In the HMM, x_t is independent from the other observations given z_t, and we assume the emission distribution $p(x_t | z_t = i)$ in state i to be parameterized by its mean vector $\mu_i \in \mathbb{R}^p$.

3.1. Emission distributions

Because our representation is normalized, a natural choice for emission distributions is the multinomial. This corresponds to having different probabilities associated with every frequency bin. Although the multinomial is discrete, we can approximate its mass function with our non-integer representation by normalizing our vectors x_t to a large enough $N = \sum_j x_{t,j}$, which corresponds to the number of trials in the multinomial, given that the combinatorial statistic $h(x) = N!/x_1! \dots x_p!$ will cancel out in the parameter updates. Note that this choice leads to a mass function of the form $p(x; \mu) = h(x) \exp(-D_{KL}(x || \mu))$, where $D_{KL}(x||y) = \sum_i x_i \log x_i / y_i$ is the KL divergence [18], which has been shown to empirically perform better than the Euclidian distance for audio signals [19], thus further justifying our choice.

More generally, we will consider *regular* exponential families, for which Banerjee et al. [18] show that there is a bijection with a class of Bregman divergences, including squared Euclidian distance, KL and Itakura-Saito divergences. Bregman divergences are defined by $D_\psi(x, y) = \psi(x) - \psi(y) - \langle x - y, \nabla \psi(y) \rangle$, where ψ is a strictly convex function, and under some assumptions [18], we have

$$p_\mu(x) = h(x) \exp(\langle x, \theta \rangle - \psi(\theta)) = h_1(x) \exp(-D_{\psi^*}(x, \mu)),$$

with $h_1(x) = h(x) e^{\psi(x)}$, where ψ^* is the Fenchel conjugate of ψ, μ is the mean parameter, $\theta = \nabla \psi(\mu)$ is the natural parameter, and x the (minimal) sufficient statistic. We use the mean parameterization, for which the maximum likelihood estimate is given by the empirical expectation (by moment-matching). Here, we use emissions of the form $p(x|z = i) = p_{\mu_i}(x)$, although it is easy to generalize our work to different emissions, such as general Gaussian distributions, as in [12], or even mixtures of Gaussians if we add mixture components to the hidden chain in a similar way to Section 4.4.

3.2. Explicit duration distributions

In an HMM, the length distribution of a segment in state i is implicitly geometric, given by $p_i(d) = A_{ii}^{d-1}(1 - A_{ii})$. In audio segmentation, we might want to enforce different duration distributions for our segments, e.g., to avoid very short segments, or to encourage having segments of a specified length. This can be done by explicitly modeling duration distributions of each state with a *hidden semi-Markov model* (HSMM), also known as *explicit duration HMM* in this specific case [8, 20]. Common examples of explicit duration distributions p_i are Poisson or negative binomial distributions. The transition matrix A then models transition across segments, and in each segment in state i, a segment length d is sampled from p_i, and d observations are sampled i.i.d. from p_{μ_i}.

4. ONLINE PARAMETER ESTIMATION

4.1. EM algorithm

The EM algorithm is probably the most standard algorithm used for maximum likelihood estimation in latent variable models, especially in HMMs. One way to see it is that it successively maximizes lower bounds on the log-likelihood, each given by a particular auxiliary distribution on the hidden variables. If $\mathbf{x} = x_{1:T}$ are the observed variables, $\mathbf{z} = z_{1:T}$ the hidden variables and θ the parameters, then using Jensen's inequality, for any probability distribution q on the hidden variables, we have (see, e.g., [13]):

$$\log p(\mathbf{x}; \theta) = \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} \geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})}.$$

The E-step takes $q(\mathbf{z}) = p(\mathbf{z}|\mathbf{x}; \theta)$, which makes the bound tight by maximizing it with respect to the distribution q. The

M-step then maximizes this new lower bound w.r.t. θ, which is equivalent to maximizing $\mathbb{E}_{\mathbf{z}\sim q}[\log p(\mathbf{x}, \mathbf{z}; \theta)]$. In HMMs, the E-step is carried by computing posterior marginals on each state and pair of states in the hidden sequence using, e.g., a forward-backward algorithm, and these quantities are then used in the M-step to update parameters.

In the case of independent observations, we can use a factored distribution $q(\mathbf{z}) = \prod_t q_t(z_t)$ in the lower bound since the maximizer takes this form. The incremental EM algorithm of Neal and Hinton [13] then performs partial E-steps by maximizing the lower bound w.r.t. a single q_t.

4.2. Incremental EM for HMMs

In order to derive an incremental EM algorithm for HMMs, we would like to update the distribution q incrementally with every new observation. Using the chain rule and the conditional independencies of HMMs, the maximizing distribution factorizes as $q(z_{1:T}) = p(z_1|x_{1:T}) \prod_{t\geq 2} p(z_t|z_{t-1}, x_{1:T})$. Thus, we can restrict ourselves to distributions q of the form $q(z_{1:T}) = q_1(z_1) \prod_{t\geq 2} q_t(z_t|z_{t-1})$, with $\sum_j q_t(j|i) = 1$ for all t and i. For this choice of q, the marginals on z_t are given by $\phi_t(z_t) = \sum_{z_{1:t-1}} q_1(z_1) \dots q_t(z_t|z_{t-1})$, which can be computed incrementally with $\phi_1(i) = q_1(i)$ and $\phi_t(j) = \sum_i \phi_{t-1}(i) q_t(j|i)$ for $t \geq 2$. The pairwise marginals then take the form $q(z_{t-1}, z_t) = \phi_{t-1}(z_{t-1}) q_t(z_t|z_{t-1})$. We obtain the following lower bound on the log-likelihood:

$$\hat{f}_T(\theta) = \mathbb{E}\left[\log \frac{p_\theta(x_{1:T}, z_{1:T})}{q(z_{1:T})}\right] = \sum_{t=1}^T \mathbb{E}\left[\log \frac{p_\theta(x_t, z_t|z_{t-1})}{q_t(z_t|z_{t-1})}\right]$$
$$= \sum_{t=1}^T \sum_{z_{t-1}, z_t} \phi_{t-1}(z_{t-1}) q_t(z_t|z_{t-1}) \log \frac{p_\theta(x_t, z_t|z_{t-1})}{q_t(z_t|z_{t-1})},$$

with expectations taken w.r.t. the distribution $q(z_{1:T})$. Hence, we can perform a partial E-step on a new observation x_T by maximizing the last term in $\hat{f}_T$ w.r.t. q_t, leading to $q_t(j|i) \propto A_{ij} p(x_T|z_T = j; \theta)$, and updating ϕ_t accordingly. The M-step then maximizes $\hat{f}_T$ w.r.t. θ.

Existing approaches apply the independent incremental EM algorithm [13] to blocks of the sequence that are considered independent (e.g., [21]), which does not provide an incremental model for transitions unlike our method. Other approaches try to approximate expected sums of sufficient statistics under the posterior (e.g., [22]), while our algorithm involves exact expectations under the custom distribution q.

4.3. Example with Bregman divergence emissions

We now consider the HMM with emission distributions described in Section 3.1. The M-step requires the following expected sums of sufficient statistics in order to estimate the new parameters $A^{(T)}$ and $\mu_i^{(T)}$[1]: $S_T^A(i,j) = \sum_{t=2}^T \phi_{t-1}(i) q_t(j|i)$, $S_T^{\mu,0}(i) = \sum_{t=1}^T \phi_t(i)$, and $S_T^{\mu,1}(i) = \sum_{t=1}^T \phi_t(i) x_t$, which can trivially be updated incrementally after each observation. The parameter updates are then:

$$A_{ij}^{(T)} = \frac{S_T^A(i,j)}{\sum_{j'} S_T^A(i,j')} \quad \mu_i^{(T)} = \frac{S_T^{\mu,1}(i)}{S_T^{\mu,0}(i)}.$$

The time complexity of each online iteration is thus $O(K^2 + Kp)$ (where p is the dimensionality of the observations), much less than the online EM iterations of Cappé [12] which cost $O(K^4 + K^3 p)$, but get some theoretical guarantees.

4.4. Semi-Markov extension

In order to extend the algorithm to HSMMs, we can follow [23] and parameterize the HSMM as an HMM with two hidden variables, namely the state of the current segment, z_t, and a counter of the time steps passed since the beginning of the segment, z_t^D. The transitions are then given by

$$p(z_t = j | z_{t-1} = i, z_t^D = d) = \begin{cases} A_{ij}, & \text{if } d = 1 \\ \delta(i,j), & \text{otherwise} \end{cases}$$
$$p(z_t^D = d' | z_{t-1} = i, z_{t-1}^D = d) = \begin{cases} \lambda_i(d), & \text{if } d' = d+1 \\ 1 - \lambda_i(d), & \text{if } d' = 1 \\ 0, & \text{otherwise.} \end{cases}$$

If we take $\lambda_i(d) = D_i(d+1)/D_i(d)$, where $D_i(d) := \sum_{d' \geq d} p_i(d')$, then the prior probability of having a segment of length at least d is equal to $\lambda_i(1) \dots \lambda_i(d-1) = D_i(d)$, and that of having a segment of length exactly d is $\lambda_i(1) \dots \lambda_i(d-1)(1 - \lambda_i(d)) = p_i(d)$, which matches the HSMM (for any choice of duration distribution).

The incremental EM algorithm in this model uses quantities of the form $q_t(z_t, z_t^D | z_{t-1}, z_{t-1})$ and $\phi_t(z_t, z_t^D)$. Thanks to the deterministic transitions, we get $q_t(j, d'|i, d) = 0$ if $d' \notin \{1, d+1\}$, or $d' = d+1$ and $i \neq j$, and we obtain a time complexity per observation of $O(K^2 D + KDp)$, where D is the maximal duration of a segment, versus $O(K^4 D + K^3 Dp)$ if we adapt the algorithm of Cappé [12] to this model.

5. AUDIO SEGMENTATION EXPERIMENTS

We applied our algorithm to a musical note segmentation task and an acoustic scene segmentation task. We use the short-time spectral representation described in Section 3, with a sampling rate of 44.1KHz, computed using Hamming windows of size 4096 with an offset size of 512 samples, limited to $p = 1024$ frequency components. As explained in Section 3.1, our emissions are multinomials, initialized to uniform multinomials with a small added noise term to break symmetry. As in [12], we wait some time ($t = 100$) before performing M-steps in the online algorithms.

[1] Note that we do not attempt to estimate the initial distribution π online, as justified in [12].

	HMM	HSMM	HMM long sequence
Batch EM (5 iter.)	0.60	1.06	18
Cappé [12]	1.52	108.6	325 (231)
Incremental	0.51	0.91	13.5 (10.5)

Table 1: Running time comparison (in seconds). The time in parenthesis is obtained with an M-step every 10th iteration.

5.1. Musical note segmentation

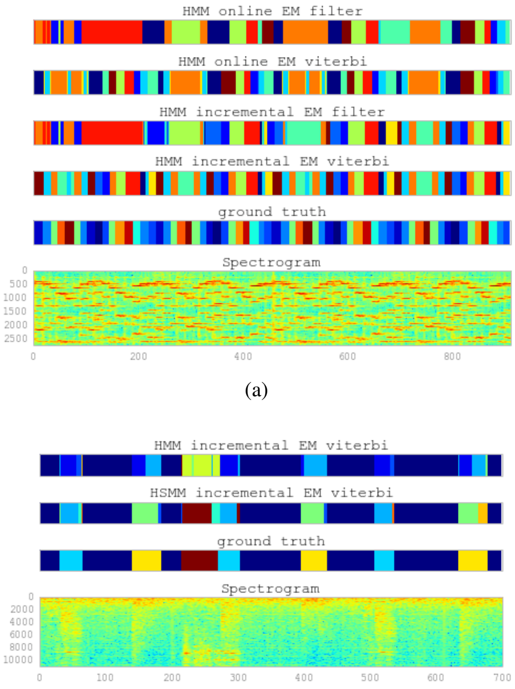
We compared the results of our incremental EM algorithm to the online EM approach of Cappé [12] on a short music sequence from J.S. Bach's second violin sonata, which is repeated twice (6s total) in order to see the improvements of the online algorithms after "hearing" it twice. Figure 1a shows the filtering estimates obtained in real-time as well as the Viterbi sequences with final parameters for both algorithms, along with the ground truth note sequence and spectrogram. We can see how the real-time segmentation improves over time, and especially when the sequence is repeated. Our incremental EM algorithm seems to find a more granular segmentation with more notes. Table 1 (first two columns) shows a comparison of running times on the same sequence for batch EM (5 iterations, enough to converge) and the two online algorithms. Duration distributions are $NB(5, 0.2)$ (mean 20) truncated to $D = 100$. We see that online EM [12] is slower than our incremental EM, and prohibitively so for the HSMM.

On a specific test run using a recording of the 3rd movement from Beethoven's first piano sonata (duration of approximately 150 seconds), we compared the running time of batch EM and incremental EM ($K = 40$, $T = 6500$), shown in the last column of Table 1. Performing an M-step every 10 iterations, incremental EM runs in 10.5s, while it takes 5 iterations for batch EM to surpass the final likelihood, for a total of 18s, which is significantly slower. In comparison, online EM [12] runs in 231s, and the final likelihood value is lower than the value obtained with incremental EM.

5.2. Acoustic scene segmentation

We applied our segmentation algorithms to auditory scenes, using examples from the Office Live Dataset [15]. The goal is to detect acoustic events such as a coughing sound, a door slam, or the sound of keys being dropped on a table. Because most of these sounds are not homogeneous, we do not expect each event to be represented by a single segment, but rather by a sequence of smaller segments, which is similar for different instances of the event. The events can then easily be detected from the sequence of segments with a higher-order algorithm.

In Figure 1b, we used a sound sequence which alternates between keys dropping and door slam sounds. The audio content is quite different across examples, as we can see in the spectrogram, and in particular the second keys drop is preceded by a sound of shaking keys. Nonetheless, our online algorithms result in similar segment sequences for the different instances, and the shaking keys sound has its own separate cluster. The HSMM duration distributions are $NB(5, 0.2)$, and we can see how this encourages longer segments and avoids the very short segments obtained with the HMM.



(a)

(b)

Fig. 1: (a) Online vs incremental EM for HMMs on a musical note segmentation task ($K = 10$, $T = 910$). (b) Scene segmentation results: keys dropping and door slam. Background states hidden for clarity. Best seen on screen.

6. CONCLUSION AND DISCUSSION

In this paper, we proposed a framework for online joint segmentation and clustering of audio signals. We employed hidden Markov and semi-Markov models and proposed an incremental EM algorithm for online parameter learning, with a complexity comparable to a single batch EM iteration. A convergence analysis of the algorithm is left for future work.

Our main motivation is to enable segmentation and clustering on audio signals in real-time and solely based on their unfolding statistical properties with no specific content-based analysis as observed in most front-end applications. In Section 5, we provided preliminary results on realistic music and auditory scene signals. Future research will attempt to further enhance the proposed framework for automatic structure discovery, audio surveillance systems, and music indexing.

7. REFERENCES

[1] R. Radhakrishnan and A Divakaran, "Generative process tracking for audio analysis," in *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2006, vol. 5.

[2] Dan P. W. Ellis, *Prediction-driven Computational Auditory Scene Analysis*, Ph.D. thesis, Massachusetts Institute of Technology, MA, USA, June 1996.

[3] Meinard Müller, *Information retrieval for music and motion*, Springer, 2007.

[4] Wei Chai, "Semantic segmentation and summarization of music: methods based on tonality and recurrent structure," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 124–132, 2006.

[5] George Tzanetakis and Perry Cook, "Multifeature audio segmentation for browsing and annotation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 1999, pp. 103–106.

[6] Arnaud Dessein and Arshia Cont, "An information-geometric approach to real-time audio segmentation," *Signal Processing Letters, IEEE*, vol. 20, no. 4, pp. 331–334, 2013.

[7] Jonathan Foote, "Automatic audio segmentation using a measure of audio novelty.," in *IEEE International Conference on Multimedia and Expo (I)*, 2000.

[8] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[9] Olivier Cappé, Eric Moulines, and Tobias Ryden, *Inference in Hidden Markov Models (Springer Series in Statistics)*, Springer-Verlag New York, 2005.

[10] Jean-julien Aucouturier and Mark Sandler, "Segmentation of musical signals using hidden Markov models," in *In Proc. 110th Convention of the Audio Engineering Society*, 2001.

[11] Mark Levy and Mark Sandler, "New methods in structural segmentation of musical audio," in *In Proc. EUSIPCO*, 2006.

[12] Olivier Cappé, "Online EM algorithm for hidden Markov models," *Journal of Computational and Graphical Statistics*, vol. 20, no. 3, pp. 728–749, Jan. 2011.

[13] Radford Neal and Geoffrey E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*. 1998, pp. 355–368, Kluwer Academic Publishers.

[14] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, pp. 19–60, Mar. 2010.

[15] Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange, and Mark D. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[16] Wael Khreich, Eric Granger, Ali Miri, and Robert Sabourin, "A survey of techniques for incremental learning of HMM parameters," *Information Sciences*, vol. 197, pp. 105–130, 2012.

[17] P. Brossier, J. P. Bello, and M. D. Plumbley, "Real-time temporal segmentation of note objects in music signals," in *Proceedings of the International Computer Music Conference (ICMC2004)*, Nov. 2004.

[18] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, Dec. 2005.

[19] Y. Stylianou and AK. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, vol. 2, pp. 837–840.

[20] Yann Guédon, "Estimating hidden semi-Markov chains from discrete sequences," *Journal of Computational and Graphical Statistics*, vol. 12, no. 3, pp. 604–639, 2003.

[21] Yoshihiko Gotoh, Michael M. Hochberg, and Harvey F. Silverman, "Efficient training algorithms for HMMs using incremental estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 6, pp. 539–548, 1998.

[22] V. Krishnamurthy and J.B. Moore, "On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure," *IEEE Transactions on Signal Processing*, vol. 41, no. 8, pp. 2557–2573, Aug. 1993.

[23] S. Yildirim, S. S. Singh, and A. Doucet, "An online expectation-maximization algorithm for changepoint models," *Journal of Computational and Graphical Statistics*, 2012.