Speech Acoustic Unit Segmentation Using Hierarchical Dirichlet Processes

Amir Hossein Harati Nejad Torbati[1], *Joseph Picone*[1] *and Marc Sobel*[2]

[1]Department of Electrical and Computer Engineering, Temple University, Philadelphia, USA
[2]Department of Statistics, Temple University, Philadelphia, USA

amir.harati@gmail.com, picone@temple.edu, marc.sobel@temple.edu

Abstract

Speech recognition systems have historically used context-dependent phones as acoustic units because these units allow linguistic information, such as a pronunciation lexicon, to be leveraged. However, when dealing with a new language for which minimal linguistic resources exist, it is desirable to automatically discover acoustic units. The process of discovering acoustic units usually consists of two stages: segmentation and clustering. In this paper, we focus on the segmentation portion of this problem. We introduce a nonparametric Bayesian approach for segmentation, based on Hierarchical Dirichlet Processes (HDP), in which a hidden Markov model (HMM) with an unbounded number of states is used to segment the utterance. This model is referred to as an HDP-HMM. We compare this algorithm to several popular heuristic methods and demonstrate an 11% improvement in finding boundaries on the TIMIT Corpus. A self-similarity measure over segments shows an 88% improvement compared to manual segmentation with comparable segment length. This work represents the first step in the development of a speech recognition system that is entirely based on nonparametric Bayesian models.

Index Terms: nonparametric Bayesian models, hierarchical Dirichlet processes, speech segmentation

1. Introduction

Acoustic unit selection is a critical issue in many speech recognition applications where there are limited linguistic resources or limited training data is available for the target language. For example, recently IARPA's Babel program **Error! Reference source not found.** sponsored a competition to create a speech to text system in a mystery language in one week of time using very limited resources. Though traditional context-dependent phone models perform well when there is ample data, automatic discovery of acoustic units offers the potential to provide good performance for resource deficient languages with complex linguistic structures (e.g., African click languages).

Most approaches to automatic discovery of acoustic units [2]-[4] do this in two steps: segmentation and clustering. Segmentation is accomplished using a heuristic method that detects changes in energy and/or spectrum. Similar segments are then clustered using an agglomerative method such as a decision tree. Advantages of this approach include the potential for higher performance than that obtained using traditional linguistic units, and the ability to automatically discover pronunciation lexicons.

In this paper, we propose the use of a nonparametric Bayesian model (NPBM) for segmentation. In this formulation of the problem, the number of units (or segments) is unknown. One approach is to exhaustively search through a model space consisting of many possible parameterizations. An alternative approach is based on a nonparametric Bayesian statistical model [5][6] in which the model complexity can be inferred directly from the data. Segmenting an utterance into acoustic units can be approached in a manner similar to that used in speaker diarization, where the goal is to segment audio into regions that correspond to a specific speaker. Fox et al. [7] used one state per speaker and demonstrated segmentation without knowing the number of speakers a priori. Here, we demonstrate that a similar approach can be used to segment an utterance into acoustic units.

Our approach is demonstrated in Figure 1 for an example utterance from the TIMIT Corpus [8]. The segmentation is performed using an extension of a hidden Markov model (HMM) with an unbounded number of states and an unbounded number of mixture components. This model is known as an infinite HMM or more recently a Hierarchical Dirichlet Process HMM (HDP-HMM) [7]. It uses a hierarchical Bayesian model to define an NPBM [9].

Relation to Prior Work: We propose a new algorithm for segmentation of speech based on an HDP-HMM [7]. Previously a dynamic programming method was applied that incorporated a heuristic stopping criterion [2]-[4]. Recently, Lee & Glass [10] proposed an NPBM for unsupervised segmentation of speech. A Dirichlet Process Mixture (DPM) model was used. In order to obtain phoneme-like segments, a 3-state HMM was used to model each segment. A Gibbs sampler was employed to estimate the segment's boundaries. In our model, each segment is modeled using one state of an HMM. We use HDP-HMM to discover the optimal number of segments. Note that using a parametric HMM is not possible since the number of segments is unknown.

2. Hierarchical Dirichlet Processes

Hidden Markov models (HMMs) are a class of doubly stochastic processes in which discrete state sequences are

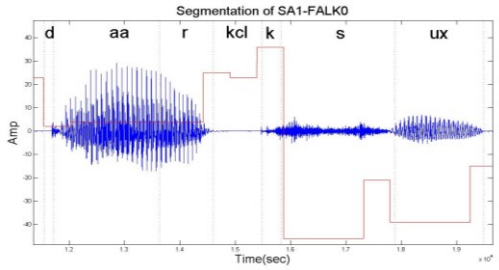

Figure 1: *Segmentation of a speech utterance produced using the proposed algorithm is shown by overlaying the duration and index of each unit on the waveform. The height of each rectangle overlay simply indicates the index of that unit.*

modeled as a Markov chain [11]. We can denote a continuous distribution HMM based on Gaussian mixtures as $\lambda = (\boldsymbol{A},\boldsymbol{B},\pi)$, where $\boldsymbol{A}$ represents the transition probability matrix, $\boldsymbol{B}$ represents the output distributions and π represents the initial state probabilities. The state of a Markov chain at time t is denoted by s_t. Observations are conditionally independent given s_t and are denoted by $o_t \mid s_t, m_t \sim b\left(\theta_{s_t,m_t}\right)$ where s_t and m_t are the state and mixture component indices respectively, and θ_{s_t,m_t} represents the distribution's parameters for state s_t and mixture component m_t. The transition probability from s_t to the next state is denoted by $s_t \mid s_{t-1} \sim a_{s_t,s_{t-1}}$, which implies the current state is only a function of the previous state.

A Dirichlet process (DP) [12][13] is a discrete distribution that is composed of a weighted sum of impulse functions. Although there are many different representations for a DP, we will focus here on two sets of parameters related to the discrete distribution: locations of the impulse functions and their corresponding weights. The impulse functions are often referred to as atoms. For example, in a binomial distribution, there are exactly 2 atoms, x=0 and x=1, and two corresponding weights, P(x=0) and P(x=1). A DP, on the other hand, consists of an infinite number of atoms and corresponding weights, though the sum of the weights of these atoms is constrained to be one.

The distribution from which these atoms can be sampled is known as the base distribution. Weights for these atoms can be generated using a process referred to as stick breaking [14]:

$$c_k = c'_k \prod_{l=1}^{k-1}\left(1-c'_l\right), \qquad c'_k \sim Beta(1,\nu) . \tag{1}$$

We initialize the recursion by drawing a sample, c_1, in the range [0,1], from a Beta distribution in which ν is a concentration parameter. We break the remaining part of the stick, of length 1-c_1, by sampling from another Beta distribution. Each successive break, c_k, represents a weight for a new atom. A small ν means the low order c_k would be much larger on the average than the higher order terms. Hence, ν is one way we control the complexity of the model. This stick-breaking process is also known as the Griffiths, Engen and McCloskey (GEM) model [15].

For reasons that will become clear shortly, we desire to have a collection of DPs that share a set of atoms associated with a base distribution. Specifically, we want these DPs to share the locations of the impulse functions but not the weights. For example, suppose we want to model individual phones using a collection of features. We can employ a DP to model each phone independently much like we would use a standard Gaussian mixture model. Instead of allocating a number of mixture components to each phone, each atom in our DP would represent parameters associated with one mixture component (e.g., means and variances). The weight of the atom would correspond to the mixture weight.

Since each phone in a pure DP approach is modeled independently, no atoms would be shared between them. If we desire to share parameters, we can constrain the phones to use the same set of atoms. This structure, in which we use a common DP for a base distribution, and then model each phone using a DP that shares atoms with other DPs, is known as an Hierarchical DP (HDP) [9]. Note that in an HDP, we still recompute the weights of each shared atom.

An HDP-HMM [7][9][16] is an HMM with an unbounded number of states. An overview of an HDP-HMM is given in Figure 2. In a typical ergodic HMM, the number of states is fixed so a matrix of dimension N states by N transitions per state is used to represent the transition probabilities. In an HDP-HMM, the transition matrix is replaced by an infinite, but discrete transition distribution, modeled by an HDP for each state. This lets each state have a different probability distribution for its transitions while the set of reachable states would be shared among all states.

One side effect of this model is that it does not differentiate between different states. Therefore the probability of self-transition would be small. As a result an HDP-HMM has a tendency to create many redundant states and switch rapidly amongst them. This implies we will have less data per state for parameter estimation (e.g. the mean and covariance of the emission distributions) and therefore the estimates would be less reliable. This has been mitigated by introducing a sticky parameter, κ, to the definition of an HDP-HMM (known as a sticky HDP-HMM) that encourages the probability of a self-transition by introducing a bias term (delta function) favoring the current state [7][9]:

$$\begin{aligned}
&c \mid \nu \sim GEM(\nu)\\
&a_j \mid \eta, c \sim DP(\eta+\kappa, \frac{\eta c + \kappa\delta_j}{\eta+\kappa})\\
&\psi_j \mid \sigma \sim GEM(\sigma)\\
&\theta_{kj} \mid H, \xi \sim H(\xi)\\
&s_t \mid s_{t-1}, \{a_j\}_{j=1}^{\infty} \sim a_{s_{t-1}}\\
&m_t \mid \{\psi_j\}_{j=1}^{\infty}, s_t \sim \psi_{s_t}\\
&o_t \mid \{\theta_{kj}\}_{k,j=1}^{\infty}, s_t, m_t \sim b\left(\theta_{s_t m_t}\right) .
\end{aligned} \tag{2}$$

The state, mixture component and observation are represented by s_t, m_t and o_t respectively. The indices j and k are indices of the state and mixture components respectively. The base distribution that links all DPs together is represented by c and can be interpreted as the expected value of state transition distributions. The transition distribution for state j is a DP denoted by a_j with a concentration parameter η. Another DP, Ψ_j, with a concentration parameter σ, is used to model an infinite mixture model for each state (s_j). The distribution H is the prior for the parameters θ_{kj}. If we want the posterior distribution over the parameters to remain in the same family

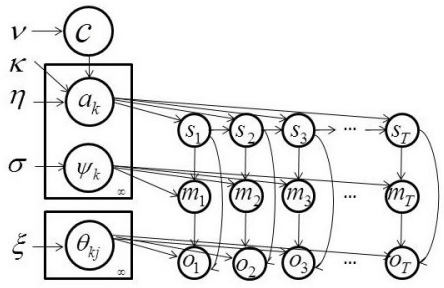

Figure 2: *A graphical representation of an HDP-HMM is shown in which s_t, m_t and o_t represent state, mixture component and observation respectively.*

as the prior, then H should be chosen to be a conjugate prior to the observation likelihood. Since the likelihood is a multivariate normal, we use a normal inverse Wishart distribution for H.

The final ingredient in this model is an inference algorithm. Inference algorithms are used to infer the values of the latent variables, in this case s_t and m_t. Equation (2) describes a generative model. There are several popular approaches for inference including the block sampler [7] used in this work. The block sampler employs a Markovian structure to improve its performance. It has been shown that if all states are sampled at once, algorithm will be much more computationally efficient. In the block sampler a variation of the forward-backward procedure used to sample the state sequence in one step. A block sampler needs a fixed truncation level, K_s, to be specified in advance, representing the maximum number of states that could be found.

It should be noted that the resulting algorithm is different from a parametric Bayesian HMM because it induces a sparse subset of the K_s possible states, while a parametric model always finds K_s states. Similarly, a fixed truncation level, K_m, is used for the number of mixture components per state. In practice, if both K_s and K_m are sufficiently large, the results will be the same as if we use an infinite truncation level.

In our model, each state of the HMM represents a segment. Since an HDP-HMM has an unbounded number of states, the model can infer the number of segments automatically from the data. Modeling each segment with a state of an HMM means that the algorithm segments speech into semi-stationary parts.

3. Experiments

To evaluate the proposed algorithm, we used the TIMIT Corpus [17] because of the existence of highly accurate manual segmentations. Each utterance was converted into a standard 39-dimensional MFCC feature stream (12 MFCC coefficients plus energy as well as the first and second derivatives) computed using 10 ms frames. Next, L consecutive frames of data are averaged to produce one output frame of features per L frames. This averaging process is done to ensure that segments have a minimum duration of L frames. Typically, L varies from 1 to 3.

In an HDP-HMM model, there are several parameters that must be adjusted, including the truncation level for the number of states (K_s) and the truncation level for the number of mixture components (K_m) per state. These should be set to be larger than the expected number of states and mixture components per state but not too large since computational complexity increases quadratically with K_s and K_m.

To measure the performance of the segmentation we followed the approach used in [10] with a tolerance window of 20 ms. In this approach the discovered boundaries of the segments are compared to the manually segmented reference boundaries. The number of co-occurrences of segments boundaries and phoneme boundaries is called recall. The percent of declared boundaries that coincide with phoneme boundaries is called precision. A single numeric score is referred to as the F-score and defined as:

$$\text{F-score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} . \quad (3)$$

Table 1. *Segmentation performance of HDP-HMM improves recall while maintaining an acceptable precision.*

Algorithm	Recall	Prec.	F-score
Dusan & Rabiner (2006) [18]	75.2	66.8	70.8
Qiao et al. (2008) [19]	77.5	76.3	76.9
Lee & Glass (2012) [10]	76.2	76.4	76.3
HDP-HMM	**86.5**	68.5	76.6

A comparison of our proposed method to other state of the art systems is shown in Table 1. The first row represents a system that performs unsupervised segmentation with no prior information about number of segments for each utterance. The second row represents a system that implements a semi-supervised approach. The third row provides results for a recently proposed nonparametric Bayesian approach [10].

The HDP-HMM algorithm performs particularly well on recall, which implies that it is finding boundaries that better match the reference phoneme boundaries. The improvement in recall is over 11% even though; our approach unlike [19] is completely unsupervised. The precision is lower, however, which means there are slightly more false alarms. This is not unexpected since its determination of acoustic units is driven by the complexity of the data.

The quality of the segments can be measured using a similarity score, defined as:

$$S(s_1, s_2) = \begin{cases} s_1 = \dfrac{1}{MN} \sum\limits_{\substack{I=\{i\} \\ M=|I|}} \sum\limits_{\substack{J=\{j\}=\{j:class(j)=class(i), i\neq j\} \\ N=|J|}} \left| corr(x_i, x_j) \right| \\ s_2 = 1 - \dfrac{1}{M'N'} \sum\limits_{\substack{I=\{i\} \\ M'=|I|}} \sum\limits_{\substack{J=\{j\}=\{j:class(j)\neq class(i)\} \\ N'=|J|}} \left| corr(x_i, x_j) \right| \end{cases} \quad (4)$$

This score, S, is an indicator of consistency. It has two main components: (1) s_1 is the in-class similarity score and is defined as the average over the correlation between different instances of segments with identical labels; (2) s_2 is the out-of-class dissimilarity score. The quality of segmentation is higher when both numbers are closer to one. It should be noted that the similarity score functions much like a likelihood score – it increases monotonically with an increase in the number of classes. Therefore, for a meaningful comparison, the number of classes being compared for two algorithms must be the same (defined as the number of segments with same identity) or equivalently the average length of segments produced by the two algorithms should be comparable.

In Table 2, we demonstrate segmentation performance. N_c is the number of discovered classes. Similarity scores for the manual segmentations and the proposed algorithm are shown in the last two columns of Table 2. Each cell consists of a pair of numbers corresponding to s_1 and s_2. The number of classes for the manual segmentations is fixed to 61, the number of phones used to mark the corpus. For HDP-HMM, N_c varies between 23 and 139.

Note that increasing the number of classes results in an increase in the in-class similarity scores, but the out-of-class dissimilarity scores remain relatively constant. If we consider the last row of the table, we observe that the number of classes (51) is roughly comparable to the number of phones (61). In this case the average length of discovered segments is slightly longer than standard phonemes, yet the similarity score for our

Table 2. *Performance of the HDP-HMM approach to automatic discovery of acoustic units on TIMIT is shown. The in-class similarity scores for the proposed algorithm are significantly higher than those for the manual segmentations.*

Experiment	N_c	Manual	HDP-HMM
K_s=100, K_m=1, L=1	70	(0.44,0.72)	(0.82,0.73)
K_s=100, K_m=1, L=2	33	(0.44,0.72)	(0.77,0.73)
K_s=100, K_m=1, L=3	23	(0.44,0.72)	(0.75,0.72)
K_s=100, K_m=5, L=1	139	(0.44,0.72)	(0.90,0.72)
K_s=100, K_m=5, L=2	73	(0.44,0.72)	(0.87,0.72)
K_s=100, K_m=5, L=3	51	(0.44,0.72)	(0.83,0.72)

algorithm is 88% larger (0.83 vs. 0.44). This means segments discovered by our algorithm are more self-consistent and therefore can be modeled using simpler models.

We have also performed a qualitative analysis of our proposed algorithm. In Table 3, excerpts from automatically discovered lexicons are shown for four different parameter configurations. This data was the result of processing utterance SA1 for speakers FALK0 and FCJF0. The labels shown are arbitrarily assigned during the discovery process. Though we don't expect the value of the label to be repeated for a different set of data, we can see that there is a general similarity in the sequence of labels for similar words spoken by different speakers. For example, word "all" in the first row of the table is represented by segments "60-54-80-41" for FALK0 and "29-54-80-41" for FCJF0.

Further analysis revealed that the segments 60 and 29 are also acoustically close. The normalized distance between the mean of the Gaussian distributions that represent each segment is 11.6 while the average distance between two arbitrary segments is 41.1. This indicates that segments 29 and 60 are accounting for different pronunciations of the initial phone; e.g. it is expected further clustering step merge these two segments into one cluster.

Segments derived using the proposed algorithm follow an *N*-gram statistical structure. For example, in the second row of Table 3, segment 79 always follows segment 18, and segment 12 always follows segments 70, 79 and 68, which are very close in terms of acoustic distance. This suggests that the discovered segments are similar to phoneme like units.

The first two experiments use a single Gaussian emission for each state (K_m=1). The last two experiments use Gaussian mixtures (K_m=5) where the maximum number of components per state is K_m. The flexibility added by the mixture model improves the consistency of the segmentation. For example, by comparing the word "she" for the first and third experiments in Table 3, we observe that the segmentations for both speakers are much more similar in the third experiment (K_m=5) than the first experiment (K_m=1). Recall that in this model the number of mixture components per state can vary, and the number of derived classes grows only as needed based on the complexity of the data. Hence, the model essentially adapts to the data.

Figure 1 demonstrates that the boundaries found by HDP-HMM approximately coincide with boundaries found from manual segmentation, supporting the results reported in Table 1. However, in some cases the discovered segments cover several phonemes (e.g., /aa r/) while in other instances a single phoneme is divided into more than one segment (e.g., /s/). This splitting does not violate the phoneme boundaries and can be interpreted as a finer representation of the phoneme. This is supported by the fact that for a comparable number of classes (e.g. segment identity) the similarity score is higher for the automatically discovered segments. This suggests that the splitting/merging phenomenon inherent to the HDP-HMM improves the segmentation process and we expect the resulting segments to generate a set of acoustic units that represent the data more consistently.

Table 3. *Samples of the lexicons are shown for several parameter configurations. The labels in the third and fourth columns are arbitrarily assigned to acoustic units. There is a reasonable amount of consistency between words with similar phonetic transcriptions.*

Exp.	Word	FALK0	FCJF0
K_s=100 K_m=1 L=1	She	81-2-7-41	27-67-40-41-68
	Wash	45-25-29-54-59-30-94-81	41-45-25-29-54-73-8-4-27-81-17
	Water	29-54-59-28-71-72-98	29-54-28-98
	All	60-54-80-41	29-54-80-41
K_s=100 K_m=1 L=2	She	60-18-79-70	27-67-40-41-68
	Wash	75-10-51-91-52-60-61	75-10-51-91-19-54-60-61
	Water	10-51-3-99	10-51-3
	All	10-51-70	10-51-70
K_s=100 K_m=5 L=1	She	35-75-43-89	35-76-43-89
	Wash	70-29-48-47-88-7-100-35-41	70-48-47-88-7-15-6-35-41
	Water	48-47-88-73-50-57-45	47-88-39-47
	All	25-87-7-43	47-30-43
K_s=100 K_m=5 L=3	She	24-6-86	17-38-6-30-58
	Wash	43-26-30-73-24	5-43-26-30-76-10-17-59-78
	Water	43-26-30-50-69	26-50-80
	All	26-30-69-55	26-69

4. Conclusions

We have investigated the application of an HDP-HMM model to segmentation of speech. It was shown that this segmentation model produces meaningful and consistent results. Discovered boundaries generally coincide with the boundaries for manually segmented phonemes. It was shown that for a comparable number of classes (e.g. phonemes), the proposed algorithm improves segmentation self-similarity score by more than 88% over the manual segmentation despite the fact that average length of automatically discovered segments are slightly longer than average length of phonemes. Moreover, we have shown that our algorithm improves recall rate over other state of the art algorithms by more than 11%.

Future research will be focused on clustering segments produced by HDP-HMM and automatic generation of a corresponding lexicon. This step can also be implemented using a nonparametric Bayesian approach. This is the last crucial step in achieving our goal of a system based entirely on nonparametric Bayesian approaches.

5. Acknowledgements

This research was supported in part by the National Science Foundation through Major Research Instrumentation Grant

No. CNS-09-58854.

6. References

[1] M. Harper, "IARPA Solicitation IARPA-BAA-11-02," *IARPA BAA*, 2011. Available: http://www.iarpa.gov/solicitations_babel.html.

[2] M. Bacchiani and M. Ostendorf, "Joint lexicon, acoustic unit inventory and model design," *Speech Communication*, vol. 29, no. 2–4, pp. 99–114, 1999.

[3] B. Ma, et al., "An Acoustic Segment Modeling Approach to Automatic Language Identification," in *Proc. of INTERSPEECH*, 2005, pp. 2829–2832.

[4] K. Paliwal, "Lexicon-building methods for an acoustic sub-word based speech recognizer," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1990, pp. 729– 732.

[5] P. Muller and F. A. Quintana, "Nonparametric Bayesian Data Analysis," *Statistical Science*, vol. 19, no. 1, pp. 95–110, Feb. 2004.

[6] A. Harati, J. Picone, and M. Sobel, "Applications of Dirichlet Process Mixtures to Speaker Adaptation," in *Proceedings of ICASSP*, 2012, pp. 4321–4324.

[7] E. Fox, et al., "A Sticky HDP-HMM with Application to Speaker Diarization.," *The Annalas of Applied Statistics*, vol. 5, no. 2A, pp. 1020–1056, 2011.

[8] V. Zue, et al., "Acoustic Segmentation and Phonetic Classification in the SUMMIT System," in *Proceedings of ICASSP*, 1989, pp. 389–392.

[9] Y. Teh and M. Jordan, "Hierarchical Bayesian Nonparametric Models with Applications," in *Bayesian Nonparametrics: Principles and Practice*, Cambridge-UK: Cambridge University Press, 2010, pp. 158–207.

[10] C. Lee and J. Glass, "A nonparametric Bayesian approach to acoustic model discovery," in *Proceedings of the ACL*, 2012, pp. 40–49.

[11] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 879–893, 1989.

[12] C. Antoniak, "Mixtures of Dirichlet Process with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, vol. 2, no. 7, pp. 1152–1174, 1974.

[13] C. E. Rasmussen, "The Infinite Gaussian Mixture Model," in *In Advances in Neural Information Processing Systems*, MIT Press, 2000, pp. 554–560.

[14] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[15] J. Pitman, "Random Discrete Distributions Invariant under Size-Biased Permutation," *Advances in Applied ProbabilityApplied Probability Trust*, vol. 25, no. 2, pp. 525–539, 1996.

[16] M. Beal, Z. Ghahramani, and C. E. Rasmussen, "The Infinite Hidden Markov Model," in *Advances in Neural Information Processing Systems*, 2002, pp. 577–584.

[17] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, N. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *The Linguistic Data Consortium Catalog*, 1993.

[18] S. Dusan and L. Rabiner, "On the relation between maximum spectral transition positions and phone boundaries," in *Proceedings of INTERSPEECH*, 2006, pp. 1317–1320.

[19] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: Objectives, algorithms and comparisons," in *Proceedings of ICASSP*, 2008, pp. 3989–3992.