Automatic Speech Segmentation Using Probabilistic Latent Component Modeling

Sayan Ghosh and T.V. Sreenivas

Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India

sayghosh@ece.iisc.ernet.in, tvsree@ece.iisc.ernet.in

Abstract

Latent variable methods, such as PLCA (Probabilistic Latent Component Analysis) have been successfully used for analysis of non-negative signal representations. In this paper, we formulate PLCS (Probabilistic Latent Component Segmentation), which models each time frame of a spectrogram as a spectral distribution. Given the signal spectrogram, the segmentation boundaries are estimated using a maximum-likelihood approach. For an efficient solution, the algorithm imposes a hard constraint that each segment is modelled by a single latent component. The hard constraint facilitates the solution of ML boundary estimation using dynamic programming. The PLCS framework does not impose a parametric assumption unlike earlier ML segmentation techniques. PLCS can be naturally extended to model coarticulation between successive phones. Experiments on the TIMIT corpus show that the proposed technique is promising compared to most state of the art speech segmentation algorithms.

Index Terms:Speech segmentation, PLCA, Spectrograms, Coarticulation, Dynamic Programming

1. Introduction

Automatic speech segmentation is a classical signal processing problem which is important for several applications. An automated approach to segmentation is necessary for speech data too large to be segmented manually. There are two primary approaches to segmentation. The *homogeneity* approach strives to find the optimal segment boundaries so that the distortion within each segment is minimized. The ML segmentation proposed by Svendsen et al. [1] is of this category. The *heterogeneity* approach to segmentation makes use of the dissimilarity between successive segments, which is expected to be highest at the segment boundaries. The STM approach [2] is an example of heterogeneity based segmentation. More recent approaches to speech segmentation include MMC (Maximum Margin Clustering) by Estevan et al. [3] and the usage of model selection criteria [4]. In [10], Aversano et al. develop a segmentation method which does not rely on prior knowledge of the phoneme sequence. Qiao et al [5] formulate the segmentation problem using a probabilistic framework. They formulate three optimal objective functions for segmentation: Mean Square Error (MSE), Log Determinant (LD) and Rate Distortion (RD). In [6], Rasanen et al. propose an improved quality measure, the R-value to evaluate the performance of unsupervised speech segmentation algorithms. The difficulty for segmentation is mainly due to the non-stationarity within the segments which poses a problem for both homogeneity based methods as well as heterogeneity based methods. The mixture of stochastic models proposed below has the promise of handling non-stationarity better.

The last decade has witnessed a growing interest in analysis of non-negative data and non-negative signal representations. Unsupervised learning methods, such as NMF [7] and PLCA (Probabilistic Latent Component Analysis) [8] have been successfully applied to a variety of problems such as music transcription, acoustic source separation, speech and audio denoising, spectrogram superresolution and speech dereverberation [9], [11], [12]. Given a non-negative time-frequency representation $N(t, f)$ of a signal, PLCA decomposes the spectral distribution $P(f|t) = \frac{N(t,f)}{\sum_f N(t,f)}$ at each time frame t as a sum over a set of M latent components:

$$P(f|t) = \sum_{z=1}^{M} P(f|z)P(z|t) \tag{1}$$

Our motivation in applying the PLCA based decomposition to the speech segmentation task is to model a speech spectrogram as a superposition of different segments, where each speech segment can be modelled by a separate latent component. We note that unlike in most conventional segmentation algorithms, we do not make any underlying assumption that the time frames in each segment are drawn from a parameterized distribution, such as the Gaussian density. Thus the PLCS framework would be valid even when the segments are drawn from arbitrary distributions, or when there are too few frames in each segment to reliably learn a parameterized distribution. Conventional approaches to segmentation generally extract relevant frame based features from the speech signal prior to segmentation. Use of different sets of features would yield different results, with MFCCs shown to be the most robust for phoneme segmentation. In contrast, our method does not involve the extraction of specialized features. The only input required is a spectrogram of the signal to be segmented. Features such as LPCCs, MFCCs and LSFs are not necessarily non-negative, whereas the spectrogram is a non-negative signal representation. The spectrogram is obtained by computing the squared absolute value $|S_x(t, f)|^2$ of the STFT (Short-Time Fourier Transform) $S_x(t, f)$ of signal $x(t)$, which makes it a non-negative representation.

The advantage of imposing the non-negativity constraint is that the input data is decomposed into a *parts based representation*. Speech signals exhibit the phenomenon of coarticulation, where adjacent segments merge into one another, thus posing an additional problem to the automated segmentation algorithm. In the proposed framework, the spectral vectors at the phone boundaries can be modelled as a linear combination of the preceding and successive segment latent components. This is facilitated by the non-negative signal representation, which is additive in nature. Conventional features such as MFCCs do not follow linear superposition, and hence do not lend themselves readily to such modeling.

2. Formulation of PLCS

Let the spectrogram of the signal which we wish to segment into M successive parts be $N(t, f)$. Let $t \in \{1, 2, ..., T\}$ and $f \in \{1, 2, ..., F\}$. Thus we are considering T time bins and F frequency bins in the TF-representation. Let us treat the spectrogram as a co-ocurrence matrix of time and frequency (a 2D discrete pmf), ie., let us assume that we have N ”virtual” observations of time-frequency tuples, of the form $(t_1, f_1), (t_2, f_2), (t_3, f_3), ...(t_N, f_N)$, where the number of observations of the tuple (t', f') out of N observations is given by $N(t', f')$. Hence we have:

$$\sum_{t=1}^{t=T} \sum_{f=1}^{f=F} N(t, f) = N \tag{2}$$

By using the symmetric PLCA model [8], we can express:

$$P(t, f) = P(t) \sum_{z=1}^{z=M} P(f|z)P(z|t) \tag{3}$$

where $P(t, f)$ is the joint probability of time bin t and frequency bin f. If we assume that the N observations of time-frequency tuples are mutually independent, then the joint likelihood of all L observations is given by:

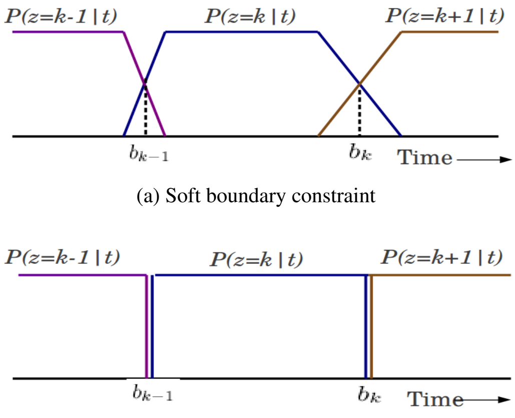
$$L = \prod_{i=1}^{i=N} P(t_i, f_i) = \prod_{t=1}^{t=T} \prod_{f=1}^{f=F} \left[P(t) \sum_{z=1}^{M} P(f|z)P(z|t) \right]^{N(t,f)} \tag{4}$$

In the above relation, we have grouped together all time-frequency tuples with the same time-frequency bin (t, f). The log-likelihood of the data is now given by:

$$LL = \sum_{i=1}^{i=N} \sum_{f=1}^{f=F} N(t, f) \left[logP(t) + log \sum_{z=1}^{M} P(f|z)P(z|t) \right] \tag{5}$$

2.1. Towards a model for segmentation

Consider the signal to be segmented to have M segments, where we can model the m-th segment to correspond to a latent basis $P(f|z = m)$. Let the M segment boundaries be denoted as $\{b_0, b_1, b_2, ..., b_{M-1}\}$, where the m-th segment is in the range $b_{m-1} + 1 \leq t \leq b_m$. In this model no attempt is made to find similarity between segments that are not contiguous. Repetition of a similar segment elsewhere in the signal is considered a distinct segment and a distinct latent component. The weighing parameter $P(z = m|t)$ of the segmentation model is a measure of the degree of excitation of the m-th segment at the t-th time frame. This is well suited for modeling time frames at phoneme boundaries. For example, consider the k-th phone segment in an utterance, where it is coarticulated with the $(k - 1)$-th segment as well as the $(k + 1)$-th segment. In Figure 1a, we show how our model can enable the spectra $P(f|t)$ in the region $b_{k-1} - \delta \leq t \leq b_k + \delta$ (region of coarticulation) to be expressed as a convex linear combination of the latent bases of the $(k-1)$-th and k-th segments; a similar combination holds in the coarticulation region of the k-th and the $(k + 1)$-th segments. This is *physically intuitive* since the spectrogram is an additive representation when neglecting the cross terms. Conventional features such as MFCCs are non-linear with respect to their additive signal components and hence do not follow such a superposition property. However for computational simplicity, we make use of a hard constraint instead of a soft linear combination, as we shall describe in the next section.



(a) Soft boundary constraint

(b) Hard boundary constraint

Figure 1: Hard vs. soft segment boundary constraints

2.2. Incorporating a hard constraint

We impose a hard constraint that the spectra $P(f|t)$ within the m-th segment $b_{m-1} + 1 \leq t \leq b_m$ corresponds to only one latent component, which is the basis $P(f|z = m)$ for the m-th segment. This is depicted in Figure 1b. Thus, the segmentation model reduces to:

$$P(f|t) = P(f|z = m), \forall\, b_{m-1} + 1 \leq t \leq b_m \tag{6}$$

The hard constraint restricts the spectra in each segment to belong to only one latent component, thus imposing a form of spectral stationarity within each segment. While this constraint does not account for coarticulation, it is found to be reasonable and yields good results when applied to speech data. The objective function to be maximized now becomes:

$$\sum_{m=1}^{M} \sum_{t=b_{m-1}+1}^{b_m} \sum_{f=1}^{f=F} N(t, f) logP(f|z = m) \tag{7}$$

subject to the condition $\sum_{f=1}^{F} P(f|z = m) = 1$. To maximize the likelihood, we have to solve for the segment boundaries $\{b_0, b_1, ..., b_M\}$ and the latent bases $P(f|z = m) \forall\, m$. It is straightforward to show that given the boundaries, the “centroid” (latent basis) for the m-th segment is given by:

$$\mu_m = \hat{P}(f|z = m) = \frac{\sum_{t=b_{m-1}+1}^{b_m} N(t, f)}{\sum_{f=1}^{F} \sum_{t=b_{m-1}+1}^{b_m} N(t, f)} \tag{8}$$

We can further show that the algorithm strives to minimize the following cost function in terms of the segment boundaries:

$$\sum_{m=1}^{M} \sum_{t=b_{m-1}+1}^{b_m} d_{spec}(t, m) \tag{9}$$

The distance measure $d_{spec}(t,m)$ is given by:

$$d_{spec}(t,m) = \sum_{f=1}^{f=F} -N(t,f)log\mu_m(f) = N(t)\mathbf{H}[P(f|t)||\mu_m] \tag{10}$$

$\mathbf{H}[P(f|t)||\mu_m]$ is the cross-entropy between the spectral distribution at time t, and the spectral centroid $\mu_m = \hat{P}(f|z=m)$ of the m-th segment as obtained from Equation (8). This cross entropy is weighted by the energy term $N(t)$, which is the time-marginal of $N(t,f)$. Even though the energy term appears in Equation (10), experiments on synthetic data as well as speech clearly show that it does not aid the segmentation task by utilizing energy information of each time frame. The cross-entropy term incorporates the normalized spectral distribution $P(f|t)$ and mean basis μ_m, and only the spectral information is taken here into account. Similar to the approach in [1], we resort to dynamic programming to solve for the optimum segment boundaries. We also impose a condition that the duration of each speech segment is between $d_{min} = 25ms$ and $d_{max} = 200ms$, which reduces the complexity of the DP search procedure.

2.3. Taking energy into account

In the PLCS framework above, we use only the spectral information contained in the input spectrogram $N(t,f)$. However, the energy information is important for the segmentation task, especially when marking regions of silences and stops in the speech signal. We would like to utilize the homogeneity in energy information also for better segmentation. We propose to use an energy cost $d_{ener}(t,m)$ where

$$d_{ener}(t,m) = (N(t) - \bar{N}_m)^2 \tag{11}$$

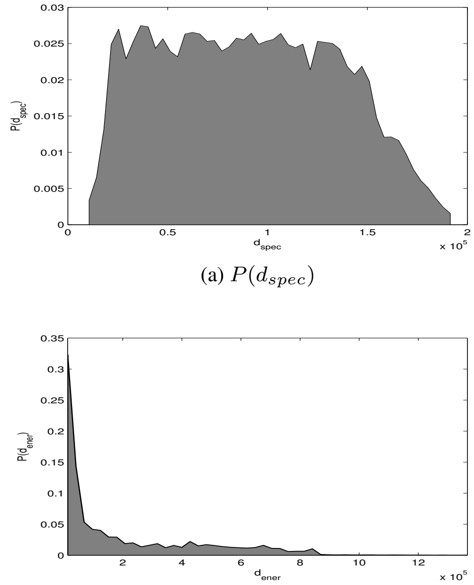
where $\bar{N}_m = \frac{1}{b_m - b_{m-1}} \sum_{t=b_{m-1}+1}^{b_m} N(t)$ is the mean energy of the m-th segment. We note that the spectral distance $d_{spec}(t,m)$ is of a different order of magnitude than $d_{ener}(t,m)$. Figure 2 shows the distribution of d_{spec} and d_{ener} obtained from a training set of 50 TIMIT sentences. We transform d_{spec} and d_{ener} to d'_{spec} and d'_{ener} respectively so that the resulting distributions are normalized to zero mean and unit variance. We modelled the distributions d_{spec} and d_{ener} non-parametrically, and did not use a parameterized distribution, such as a Gaussian. Parameterized distributions would not model these distributions well, and would result in a decrease in performance. The modified cost function to be maximized is now:

$$\sum_{m=1}^{M} \sum_{t=b_{m-1}+1}^{b_m} \left[d'_{spec}(t,m) + \lambda d'_{ener}(t,m)\right] \tag{12}$$

where λ is a tuning parameter designed to give higher weightage to the spectral variations after the normalization. We have experimented with a set of 50 TIMIT sentences and found that $\lambda = 0.05$ gives the best segmentation performance.

3. Experiments on Speech Data and Results

We perform experiments on the TIMIT corpus to test the effectiveness of the proposed segmentation framework. The algorithms we have implemented in our experiments are :(1) PLCS (Probabilistic Latent Component Segmentation) with a hard boundary constraint (2) PLCS with an energy based cost (PLCS-E). The performance measures we have used are :(1) % Hit rate (H.R) (2) % Insertion rate (I.R) (3) % Successive segment hit rate (S.H.R). Following the widely used convention in the literature [5], [3], we consider an automatic segment to be a hit only if it lies within 20 ms. of a TIMIT boundary. The % H.R is the fraction of TIMIT boundaries which have been correctly detected. % S.H.R denotes the percentage of consecutive hits. All experiments were performed on five dialects of the training set of TIMIT corpus. We have used wideband log-spectrograms, with a frame shift of 5 ms and a window length of 20 ms. In all experiments, the target number of segments for DP search is set equal to that in the TIMIT transcriptions.



(a) $P(d_{spec})$

(b) $P(d_{ener})$

Figure 2: Distributions of $P(d_{spec})$ and $P(d_{ener})$

Table 1: *Performance of PLCS framework*

Algorithm	*% H.R.*	*% I.R.*	*% S.H.R*
PLCS	74.55	27.03	49.6
PLCS-E	76.36	22.75	51.3

The results are reported in Table 1. From the Table, we observe that the proposed PLCS algorithm and its variant attains an average hit rate of 75.45%, with a mean insertion rate of 24.89% and a mean successive segment hit rate of 50.45%. Variant PLCS-E is slightly better than the original PLCS in terms of a 1.81% increase in segment matches, but has a decrease of 4.28% in terms of spurious insertions. Since we have used the same number of segments as the TIMIT transcription, our results correspond to an over-segmentation (o.s. rate) of 0%. The Gaussian ML segmentation in [1] yields a hit rate of 80% with an insertion rate of 21% for no oversegmentation. In [3], the reported c.d.r (equivalent to % Hit rate) is 76%. In [10], the reported % match is 73.6% at an o.s. of 0%. In the STM approach of [2], the authors report 84.6% correctly detected boundaries, of which 89% lie within a margin of 20 ms, which is equivalent to a hit rate of 75.2%, and an insertion rate of 28.2%. In [5], the authors report a best performance of 77.5% using a Rate-Distortion approach. Thus, the PLCS framework

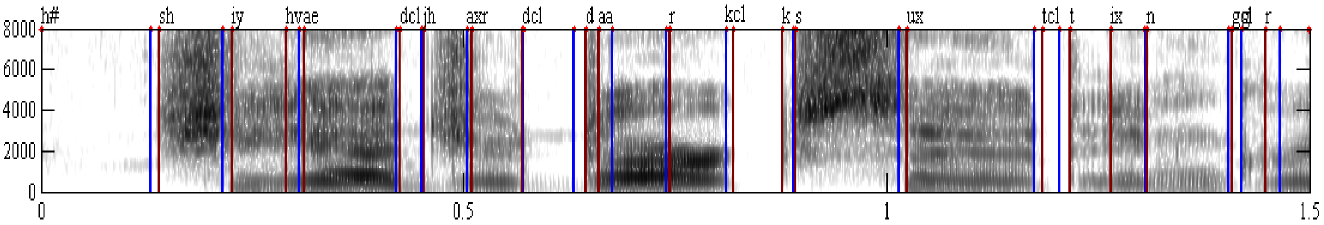


Figure 3: Spectrogram of a section of a TIMIT utterance with manual (red) and automatic (blue) boundaries superimposed

yields results comparable to the latest state of the art segmentation algorithms.

In Figure 3, we show the spectrogram of the initial 1.5 seconds of a TIMIT sentence ('she had your dark suit in gr'), with the manual and automatic boundaries superimposed. It is observed in Figure 3 that the boundaries corresponding to the start of closures \dcl\and \gcl\, the fricatives \jh\and \s\, the nasal \n\, and the semivowels \r\and \axr\are very close to the TIMIT boundaries. The algorithm groups \iy\hv\and \t\ix\into a single segment. Overall, the boundaries are very similar to that obtained by a visual inspection of the spectrogram.

To obtain a better insight into the PLCS framework, we have grouped all TIMIT phones into six broad categories : Stop Consonants (C), Fricatives (F), Nasals (N), Semivowels (SV), Vowels (V), Pause/Silence (P) and presented the proportion of boundaries correctly detected (for one TIMIT dialect) for different classes of preceding and following segments in Table 2. It is observed that PLCS achieves high accuracy for dissimilar segment classes, such as vowel-consonant and vowel-nasal boundaries. Even for similar classes, such as vowel-vowel boundaries, the boundary detection accuracy is over 50%. Modeling coarticulation in the PLCS framework would be expected to further increase accuracy.

Table 2: *Boundary Detection Accuracy for different phone classes (best entries in the table are shaded)*

Class	C_t	F_t	N_t	SV_t	V_t	P_t
C_{t-1}	$\frac{3}{4}$	$\frac{133}{227}$	$\frac{39}{56}$	$\frac{203}{253}$	$\frac{615}{859}$	$\frac{27}{76}$
F_{t-1}	$\frac{85}{189}$	$\frac{18}{40}$	$\frac{32}{32}$	$\frac{81}{92}$	$\frac{570}{650}$	$\frac{93}{141}$
N_{t-1}	$\frac{164}{202}$	$\frac{84}{97}$	$\frac{8}{15}$	$\frac{17}{31}$	$\frac{233}{278}$	$\frac{21}{34}$
SV_{t-1}	$\frac{77}{81}$	$\frac{44}{46}$	$\frac{14}{20}$	$\frac{23}{42}$	$\frac{444}{717}$	$\frac{8}{11}$
V_{t-1}	$\frac{778}{911}$	$\frac{563}{624}$	$\frac{418}{495}$	$\frac{233}{428}$	$\frac{111}{195}$	$\frac{50}{76}$
P_{t-1}	$\frac{81}{88}$	$\frac{66}{110}$	$\frac{30}{39}$	$\frac{47}{71}$	$\frac{24}{30}$	-

4. Conclusions

We have proposed a novel segmentation framework, based on Probabilistic Latent Component Analysis (PLCA) which is fundamentally different from existing parametric pdf-based segmentation schemes. The framework is based on the probabilistic decomposition of a non-negative signal representation which can also model coarticulation between successive phonemes. Using a hard constraint on the segment boundaries to simplify the framework, a DP-search based procedure is employed to solve for the segment boundaries. Performance is further improved by adding an energy cost to the objective function. Experiments on speech data from the TIMIT corpus show results comparable to contemporary results in the literature. The present work can be extended to incorporate soft constraints on the boundaries, and to speaker segmentation, where each segment can be modelled as a mixture of latent components.

5. References

[1] T. Svendsen and F.K. Soong, On the Automatic Segmentation of Speech Signals, Proceedings. of ICASSP-Dallas,1987,

[2] S. Dusan and L. Rabiner, On the Relation between Maximum Spectral Transition Positions and Phone Boundaries, Proc. Inter-Speech, Pittsburg, PA, Sept. 2006.

[3] Y. P. Estevan, V. Wan, and O. Scharenborg, Finding Maximum Margin Segments in Speech, ICASSP, pp. 937940, 2007.

[4] George Almpanidis, Margarita Kotti, Constantine Kotropoulos, Robust Detection of Phone Boundaries Using Model Selection Criteria With Few Observations., IEEE Transactions on Audio, Speech and Language Processing 17(2): 287-298 (2009)

[5] Y. Qiao, N. Shimomura, N. Minematsu, Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons, Proc. ICASSP, pp.3989-3992, 2008

[6] O. J. Rasanen, U. K. Laine, and T. Altosaar. An improved speech segmentation quality measure: the R-value. In Interspeech, pages 18511854, 2009

[7] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature, vol. 401, no. 6755, pp. 78891, 1999.

[8] P. Smaragdis, B. Raj, and M. Shashanka, A probabilistic latent variable model for acoustic modeling, in In Workshop on Advances in Models for Acoustic Processing at NIPS, 2006.

[9] G. J. Mysore, P. Smaragdis, and B. Raj, Non-negative hidden markov modeling of audio with application to source separation, in LVA/ICA, 2010, pp. 140148.

[10] Aversano, Guido et al. A New Text-Independent Method for Phoneme Segmentation. Proceedings of the 44th IEEE Midwest Symposium on Circuits and Systems (2001) : 516 - 519.

[11] G. Grindlay, D.P.W. Ellis, Transcribing Multi-Instrument Polyphonic Music With Hierarchical Eigeninstruments, IEEE Journal of Selected Topics in Signal processing, Volume 5, Issue 6, Oct 2011

[12] J.Nam, G. J. Mysore, J. Ganseman, K. Lee, J.S. Abel, A Super-Resolution Spectrogram Using Coupled PLCA, Proceedings of Interspeech 2010, pg. 1696-1699.