**COLLEGE OF ENGINEERING**

**Internal Reports**

**Segmentation of Speech Utterances using HDP-HMM: The Pilot experiment**

Amir Harati and Joseph Picone

Amir.harati@gmail.com, picone@temple.edu

Institute for Signal and Information Processing

, Department of Electrical Engineering,

Temple University

July 2012

****



**Executive Summary**

One of the important assumptions in designing a speech recognizer is the availability of some sort of acoustic units and a dictionary based on these units. For languages like English, these predesigned units are usually phonemes and dictionaries are existed that map words into these units. However, there are two important problems related to predefined units similar to phonemes. First, these units are designed based on linguistic knowledge by human experts; as a result they are not optimal from an acoustical modeling perspective. Secondly, for many languages (and even accent within a language) there is not enough linguistic knowledge and resources to design a unit inventory and build a dictionary upon it. Therefore, during past years some researchers tried several approaches to generate inventory and lexicon directly from acoustic data without using any linguistic knowledge. These efforts were somewhat successful; best reported results match of those obtained with phoneme based system (Bacchiani and Ostendorf, 1999).

The problem of automatic discovery of acoustical units can be decomposed into several sub-problems. Particularly, it can be seen as a segmentation problem followed by clustering. In other words, the acoustic data is first segmented into several parts and then these parts are clustered into several units. The difference between these algorithms often is related to the way they approach segmentation and clustering problems. Most algorithms use lots of heuristics in both of these stages. Segmentation is often implemented with some kind of dynamic programming algorithm with criteria such as changes in the spectrum (Bacchiani and Ostendorf, 1999). Clustering often implemented based on Kmeans, tree or other more sophisticated (yet heuristic) approaches (Paliwal, 1990) and (Bacchiani et al, 1996). People also use different ways to prune the results and tune the algorithm.

In this pilot project, we investigate the segmentation problem. Our approach, unlike most other works, is not heuristic and is based on a very well-grounded mathematical theory. We propose to use a nonparametric Bayesian method to segment the speech. This algorithm is based on assuming relatively homogenous segments. Initially the number of segments along with their boundaries is unknown. This means the problem fits within the nonparametric class of problems. In case of nonparametric Bayesian we use a proper prior (belief) for the model. In other words, we restrict the model without clarifying the number of segments (parameters) in advance. If we use good priors Bayesian approach can lead to a better result in compare to non-Bayesian ones.

For segmentation problems, infinite Hidden Markov Models (HMMs) also known as HDP-HMMs have been recently proposed (Fox et al., 2011). This model has been used successfully in speaker Diarization problem which is in some ways is similar to segmenting speech into acoustical units (Fox et al., 2011). In this project we also used HDP-HMM. Each segment is molded by a single state of the HMM but since the number of states in unbounded the model can automatically find the appropriate number of states without using any kind of heuristic or model comparison techniques.

In this project we have used a small dataset extract from SA part of TIMIT data. Two speakers FALK0 and FCJF0 with different accents have been selected. Our results show HDP-HMM can consistently segment speech utterances. Despite of very different accents, the algorithm can find somewhat similar segments for words spoken by different speakers. We also find that if we use a Dirichlet process mixture (DPM) to model the observation of each state the results become more consistent. However, in this case the interpretation is somewhat more difficult since each segment can potentially model different acoustical events. Moreover the clustering becomes a more difficult problem. One solution to this problem is to label and define segments based on both state and mixture numbers. This will make the final segments more homogenous and well defined.

The next step, after segmentation, is to cluster the segments and also to generate a dictionary. Generating the dictionary in case of TIMIT dataset is relatively simple because the dataset is manually transcribed and we can simply align the final string of acoustical units with each word and generate an entry for our dictionary. For the clustering step we propose to use another nonparametric Bayesian model named DPM. This technique which has been used to model the output distribution of states in this project has been extensively used in clustering problems (Kurihara, Welling and Vlassis, 2006).

Table of Contents

[1. Introduction 1](#_Toc330501076)

[2. HDP-HMM 3](#_Toc330501077)

[2.1. Inference and Learning 4](#_Toc330501078)

[3. Data 4](#_Toc330501079)

[4. Experiments 5](#_Toc330501080)

[4.1. Experiment parameters 5](#_Toc330501081)

[4.2. Experiment 0 (exp000) 5](#_Toc330501082)

[4.3. Experiment 1(exp001) 7](#_Toc330501083)

[4.4. Experiment 2 (exp002) 8](#_Toc330501084)

[4.5. Experiment 3 (exp015) 8](#_Toc330501085)

[4.6. Experiment 4 (exp010) 10](#_Toc330501086)

[5. Discussion 11](#_Toc330501087)

[6. Reference 11](#_Toc330501088)

# Introduction

Automatic discovery of acoustical units is of some interest in speech recognition applications. Most algorithms approach this problem in two steps (Paliwal, 1990), (Bacchiani et al, 1996), (Bacchiani and Ostendorf, 1999) and (Goussard and Niesler, 2010):

1. segmentation
2. clustering

Usually segmentation is accomplished using a heuristic method (e.g. energy changes of the spectrum). After the initial segmentation, similar segments are clustered in several groups using a heuristic clustering algorithm (e.g. tree). Each group represents an acoustical unit. Another important part of all algorithms is to assign a string of these acoustical units to each word in the transcription and generate a dictionary. It is expected that acoustical units and dictionary generated directly from speech works better than traditional linguistic based units (e.g. phonemes); moreover, for many languages there is not enough linguistic or even written resources and therefore automatically generated units are the only available option..

In this pilot project, the goal is to investigate the first step by a simplified dataset using HDP-HMM approach. Unlike most other works, our approach is not heuristic and has a very deep mathematical background. In this approach we use HMMs to segment an utterance into several homogenous regions. HMMs provide a very powerful toolkit for segmentation and their wide spread applications is usually because of this power. However, the number of states should be known in advance. Since the number of segments (acoustical units) is not known a priori, we cannot use HMMs directly.

One solution is to use many different HMMs with different number of states and then select the best using model selection techniques. However, model selection is computationally expensive and moreover, the criteria used for selecting the best model is usually a heuristic one. Fortunately, another approach recently proposed based on nonparametric Bayesian statistics (Teh et al, 2006), (Teh and Jordan, 2010) and (Fox et al., 2011). This approach is named infinite HMM or HDP-HMM and provide a variant of HMM with unbounded number of states. The basic idea for segmenting an utterance into acoustical units is similar to speaker diarization problem (Fox et al., 2011). In that problem, the goal is to segment an audio file into speaker homogenous areas. One of the recent and successful approaches toward that problem was to use an HMM with unbounded number of states (Fox et al., 2011). In that application, each state represents a speaker, and in theory (and also experiment) the model successfully can capture speakers without knowing the number of speakers a priori.

The current problem is very similar to the speaker diarization problem and the only difference is to substitute speakers with acoustical units. One important assumption is to assume that acoustical units consist of relatively homogenous regions. This assumption automatically encourages the using of more units since each unit is relatively simple and cannot be accounted for complex acoustical events individually.

In the subsequent sections, we show that HDP-HMM can handle the segmentation problem and despite of different accents can find similar results (to some extend) for words spoken by different speakers. We also show that in order to obtain consistent and meaningful results we should assume each unit is relatively homogenous. Figure 1 shows the result of segmentation using HDP-HMM for SA1 and speaker FALK0 and with minimum segment length equal to 30 msec. As it can be deduced from this figure the findings of the algorithm is consistent with the intuition. It should be noted that these segments are not the final acoustical units and might be regrouped after the clustering stage.

The next step for generating acoustical unit from data is to cluster segments (along with generating a dictionary). For a moment, we can assume that dictionary can be generated separately. The clustering problem can be accomplished using any clustering method. In fact, many people have used K-means or trees (with some variations and heuristics) for clustering (Paliwal, 1990), (Bacchiani et al, 1996), and (Goussard and Niesler, 2010).

In this project, we are not concerned with the clustering problem but it seems natural to use a nonparametric Bayesian approach for clustering too. A Dirichlet process mixture (DPM) has been used extensively for clustering problems (Kurihara, Welling and Vlassis, 2006) and (Harati, Picone and Sobel, 2012). We propose to use HDP-HMM for segmentation and DPM for clustering in a future project. The final system will be a completely nonparametric Bayesian solution to the problem of acoustical unit discovery.

This report is organized in five sections. In section ‎2 we review HDP-HMM without going into details. In section ‎3 dataset is introduced. In section ‎4, experimental results will be presented, and finally in section‎ ‎5 the results are summarized and a brief discussion will be presented.



Figure 1-Segmenting SA1 (FALK0) using HDP-HMM with minimum segment length of 30 msec.

# HDP-HMM

In this section, we review the definition of HDP-HMM; however, the backgrounds and details are not discussed in this report. Interested reader can refer to (Teh and Jordan, 2010) and (Fox et al., 2011) ; for a more compact review refer to (Harati, 2012).

Hidden Markov models (HMMs) are a class of doubly stochastic processes in which discrete state sequences are modeled as a Markov chain (Rabiner, 1989). In the following discussion we will denote the state of the Markov chain at time  with  and the state-specific transition distribution for stateby.The Markovian structure means. Observations are conditionally independent given the state of the HMM and are denoted by.

HDP-HMM is an extension of HMM in which the number of states can be infinite. The idea is relatively simple; at each statewe should be able to go to an infinite number of states so the transition distribution should be a draw from a Dirichlet Process (DP). On the other hand, we want reachable states from one state to be shared among all states so these DPs should be linked together. The result is an Hierarchical Dirichlet Process (HDP). In an HDP-HMM each state corresponds to a group and therefore, unlike HDP in which an association of data to groups is assumed to be known a priori, we are interested to infer this association. The major problem with original HDP-HMM is the state persistence. HDP-HMM has a tendency to make many redundant states and switch rapidly among them (Fox et al., 2011). This problem is solved by introducing a sticky parameter to the definition of HDP-HMM (Fox et al., 2011). Equation shows the definition of a sticky HDP-HMM with unimodal emissions.is a sticky hyper-parameter and generally can be learned from data. Original HDP-HMM is a special case with. From this equation we can see for each state (group) we have a simple unimodal emission distribution. This limitation can be addressed using a more general model defined in . In this model, a DP is associated with each state and a model with augmented stateis obtained. Figure 2 shows a graphical representation.

 

 



Figure 2- Graphical model of HDP-HMM (Fox et al.m 2011)

#

## Inference and Learning

Inference can be implemented in several ways. In this project we have used a block sampler which proposed by (Fox et al., 2011). This sampler uses Markovian structure of the model to improve its performance. In this algorithm a fixed truncation level Kz is used for the number of states. However, it should be noted that the result is different from a classical parametric Bayesian HMM since the truncated HDP priors induce a shared sparse subset of the Kz possible states (Fox et al. , 2011). In short, we obtain an approximation to the nonparametric Bayesian HDP-HMM with maximum number of possible states set to Kz. However, for almost all applications this should not cause any problem if we set reasonably high. Similarly, a fixed truncation level Ks is used for number of mixtures.

# Data

For this pilot project, the data is extracted from TIMIT database (Zue et al., 1989) and consists of SA1 sentence for 2 speakers: FALK0 and FCJF0.

The reason to keep the dataset small was to investigate the HDP-HMM segmentation approach quickly before using this approach for a more challenging dataset. The existence of the manual transcription for TIMIT was another important factor. This helps us to align the segments with words rather easily. In other words, dictionary can be generated with just aligning. For other datasets we have to use a forced-alignment procedure to do the same thing.

# Experiments

In this section, some of the important results are presented. We have used Matlab to run these experiments.

First each utterance is converted into MFCC features (window length=25 msec, frame rate=10 msc), then frames converted into block of size L (without overlap) by averaging. The result is fed into HDP-HMM and segmented into several states. Each state is a segment of the input utterance and presents a potential acoustic unit.

## Experiment parameters

There are several parameters that we change through these experiments.

* block\_size : number of frames that make a block multiplied by frame rate.
* Prior type: conjugate /non-conjugate.
* Kz: truncation level for number of states.
* Ks: truncation level for number of mixtures.

For further discussion about parameters refer to (Fox et al., 2011) and (Fox et al,2010).

## Experiment 0 (exp000)

***Parameters:*** block\_size=10 msec, conjugate priors, Kz=100,Ks=1

In this experiment Ks is set to one which enforce each state (segment) to be modeled by a Gaussian. Kz is set to 100 which allows to at most 100 segments. Totally 41 unique segments has been identified. A dictionary based on these segments is as Table 1 (labels are arbitrary):

From this table, we can see there is a general similarity (in sense of segment sequence for similar words spoken by different speakers.)

It can be seen that some of the segments are speaker specific. For example for word “all” we have:

all : 60 60 60 60 60 60 60 60 60 60 60 54 54 54 80 80 41 41 41 41

all : 29 29 29 29 29 29 29 29 29 29 29 29 54 80 80 41 41

in this case we can see segment 29 and 60 are speaker specific in each case and all other segments are identical. However, the mean of the distribution that represent segments 29 and 60 are relatively close. The normalized distance between their “mean”s is 11.64 ( average distance between two arbitrary segments is 41.13 ) and the correlation between the two means is 80.15% (the average absolute correlation is 28.77%.) In other words, segments 60 and 29 represents different flavor of the same underlying segment. Depending on amount of available data, these two segments might merged or remain separate in a later clustering stage. Other examples (among others) are segments 2-67 and 7-40 in word “greasy”. In some cases, it seems a segments split into two. For example, word “in” (speaker FALK0) the distance among segments 33 and 41 is 9.67 and the distance between segments 45 and 94 is 11.14. This splitting seems to be related to the particular manner of articulation (Again depending on amount of data might merge in later clustering stage.) Furthermore, it can be seen that consistency is generally existed. For example segment 29 (and 60) approximately coincides with phoneme “aa”. Examples are words “wash”, “water” and “all”.

Another interesting thing is the fact that some segments are relatively short (shorter than 30 msec which is the minimum length of many state of the art system for different models.) This suggests that using an appropriate variable length models might be helpful in modeling acoustic units.

Table 1-Mapping Segments to Words for Experiment 0

|  |  |
| --- | --- |
| Speaker |  |
| FALK0 | sil : 81 81 81 81 81 81 81 81 81 81 81 81 81 81she : 81 81 81 81 81 81 81 81 81 81 2 7 7 41 41 41 41 41 41 41had : 45 45 45 45 45 45 62 62 62 62 62 62 62 41 41 99 99 4 4 17 82 82 33your : 33 79 79 79 79 79 79 41 41 94dark : 94 94 94 94 70 70 33 32 32 20 20 20 20 20 20 20 20 20 20 20 20 20 30 30 99 100 100 22 81 81 81 81suit : 81 81 81 81 81 81 81 81 81 81 2 2 40 41 41 41 41 41 41 41 41 41 41 94 94 70 70in : 33 33 33 33 33 41 41 41 41 41 45 45 45 45 45 45 94greasy : 94 94 98 98 65 45 45 45 48 48 48 41 41 41 41 41 41 41 41 41 41 94 94 94 94 81 81 81 2 2 7 7 7 41 41 41 41 41 41wash : 45 45 45 45 45 45 25 25 25 25 25 25 29 29 29 54 54 59 59 59 30 30 94 81 81 81 81 81 81 81 81 81water : 25 29 29 29 29 54 54 54 59 59 59 59 28 71 71 72 72 72 72 72 72 72 72 98all : 60 60 60 60 60 60 60 60 60 60 60 54 54 54 80 80 41 41 41 41year : 41 41 41 74 74 74 74 74 74 74 74 74 74 74 74 74 74 74 74 79 79 89 89 71 72 72 72 76 76 76 76 76 76 76 76 83sil : 83 83 83 83 100 100 100 100 100 100 100 100 100 100 100 100  |
| FCJF0 | sil : 81 81 81 81 81 81 81 81 81 81 17 81 81 81 81 81 81 81 27she : 27 27 27 27 27 27 27 67 67 40 41 41 41 41 41 41 68had : 68 45 45 45 45 62 62 62 62 62 62 62 62 62 62 62 41 41 68 68 4 4 27 27 27 27 67 67 40your : 4 27 27 27 27 67 67 40 41 41 41 41 41 41 41dark : 99 99 4 4 70 70 41 41 73 73 73 73 73 73 73 73 73 73 73 68 68 100 100 17 17 17 17 17 81 81suit : 81 81 81 81 81 81 81 81 67 67 40 40 41 41 41 41 41 41 41 41 41 41 41 41 41 41 41 41 41 41in : 41 41 41 41 41 41 41 94 94greasy : 65 65 65 45 45 45 45 45 41 41 41 41 41 41 41 41 41 94 94 81 81 81 81 81 81 81 67 67 40 41 41 41 41 41wash : 41 45 45 45 25 25 25 25 25 29 29 29 29 29 29 29 54 54 54 73 73 68 68 4 27 27 27 27 27 27 81 17 17 17water : 29 29 29 29 29 29 29 29 29 54 54 28 28 28 28 28 28 28 28 98 98 98 98 98all : 29 29 29 29 29 29 29 29 29 29 29 29 54 80 80 41 41year : 41 41 74 74 74 74 74 74 74 74 74 74 89 89 89 89 89 89 89 71 71 71 71 71 71 71 71sil : 71 8 8 8 8 8 8 8 8 8 8 |

## Experiment 1(exp001)

***Parameters:*** block\_size=20 msec, conjugate priors, Kz=100,Ks=1

In this experiment the block\_size (minimum segment length) is 20 msec. Total number of unique segments is 33. Due to poorer resolution the boundaries in this case are less reliable. (To map words into segments we have used manually transcribed transcriptions however when increasing the block size, some blocks can occur in the boundary and so we need to find a better procedure to do this). Despite this from Table 2 we can see several interesting phenomenon:

Table 2- Mapping Segments to Words for Experiment 1

|  |  |
| --- | --- |
| speaker |  |
| FALK0 | sil : 37 37 37 37 37 37 37she : 60 60 60 60 60 18 79 70 70 70had : 52 52 52 90 90 90 90 23 23 54 18your : 79 25 25 25 70dark : 23 23 54 18 16 88 88 88 88 88 88 88 19 34 95 61suit : 55 55 55 55 55 55 18 68 68 68 68 12 34 54in : 31 31 31 70 70 23 23 23greasy : 23 54 63 40 40 30 25 70 70 70 70 52 55 55 55 18 79 79 79 12wash : 75 75 75 75 75 75 10 10 51 51 91 52 60 60 60 61water : 10 10 10 51 51 51 3 3 99 99 99 99all : 10 10 10 10 10 10 10 51 70 70year : 70 70 70 70 70 70 70 70 70 70 70 48 48 99 99 99 99 99 87sil : 87 87 37 37 37 37 37 37 |
| FCJF0 | sil : 37 37 37 37 37 37 37 37 37 60she : 60 60 60 18 70 70 70 70had : 52 18 16 16 16 16 16 16 70 19 34 54 60 18your : 54 60 18 70 70 70 12dark : 19 34 54 18 91 91 91 91 91 19 34 95 95 61 61suit : 61 61 61 61 61 18 70 70 70 70 70 70 70 70 70in : 70 70 23 23 54greasy : 61 18 18 18 70 70 70 70 52 61 61 61 61 18 79 12 12wash : 75 75 75 75 10 10 10 10 51 91 19 54 60 60 60 61water : 10 10 10 10 10 51 3 3 3 3 3 3all : 10 10 10 10 10 10 10 51 70year : 70 70 70 70 70 70 70 70 48 48 48 48 48sil : 48 4 4 4 4 4 |

1. It seems segments follows an n-gram statistical structure. For example, for example segment 79 always come after segment 18; Or segment 12 comes after 70,79 and 68 but 70, 79 and 68 are basically very similar (distance between 70 and 79 is 4.09 and between 68 and 70 is 12.04; the average distance is 37.33).
2. “ in greasy” shows the mapping problem in boundary areas. It seems segments “23 54” (FALK0 for word “greasy”) actually belong s to word “in”. (further analysis shows that 23 belonged to “in” and “54 belonged to “greasy”)
3. Some segments look to behave like outliers (for example segment 63 though its distance to segments 18 and 40 is small)
4. Dialect and style differences can be seen clearly in words “year” and “water” (it can also be seen for exp000.) Speaker FCJF0 clearly does not pronounced “r” in both cases (at least by listening you can barely hear the “r” while it probably pronounced very quickly.)
5. By comparing “wash” and “water” we can see in this experiment segment 75 which is closely related to “w” is not presented for word “wash” for both speakers. However for exp000 segments 45 and 25 (distance 11) are presented for “wash” and “water” for speaker FALK0 while for speaker FCJF0 segment 45 only exists for word “wash”. By listening to the audio file it seems that FCJF0 skip “w” for “water”. The fact that, exp000 can detect “w” for speaker FALK0 and for word “water”, and exp001 cannot do the same, shows that increasing the minimum length of segments can reduce the accuracy of the mapping.

## Experiment 2 (exp002)

***Parameters:*** block\_size=30 msec, conjugate priors, Kz=100,Ks=1

In this experiment we have used, block\_size=30 msec . Table 3 shows a major problem: word “water” for both speakers is represented with just one segment (42) which also used to represent other words too. This basically makes the result useless.

Table 3- Mapping Segments to Words for Experiment 2

|  |  |
| --- | --- |
| speakers |  |
| FALK0 | sil : 47 47 47 47 47she : 22 22 22 69 46 46had : 39 39 42 42 42 39 25 89your : 46 88 65dark : 39 25 42 42 42 42 42 42 40 47 87suit : 87 87 87 33 69 46 46 39 25in : 69 69 46 39 39 39greasy : 42 42 42 88 46 46 46 96 87 33 46 46 42wash : 42 42 42 42 42 42 42 96 22 59water : 42 42 42 42 42 42 42 42all : 42 42 42 42 42 42 46year : 42 46 46 46 46 46 46 31 42 42 42 42sil : 47 47 47 47 47 |
| FCJF0 | sil : 47 47 47 47 47 47she : 22 22 45 69 46 46had : 39 42 42 42 42 42 40 22 45 69your : 22 45 69 42 42dark : 40 25 42 67 67 67 40 47 47 59suit : 59 59 59 69 42 42 42 42 42 42in : 46 39 39greasy : 42 42 42 42 46 46 59 59 45 69 46wash : 42 42 42 42 42 42 42 40 22 22 59 45water : 42 42 42 42 42 42 42 42all : 42 42 42 42 42 42year : 46 46 46 46 46 31 31 31 31sil : 40 47 47 47 |

## Experiment 3 (exp015)

***Parameters:*** block\_size=10 msec, conjugate priors, Kz=100,Ks=4

This experiment is similar to exp000 but Ks is set to 4. This allows the algorithm to model each segment with a mixture of Gaussians (up to 4 components.) Letting each segment to be molded by more than one Gaussian means each segment can potentially models more than one acoustical event. It causes segments representing similar acoustical events to be different from each other. Meaning for example, a segment can represent using one Gaussian while another similar segment can be represented with two Gaussians (modes). In this case, the distance between two segments might be large. Because the distance between one of the modes of the second segment might be large to the only mode of the first segment while the other mode has a smaller distance. For example, segments 25 and 47 represent a similar event but their distance is large (32.5). Segment 47 is modeled using a single Gaussian and segment 25 using 2 Gaussians. The distance between the first mode of 25 and 47 is 10.2 and the distance between the second mode of 25 and 47 is 65.9 and since the probability of both modes is almost equal the total distance is around 32.5.

From Table 4 and by comparing to Table 1 it seems the results are more consistent, however the interpretation is more difficult.

Table 4- Mapping Segments to Words for Experiment 3

|  |  |
| --- | --- |
| speakers |  |
| FALK0 | sil : 50 50 50 50 50 50 50 50 22 22 22 22 100 35she : 35 35 35 35 35 35 35 35 35 35 75 75 43 43 43 43 43 89 89 89had : 4 4 4 4 4 4 4 40 40 40 40 40 2 2 2 73 73 73 35 21 59 59 72your : 72 91 91 91 91 91 91 2 2 45dark : 45 25 25 35 11 11 91 91 40 7 7 7 7 7 7 7 7 7 7 7 7 8 8 8 13 96 96 5 16 16 76 76suit : 76 76 76 76 76 76 76 76 76 3 3 87 76 67 26 26 26 26 26 26 26 84 84 13 35 11 11in : 59 58 58 58 58 49 49 49 89 89 67 67 67 67 67 67 25greasy : 25 35 16 16 71 71 71 71 71 2 2 67 67 67 43 43 43 43 43 89 89 51 51 51 51 51 3 3 3 75 75 43 43 43 43 43 26 26 70wash : 70 70 29 29 29 29 29 29 48 48 48 48 47 47 47 88 88 88 88 7 7 100 100 35 35 35 35 35 35 35 41 41water : 48 47 47 47 47 88 88 88 88 88 88 88 73 73 50 50 50 57 57 57 57 57 57 45all : 25 25 25 25 25 25 25 25 25 25 25 87 87 87 7 7 43 43 43 43year : 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 31 31 31 23 23 23 18 18 18 18 18 18 18 18 18 13sil : 13 13 13 12 12 12 12 12 12 12 12 12 12 12 12 50 |
| FCJF0 | sil : 100 100 100 100 100 50 50 22 100 100 100 100 100 100 100 100 100 100 35she : 35 35 35 35 35 35 35 35 76 76 67 43 43 43 89 89 89had : 4 4 4 4 4 40 40 40 40 40 40 40 40 40 40 75 75 75 15 15 31 31 6 6 6 35 35 76 76your : 31 6 6 6 35 35 76 76 42 42 42 42 42 84 84dark : 13 13 31 31 11 11 76 42 42 40 40 40 40 40 40 40 40 7 7 90 90 12 12 16 16 16 16 16 35 35suit : 35 35 35 35 35 35 35 35 41 41 76 76 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 42 43in : 43 43 86 86 86 86 86 86 86greasy : 86 59 59 59 59 59 59 59 59 67 67 67 67 67 67 89 89 52 52 35 35 35 35 35 35 35 35 75 76 67 67 42 42 70wash : 70 70 70 70 48 48 48 48 48 47 47 47 47 47 47 47 88 88 88 7 7 15 15 15 6 6 35 35 35 35 35 41 41 41water : 47 47 47 47 47 47 47 47 47 88 88 39 39 39 39 39 39 39 39 39 39 39 39 47all : 47 47 47 47 47 47 47 47 47 47 47 47 30 30 30 43 43year : 43 43 43 43 43 43 43 43 43 43 43 43 76 76 76 76 76 76 13 13 13 13 13 13 13 13 13sil : 13 13 90 90 90 90 90 90 90 90 90 |

One problem is that each segment potentially can model 2 completely different acoustic events. Another problem is related to the clustering step. It is not clear how one can cluster segments with different number of Gaussians. This second problem is a serious practical problem and should be addressed before proceeding to clustering.

One possible solution to above problem is to label segments by both the segment number and mixture number. For example, if segment 25 has 2 mixtures then it actually represents two segments: 25-1 and 25-2. In this way, each segment will represent a more homogenous region and therefore model a simpler acoustical effect. Moreover, the clustering step would be rather simple and straightforward. In the later, clustering stage these segments might merge again and represent one acoustical unit.

## Experiment 4 (exp010)

***Parameters:*** block\_size=30 msec, conjugate priors, Kz=100,Ks=10

This experiment is similar to exp002 but Ks is set into 10. The resulted dictionary is shown in Table 5. In comparison, we can say the results are more consistent but have the same problems of pervious experiment. Interestingly, unlike exp002, none of the words are represented by just one segments. This suggests allowing segments to be represented by mixture of Gaussians (instead of single Gaussian) can improve the accuracy of the mapping. Figure 1 shows the result of segmentation for this case.

Table 5- Mapping Segments to Words for Experiment 4

|  |  |
| --- | --- |
| speakers |  |
| FALK0 | sil : 45 45 45 68 68she : 24 24 24 6 86 86had : 35 22 22 84 84 35 83 83your : 57 57 80dark : 35 73 52 54 54 54 54 54 75 73 86suit : 4 4 4 29 11 11 11 35 73in : 47 47 86 35 68 68greasy : 82 78 76 46 30 30 45 4 4 99 35 68 82wash : 43 43 43 43 26 30 73 24 24 24water : 43 26 30 30 50 50 69 69all : 26 26 26 26 30 69 55year : 55 55 55 55 55 55 57 50 69 69 69 7sil : 7 73 73 73 73 |
| FCJF0 | sil : 68 68 68 68 68 68she : 17 17 38 6 30 58had : 22 22 84 84 84 69 96 9 38 70your : 9 38 70 69 69dark : 46 85 70 84 84 84 76 73 78 59suit : 59 59 59 70 69 5 5 5 5 99in : 30 45 45greasy : 7 7 7 69 69 49 59 59 38 6 5wash : 5 43 43 43 26 26 30 76 10 17 59 78water : 26 26 26 50 50 50 80 80all : 26 26 26 26 26 69year : 30 55 55 55 57 56 56 56 56sil : 73 73 73 73 |

# Discussion

In this project, we investigated the usage of HDP-HMM model for segmenting speech utterances into homogenous sections which can possibly be used as a first step of a nonparametric Bayesian approach for automatic acoustical unit discovery algorithm. It has been shown that HDP-HMM can produce meaning full and consistent results. From experiments, we can claim allowing each state of HDP-HMM to have multiple mixtures (instead of one Gaussian) can improve the consistency of the results. However this can complicate the clustering step. One solution to this problem is to define (and label) segments using both state number and mixture number. In this way, segments remain simple (presented using a Gaussian distribution) and at the same time the segmentation remains consistent and more reliable.

Another problem that is omitted in this project is related to processing large datasets. In this case, each utterance should be segmented separately. However, labels for each utterance would be completely arbitrary. In other words, we should rely on the clustering algorithm to merge identical (but labeled differently) segments together. An alternative solution is to present all utterances to the segmentation algorithm; however, the direct implementation of this solution is slow and cannot be parallelized.

# Reference

Bacchiani, M., & Ostendorf, M. (1999). Joint lexicon, acoustic unit inventory and model design. *Speech Communication*, *29*(2-4), 99–114. doi:10.1016/S0167-6393(99)00033-3

Bacchiani, M., Ostendorf, M., Sagisaka, Y., & Paliwal, K. (1996). Design of a speech recognition system based on acoustically derived segmental units. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 443- 446). Atlanta, GA, USA. doi:10.1109/ICASSP.1996.541128

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2010). Supplement to “ A Sticky HDP-HMM with Application to Speaker Diarization”. *The Annals of Applied Statistics*, *S*(2A), S1-S32. . doi:10.1214/10-AOAS395SUPP

Fox, E., Sudderth, E., Jordan, M., & Willsky, A. (2011). A Sticky HDP-HMM with Application to Speaker Diarization. *The Annalas of Applied Statistics*, *5*(2A), 1020-1056. doi:10.1214/10-AOAS395

Goussard, G., & Niesler, T. (2010). Automatic discovery of subword units and pronunciations for automatic speech recognition using TIMIT. *Proceedings of the twenty-first annual symposium of the Pattern Recognition Association of South Africa* (pp. 93-99). doi:missing

Harati, A. (2012). Hierarchical Dirichlet Processes and Infinite HMMs. *PhD Preliminary Exam, Department of Electrical and Computer Engineering, Temple University*. Philadelphia, Pennsylvania, USA. doi:http://www.isip.piconepress.com/publications/presentations\_misc/2012/phd\_prelim/dpm/

Harati, A., Picone, J., & Sobel, M. (2012). Applications of Dirichlet Process Mixtures to Speaker Adaptation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 4321-4324). Kyoto, Japan. doi:TBD

Kurihara, K., Welling, M., & Vlassis, N. (2006). Accelerated Variational Dirichlet Process Mixtures. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems* (pp. 761-768). MIT Press. doi:missing

Paliwal, K. (1990). Lexicon-building methods for an acoustic sub-word based speech recognizer. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 729- 732). Albuquerque, New Mexico, USA. doi:10.1109/ICASSP.1990.115888

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 879-893. doi:10.1109/5.18626

Teh, Y., & Jordan, M. (2010). Hierarchical Bayesian Nonparametric Models with Applications. In S. W. Hjort, C. Holmes, P. Mueller (Ed.), *Bayesian Nonparametrics: Principles and Practice* (pp. 158-207). Cambridge-UK: Cambridge University Press.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, *101*(47), 1566-1581. doi:10.1198/016214506000000302

Zue, V., Glass, J., Phillips, M., & Seneff, S. (1989). Acoustic Segmentation and Phonetic Classification in the SUMMIT System. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 389- 392). Glasgow, Scotland. doi:10.1109/ICASSP.1989.266447