A Comparative Analysis of Bayesian Nonparametric Inference Algorithms for Acoustic Modeling in English and Mandarin Speech Recognition

John Steinberg, Joseph Picone 1

1 Department of Electrical and Computer Engineering, Temple University, Philadelphia, PA

[john.steinberg@temple.edu](mailto:john.steinberg@temple.edu), joseph.picone@isip.piconepress.com

Abstract

Nonparametric Bayesian models have become increasingly popular in speech recognition for their ability to discover data’s underlying structure in an iterative manner. Dirichlet process mixtures (DPMs) are a widely used nonparametric method that do not require a priori assumptions about the structure of the data, e.g. the number of mixture components, and can learn this directly from the data. DPMs, however, require an infinite number of parameters so inference algorithms are needed to make posterior calculations tractable. The focus of this work is an evaluation of three variational inference algorithms for acoustic modeling: Accelerated Variational Dirichlet Process Mixtures (AVDPM), Collapsed Variational Stick Breaking (CVSB), and Collapsed Dirichlet Priors (CDP) [1][2].

A phoneme recognition task is chosen to more clearly assess the viability of these algorithms for acoustic modeling. Evaluation is conducted on the CALLHOME Mandarin and English corpora, consisting of two languages that, from a human perspective, are phonologically very different. It is shown in this work that AVDPM, CVSB, and CDP yield comparable results to a baseline Gaussian mixture model (GMM) but with far fewer mixture components. Furthermore, the disparity in error between the languages is small enough to be attributed to Mandarin’s relatively large number of phoneme labels.

**Index Terms**: nonparametric Bayesian, variational inference, CALLHOME, Mandarin, English, phoneme recognition

# Introduction

Nonparametric Bayesian models have become increasingly popular in speech recognition tasks due to their ability to discover data’s underlying structure in an iterative manner. Dirichlet process mixtures (DPMs) are a widely used nonparametric method that do not require a priori assumptions about the structure of data, such as the number of mixture components, and can learn this information directly from the data itself. This is ideal for acoustic modeling in speech recognition where the number of mixture components is a parameter commonly found by tuning a system using a subset of the data. Typically, the number of components is assumed to be constant since it would be tedious and time consuming to tune models for each phoneme. DPMs, however, are able to automatically determine an optimal number of mixtures for each individual model.

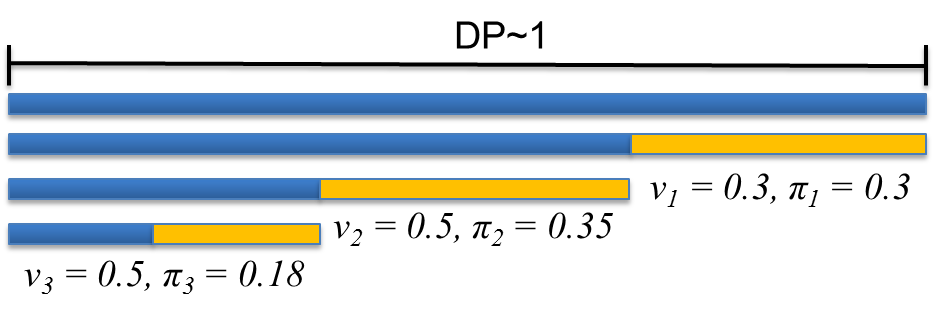


Figure 1: *A diagram showing the stick-breaking representation of the Dirichlet process. Stick breaks represent mixture components and the absolute length of each piece, πi , represents component weights.*

There are many depictions of Dirichlet processes but the algorithms in this work are all premised on the stick-breaking approach shown in Figure 1. In this representation a stick of uniform length is broken repeatedly. Each break represents a new mixture component where the fraction of the remaining stick is given by *vi* and the absolute length of each piece (i.e. the weight of the mixture component) is given by *πi*.

Aside from the automatic tuning of the number of mixtures, it is equally important to ensure that these models generalize well across different data. With the rapid expansion of globalization and social media, it is especially important that speech recognition technology be adaptable. More specifically acoustic models should not be severely influenced by the language of the training data. In this work, the performance of three Bayesian variational inference algorithms – accelerated variational Dirichlet process mixtures (AVDPM), collapsed variational stick-breaking (CVSB), and collapsed Dirichlet priors (CDP) - are assessed for both the CALLHOME English (CH-E) and CALLHOME Mandarin (CH-M) corpora.

## Variational Inference Algorithms

Nonparametric methods such as DPMs, although extremely useful for finding the underlying structure of data, often come at a cost of computational complexity. Contrary to the misleading phrase, “nonparametric”, DPMs require a potentially infinite number of parameters. This makes manipulating such distributions intractable so inference algorithms are used to approximate these models. Markov chain Monte Carlo (MCMC) methods, such as Gibbs sampling, are extremely popular for their mathematical simplicity [3]. By analyzing samples, MCMC methods can approximate complex distributions. Unfortunately, these techniques are sensitive to step size and can require careful control of the “burn-in” phase. More importantly, using MCMC as an inference algorithm requires an increasingly large number of samples as the target posterior distribution becomes increasingly complex. Thus, these methods are sometimes poorly equipped to work with DPM models.

Instead of sampling methods like MCMC, variational inference algorithms approximate a posterior, *p(y|x)*, with a simpler distribution *q(y)* by making assumptions about the independencies of the distribution’s latent variables. The task of approximating a complex distribution is transformed into an optimization problem where an optimal *q* is found from a set of variational distributions *Q={q1*, *q2*,…, *qm*} such that an objective function, i.e. Kullbach-Liebler divergence, is minimized.

Until recently, however, even variational inference methods were not computationally viable for speech recognition tasks. In this work three new variational inference algorithms – AVDPM, CVSB, and CDP [1][2] – were selected for their reasonable computational complexity to investigate their performance on an acoustic modeling task in speech recognition.



Figure 2: *A diagram showing how splitting a Dirichlet distribution infinitely many times yields discrete values.*

## English and Mandarin

As of 2009 Ethnologue reported 6,909 living languages in the world and of those Mandarin and English are numbers one and three (respectively) of the most commonly spoken [4]. Moreover, these two languages come from separate families and are linguistically and phonetically very different. For these reasons English and Mandarin are selected to ensure that the performance of AVDPM, CVSB, and CDP are not heavily influenced by any language specific artifacts.

Based on NIST benchmarks Mandarin speech recognition tasks have historically yielded worse error rates than comparable English ones [6]. There are many factors that this disparity can be attributed to such as Mandarin’s flexible grammatical structure, relatively high number of homophones – Mandarin has ~1,300 syllables as opposed to English’s ~10,000 – and, most conspicuously, the tonal nature of the language. Unlike English, whose phoneme labels are all unique, each vowel in Mandarin can take five different tones (4 distinct tones and 1 neutral tone). Thus, where English has approximately 40 phoneme labels, Mandarin actually has near 90. The scope of this work is constrained to phoneme recognition so that other factors, such as from language models, can safely be ignored.

# Background Theory

Parameterized models have been widely applied to clustering and classification problems for their ease of use, simplicity, and reasonable performance. Unfortunately, they require making assumptions about data structure and sometimes generalize poorly. Nonparametric methods, on the other hand, do not suffer from these limitations but, due to their complex nature, require inference algorithms to make posterior calculations tractable. In this section, a brief overview of one such nonparametric method, Dirichlet process mixtures, and three variational inference algorithms –AVDPM, CVSB, and CDP – are offered.

## Dirichlet Distributions and Dirichlet Processes

One of the main drawbacks of typical, parametric speech recognition systems is the assumption that the number of mixture components for each phoneme model is known and is held constant for every model. For complex data such as speech this is largely presumptuous and it would be more reasonable to assume that each phoneme model has its own unique structure.

Creating a model to characterize the optimal number of mixture components is best represented by a multinomial distribution. To model this in a statistically meaningful way priors are needed to ascertain information such as the number of mixture components and their respective weights. Dirichlet distributions act as the conjugate prior for the multinomial distribution, and in the case of this work, can be used to find the optimal number of mixture components. An extension of the Dirichlet distribution, the Dirichlet process, is used to then generate discrete priors for modeling the respective weights of these components. The next few paragraphs highlight the relationship between the Dirichlet distribution and Dirichlet process.

A Dirichlet distribution is often referred to as a distribution of distributions and is given by:



where *q* and α are a set of distributions and their respective concentration parameters (i.e. inverse variances) such that , ,  and, , and . Furthermore, the decimative property of Dirichlet distributions explains that each distribution, *qi*, can be split in such a way that:



A Dirichlet process is a Dirichlet distribution split infinitely many times, ultimately generating discrete values that serve as priors. This can be seen in Figure 4 where a Dirichlet distribution is initially set to a uniform distribution. After an infinite number of splits, the resulting distributions are infinitely narrow and essentially discrete values are obtained which serve as priors for the models in this work.

Although there are many representations of Dirichlet processes, all three algorithms used in this work focus on the stick breaking approach. In this representation a Dirichlet process can be interpreted as a stick of length 1 (i.e. the probability of the data is 1) that is split infinitely many times. For this work each break in the stick represents a new mixture component in the phoneme’s acoustic model. The length of each piece represents the component’s weight.

## Variational Inference Algorithms

Dirichlet processes use an infinite number of parameters which make posterior calculations intractable. The solution to this problem is to use inference algorithms which approximate these complex models by using simpler distributions. Markov chain Monte Carlo (MCMC) methods such as Gibbs sampling are a commonly used form of inference algorithm [3][7][8]. Although mathematically fairly simple, these methods can be sensitive to tuning and can require large numbers of samples for increasingly complex distributions.

Variational inference converts the sampling problem of MCMC methods into an optimization problem. A variational distribution, *q(y)*, which has made independence assumptions about model parameters, is used to approximate the posterior, *p(y|x)*. More specifically, these algorithms assume that the distributions that represent stick lengths (and by extension, mixture component weights), component structure (i.e. means and covariances of a Gaussian for this work), and mixture assignments are all independent. This relationship can be seen in , , and below. By using optimization techniques such as the EM algorithm and the Kullbach-Liebler (KL) divergence as a cost function, an optimal *q(y)* can be found from a set of distributions *Q = {q1, q2,… ,qk}*. Thus, new mixture components are released as the KL divergence is minimized.

Even variational inference algorithms can be computationally inefficient and often require additional constraints to make their use viable. AVDPM incorporates KD-trees which can be used during preprocessing to organize the data by partitioning them across hyperplanes in the feature vectors. This essentially allows for the tradeoff between computational resources and accuracy (i.e. larger tree depths yield more accurate results but this algorithm can manage larger datasets if the KD tree depth is limited). Moreover, AVDPM limits the number of mixture components to a truncation level, *T*, such that additional components, *L>T*, can exist but are tied to their priors. For AVDPM the factorized variational distribution is given by [2]



where *qϕ*(*vi*), *qϕ*(η*i*), and *qz(zn)* represents parametric models for stick lengths, the components’ structures (e.g. means and covariances of Gaussians), and mixture component assignments respectively. Each of the parametric models’ respective parameters are given by *ϕ*.

CVSB and CDP, on the other hand, do not incorporate KD-trees but instead use a “hard” truncation level such that the maximum number of mixture components is *T*, i.e. any additional components have zero weight. The variational distribution is almost identical to that used for AVDPM with the exception that the truncation level is capped such that any more than *T* components have probability equal to zero [1].



While CVSB can have variable stick lengths, CDP imposes a symmetric prior on the variational distributions, i.e. the lengths of *k* stick breaks are all equal and thus weights of mixture components are all equal. This essentially reduces the Dirichlet process to a Dirichlet distribution and allows for the exchangeability of labels. The factorized variational distributions for CDP is:

 .

The primary difference between and is the replacement of the *q(v)* term by *q(π)*. The stick breaks, *v*, represent the fraction of the remaining stick length and are modeled with a beta distribution [9] while *π* are the actual mixture weights (i.e. each *π* represents the fraction of the original, whole stick). Since the length of each stick break is held constant, the effect from the mixture weights can be removed from the product and replaced by *q(π)*.

# Experimental Setup

In this work, the performance of AVDPM, CVSB, and CDP are compared to standard Gaussian mixture models. This section outlines some of the key details used in this work.

Labels for the CH-E corpus consist of the 39 phonemes found in the CMU7 dictionary as well as three additional labels – sp, sil, and a garbage phoneme – which are added to account for any partial words or sounds in the data. The CH-M corpus contains 92 phoneme labels consisting of the labels found in CH-M lexicon and the 3 additional labels used in CH-E corpus. Furthermore, English words that exist in CH-M are added to the CH-M lexicon where any English vowel sounds are assigned to the neutral tone. The relatively high number of labels is due to the tonal nature of Mandarin which requires all vowel sounds to have 5 labels (e.g. vowel “a” is actually “a1”, “a2”, “a3”, “a4”, and “a5”).

Phoneme alignments are generated by training a hidden Markov model (HMM) based acoustic model using a flat start and training up to 16 monophone mixtures. Finally, a Viterbi alignment is performed to identify phoneme segments. Any utterances from the corpora that contain simultaneous speech from multiple speakers are discarded.

Using the generated segmentations, 13 MFCC features and their first and second derivatives are extracted using a frame rate and window size of 10ms and 25ms respectively. The frame based features from each phoneme segment are averaged in a 3-4-3 manner so that the number of features per segment is constant despite duration (although duration is added as a single additional feature). Models are trained for each phoneme label and predictions are generated using maximum likelihood. Diagonal covariances are used to train GMM models and the number of mixture components is held constant for all phoneme labels. Conversely, AVDPM, CVSB, and CDP find this number, and corresponding means and covariances, automatically.

The best of 10 iterations of the GMM baseline is compared to the average performance of AVDPM, CVSB, and CDP over 10 iterations. Performance is evaluated using both error rates and also the average number of mixture components per phoneme label.

These algorithms are initially evaluated on the well calibrated TIMIT corpus to confirm that this setup produces comparable performance to other published results. Following the methods in [10][11][12], the corpus is partitioned into training, validation, and evaluation sets and the 61 original phonemes that exist in the TIMIT corpus were collapsed to 39 labels. GMMs are first fit using the phoneme alignments provided with the TIMIT corpus. The number of mixture components is swept for the GMMs and optimal performance of 31.56% misclassification error was found for 4 mixture components per phoneme label. This was comparable to the results found in [12] although for a much lower number of mixture components (i.e. 4 mixtures vs. 64 mixtures). This discrepancy is due to [12] using features only from the central portion of each phoneme segment instead of the 3-4-3 approach in this work. With this confirmation, phoneme alignments are then generated for the collapsed 39 labels in the same manner used for CH-E and CH-M. These results are shown in the following section and allow for a better comparison to the performance on the CH‑E and CH‑M corpora.

# Results & Discussion

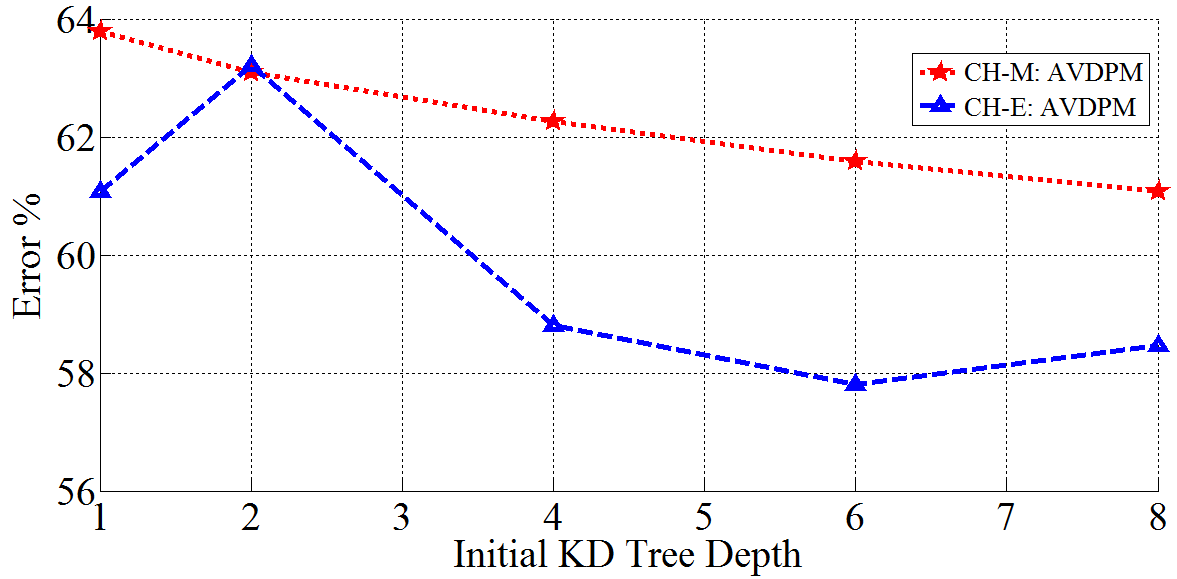


Figure 4: *A diagram showing how the performance of AVDPM is affected as the initial depth of the KD tree is varied for the validation data sets.*

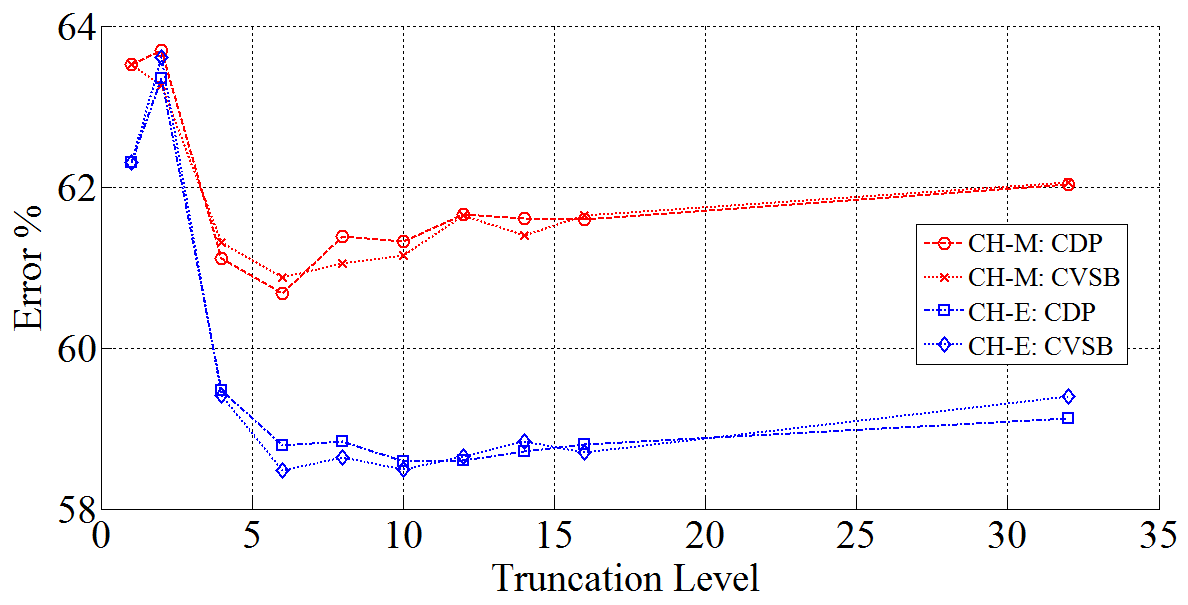


Figure 3: *A diagram showing how the performance of CVSB and CDP are affected as the truncation level is varied for the validation data sets.*

The truncation level for CVSB and CDP is swept to determine an optimal operating point for each corpus. Similarly, the initial depth of the KD tree is swept for AVDPM to determine how significantly this affects performance. Each algorithm is iterated ten times and an average misclassification error rate is calculated. The results for CH‑E and CH‑M are shown in Figure 3 and Figure 4. Furthermore a table of the best error rates on respective evaluation sets is shown in Table 1.

Table 1 shows that the average misclassification error of at least one variational inference algorithm always outperforms the baseline GMM and in all cases require significantly fewer parameters. It is interesting to note that relative performance of CVSB and CDP was worse for TIMIT than both CH-E and CH-M. This is most likely an artifact of the studio recorded, read speech of TIMIT which allows for the fixed number of mixture components of the GMM to reasonably approximate the underlying structure of the data. Conversely, CVSB and CDP are better suited to conversational telephone speech where the underlying structure is less apparent. Finally, the relatively small disparity between Mandarin and English can easily be attributed to Mandarin having more than double the number of phoneme labels as English. Thus each phoneme’s model is trained on less than half the number of segments as those for English.

All algorithms yield very comparable performance so the decision of which to implement should be decided on a case by case basis. However, AVDPM’s incorporation of KD trees might make it slightly more attractive for acoustic modeling since larger data sets can be managed by trading off the depth of the KD tree.

It can be seen in Figure 3 that both CH-E and CH-M have the same optimal truncation levels for both algorithms with the exception of CDP on CH-E. This is not unexpected since the symmetric prior CDP imposes on the lengths of the stick breaks indicates that there should be an equal or greater number of mixture components to compensate for that assumption. Furthermore, DPM’s ability to discover the underlying structure of the data makes these models less prone to overfitting and thus the maximum number of mixture components is significantly less than the baseline GMM model, where the number of components is assumed to be known a priori.

Table 1: *A comparison of misclassification error and number of mixture components for the evaluation sets of the TIMIT, CH‑E, and CH‑M corpora*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | TIMIT | CH-E | | CH-M | |
| **Error %** | **Error %** | **Avg. # Mixt.** | **Error %** | **Avg. # Mixt.** |
| GMM | **38.02%** | **58.41%** | **128** | **62.65%** | **64** |
| AVDPM | **34.69%** | **57.82%** | **5.14** | **63.53%** | **5.01** |
| CVSB | **40.30%** | **58.68%** | **5.89** | **61.18%** | **5.75** |
| CDP | **40.24%** | **57.69%** | **9.67** | **60.93%** | **5.75** |

Figure 4 shows AVDPM’s performance on CH-M improves as the initial depth of the KD tree increases. CH-E follows a slightly less expected trend which indicates that additional splitting of KD tree nodes is not always along optimal hyperplanes. However, it is expected that performance will generally improve at greater depths of the KD tree [2]. The computational complexity of this algorithm grows exponentially as depth increases [2], though, so an optimal operating point must be selected based on the computational resources available for a given task.

# Conclusions

Dirichlet distributions, and by extension DPMs, are the conjugate prior of the multinomial distribution and can be used to find underlying structure of data, e.g. the number of mixture components in a GMM. However, due to their infinite number of parameters variational inference algorithms are needed to make posterior calculations tractable. In this work, it is shown that three variational methods – AVDPM, CVSB, and CDP – all yield comparable performance to baseline GMMs but with significantly less parameters. This makes the selection of a preferred inference algorithm depend significantly on the task at hand. AVDPM may be best suited to acoustic modeling since controlling KD tree depth allows for the tradeoff between accuracy with available computational resources, thereby making training on large corpora possible. CDP, on the other hand, can provide slightly higher accuracy on noisier training data.

# Acknowledgements

This research was supported in part by the National Science Foundation through Major Research Instrumentation Grant No. CNS-09-58854. Our research group would also like to thank the Linguistic Data Consortium (LDC) for awarding a data scholarship to this project and providing the lexicon and transcripts for CALLHOME Mandarin.

# References

1. Kurihara, K., Welling, M., & Teh, Y. W. “Collapsed variational Dirichlet process mixture models” Twentieth International Joint Conference on Artificial Intelligence, 2007.
2. Kurihara, K., Welling, M., & Vlassis, N. “Accelerated variational Dirichlet process mixtures” NIPS, 2006.
3. Neal, R. “Bayesian mixture modeling by Monte Carlo simulation” Dept. Comp. Sci., University of Toronto Rep. CRG TR 91 2, 1991
4. Paul, L.M., Simons, G.F., Fennig, C.D., (2013). “Ethnologue: Languages of the World”, *Ethnologue* [online]. 17th ed. Available:   
   http://www.ethnologue.com.
5. Gu, W., Hirose, K., Fujisaki, H. “Comparison of Perceived Prosodic Boundaries and Global Characteristics of Voice Fundamental Frequency Contours in Mandarin Speech”, ISCSLP, 2006.
6. NIST, (2009). “The History of Automatic Speech Recognition Evaluations at NIST,” Available: <http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html>
7. Paisley, J., “Machine Learning with Dirichlet and Beta Process Priors: Theory and Applications.” Ph.D. dissertation, Dept. Elect. Eng., Duke Univ. NC, 2010
8. Rasmussen, C. E. “The Infinite Gaussian Mixture Model.” NIPS (pp. 554–560). MIT Press. 2000.
9. Blei, D. M., Jordan, M. I. “Variational inference for Dirichlet process mixtures.” Bayesian Analysis, 1(1), 121–143. 2006. doi:10.1214/06-BA104
10. Lee, K. F., & Hon, H.W. “Speaker-independent phone recognition using hidden Markov models.” IEEE Transactions on Acoustics, Speech, and Signal Processing, 37(11), 1641–1648. 1989. doi:10.1109/29.46546
11. Halberstadt, A., & Glass, J. “Heterogeneous acoustic measurements for phonetic classification.” Proceedings of Eurospeech 401–404. 1997.
12. Ager, M., Cvetkovic, Z., Sollich, P. “Robust phoneme classification: Exploiting the adaptability of acoustic waveform models.” EUSIPCO. 2009.