MAP Estimation of Whole-Word Acoustic Models with Dictionary Priors

Keith Kintzley[1], *Aren Jansen*[1,2], *Hynek Hermansky*[1,2]

[1]Dept. of Electrical and Computer Engineering, [2]HLT Center of Excellence
Johns Hopkins University, Baltimore, Maryland, USA

kintzley@jhu.edu, aren@jhu.edu, hynek@jhu.edu

Abstract

The intrinsic advantages of whole-word acoustic modeling are offset by the problem of data sparsity. To address this, we present several parametric approaches to estimating intra-word phonetic timing models under the assumption that relative timing is independent of word duration. We show evidence that the timing of phonetic events is well described by the Gaussian distribution. We explore the construction of models in the absence of keyword examples (dictionary-based), when keyword examples are abundant (Gaussian mixture models), and also present a Bayesian approach which unifies the two. Applying these techniques in a point process model keyword spotting framework, we demonstrate a 55% relative improvement in performance for models constructed from few examples.

Index Terms: phonetic timing, whole-word modeling, keyword spotting, point process model

1. Introduction

Isolated word recognition systems in the early days of speech recognition were often constructed by modeling entire words. While practical for limited vocabulary size, the advent of large vocabulary systems based on hidden Markov models (HMMs) necessitated the use of subword units to enable the sharing of training examples across contexts and permit the modeling of unseen words. However, if training examples are available, by maintaining the structure of the word, whole-word models have long been known to offer superior performance to subword-based systems [1].

The synthesis of words from subword units and the resulting geometric state duration distributions are partially responsible for the HMM's well-known deficiency in duration modeling. Additional constructs within the HMM framework such as segment models [2] have been introduced to address these shortcomings at the cost of increased complexity. As HMMs have been shown lacking in modeling duration, a large body of research has documented the importance of temporal cues in human speech perception [3].

Maximum a posteriori (MAP) approaches have been applied to HMM parameter estimation for purposes such as parameter smoothing and speaker adaptation [4]. Prior HMM parameter distributions based on context independent phone models can be used in the estimation of context dependent models. Likewise, speaker adaptation can be enhanced by using speaker independent prior models when speaker-specific data is limited. In both these cases, the prior is based on class-independent HMM parameter averages and MAP estimation enables smoothed estimates of class-specific HMM parameters. Unlike HMM models, we are specifically modeling the timing of phonetic events, and our prior will take a very tangible form rendered from a word's phonetic composition.

The point process model (PPM) for keyword spotting is a recently proposed whole-word modeling approach [5]. Keywords are characterized by an inhomogeneous Poisson process, and keyword detections are derived from the relative timing of a sparse set of phonetic events. Like other whole-word approaches, data sparsity is a problem. In all previous PPM applications, the Poisson rate parameters have been calculated using maximum likelihood estimation (MLE). As previously documented, system performance depends on the accurate estimation of Poisson rate parameters which in turn requires large numbers of example keywords. In this paper we address the issue of data sparsity by introducing parametric models of phonetic event distributions. Using MAP estimation with simple dictionary-based prior distributions, we overcome the need for large amounts of training data and provide a seamless transition between subword and whole-word frameworks.

2. Theory

A speech signal can be decomposed into a continuous sequence of phones, and speech corpora such as TIMIT include detailed labeling of these phonetic boundaries which can be used to train phone recognition systems. These recognizers commonly output phonetic information in the form of a *phone posteriorgram*, the vector time series of posterior distributions across the phone set $\mathcal{P}$ for each speech frame. Instead of considering vector time series estimates of $P(p|x)$ on a frame-by-frame level, in this work we will be operating on a discrete sequence of *phonetic events* represented by the time of occurrence of each phone. Ideally, there would be exactly one phonetic event located at the midpoint of each phone label. From each phone posteriorgram we generate a set of events through the use of phonetic matched filtering as explained in [7].

As in previous works on the point process model, we make the simplifying assumption that the relative timing of phonetic events within words is independent of word duration. This assumption allows us to separate the word duration model from an intra-word phonetic timing model. Additionally, by considering phonetic timing relative to a normalized word duration of 1.0, we are able to pool the keyword examples used to calculate the phonetic timing model. The normalized word dura-

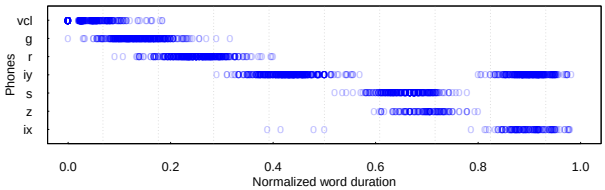

Figure 1: *Timing of phonetic events for 462 examples of the keyword "greasy" plotted against normalized word duration.*

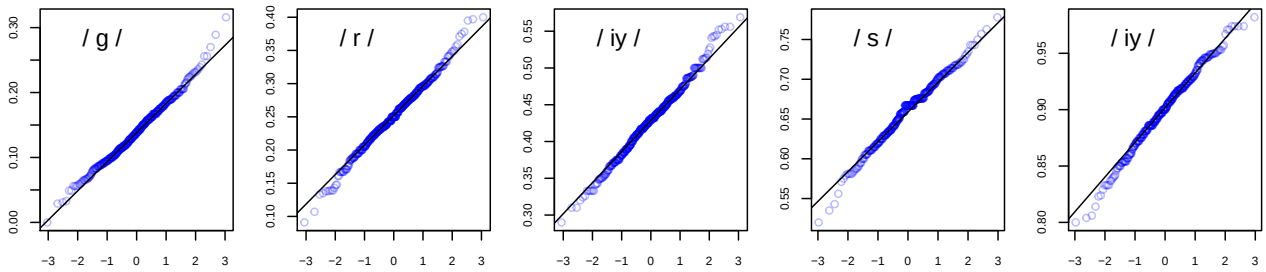

Figure 2: *Normal Q-Q plots of phonetic timing data for phones in the keyword "greasy" showing approximate normality of timing distributions. Data quantiles are shown on the vertical axis and theoretical quantiles on the horizontal axis.*

tion is then partitioned into D subdivisions (typically 10), and Poisson rate parameters are calculated for each subdivision for each phone. Maximum likelihood estimation (MLE) of these parameters is based on the total count of phonetic events observed in each subdivision over all training instances. Figure 1 illustrates the distribution of phonetic events relative to normalized word duration derived from 462 examples of the keyword "greasy." While the plot suggests to the possibility of modeling intra-word phonetic events using normal distributions, to further quantify this intuition we present Q-Q plots in Figure 2 for the phones in "greasy" comparing the empirical distribution with a Gaussian. Repeating this assessment for the other keywords, we observed that the distributions of phonetic events are reasonably well modeled by the Gaussian distribution.

2.1. Gaussian mixture model

An obvious choice for constructing a model of phonetic event timing is to use a Gaussian mixture model (GMM) for each phone. Our GMM estimation is initialized by performing k-means clustering on the phonetic events for each phone where the number of clusters k should reflect the notion that each phone instance within a word can be modeled by a single Gaussian (i.e. the phone /iy/ in "greasy" should be modeled using $k = 2$ Gaussians). While our clustering is *not* informed by the word's dictionary form, we can encourage clustering consistent with this idea by allowing k to grow as long as successive cluster means are separated by greater than 4 standard deviations. After performing expectation-maximization we obtain the mean, variance and mixture coefficients for each phone GMM. While this model does reflect the distribution of events in time, ultimately we will be using it to compute the *expected counts* of phonetic events and therefore must scale each GMM appropriately. Thus, if we observe n_p phonetic events from n_w keyword examples, applying a scale factor of n_p/n_w to the GMM for phone p allows us to compute expected counts.

A GMM-based model for the word "greasy" is shown in Figure 3a. Note that this model captures both speaker pronunciation variation and phone detector confusions. The model depicted in this figure is based on all 462 training examples. When multiple models are estimated using different ensembles at a fixed sample size, the variation across models will increase as sample size decreases. However, given that each Gaussian requires the estimation of only 2 parameters, we would expect GMM-based models to exhibit less variation than is observed when estimating 10 parameters per phone as is typically the case with MLE estimation in the point process model.

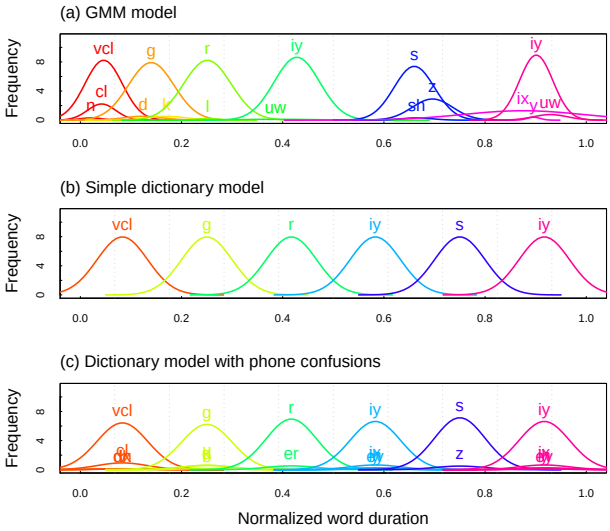

Figure 3: *Phone timing models for the keyword "greasy."*

2.2. Dictionary-based models

The GMM model of the previous section is derived entirely from keyword examples with no prior assumptions about the phonetic composition of the the word. Such models result in good keyword spotting performance when training examples are plentiful. However, consider the case where no keyword examples are available and our only knowledge comes from a word's dictionary (phonetic) form. In such a case, we could construct a naive model by assigning one Gaussian to each phone in the dictionary form with equally spaced means μ and a fixed standard deviation σ. Such a model for "greasy" is depicted in Figure 3b with σ =0.05.

Comparing the GMM and dictionary models in this figure, an obvious shortcoming of the dictionary model is the inability to accommodate pronunciation variation and likely phone confusions. Variation which arises from different speaker productions could be incorporated using weighted combinations of alternate dictionary forms. Lacking this information, a very simple alternative is to apply phone confusion matrix data associated with the phone detectors. If rows of the confusion matrix correspond to actual phone classes and columns correspond to predicted phone classes, then each matrix element $C_{ij} = \Pr(p_j|p_i)$. In this paper, we have derived a phone confusion matrix from the count matrix employed in phonetic event selection from [7]. To introduce likely confusions into our dictionary model,we replace the single Gaussian for phone p_i from a word's phonetic form with with multiple Gaussians for the

confusable phones p_j each weighted by C_{ij} but sharing a common μ and σ. The resulting model is depicted in Figure 3c.

3. Bayesian estimated parametric models

We have presented model construction at two ends of the spectrum: a model assembled without examples (dictionary-based) and an efficient parametric model built entirely from data (GMM). Now we introduce an approach to blending these models to better handle situations where only a handful of keyword examples exist. The basis of our approach is standard Bayesian inference for Gaussian distributions where both the mean μ and precision $\lambda \triangleq 1/\sigma^2$ are unknown. The conjugate prior is given by the normal-Gamma distribution:

$$NG(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) = \mathcal{N}(\mu|(\kappa_0\lambda)^{-1})\,\mathrm{Gam}(\lambda|\alpha_0, \beta_0)$$

with hyperparameters μ_0, κ_0, α_0, and rate β_0. Initial values of the hyperparameters are determined from our prior estimates of the mean and precision of μ and λ. Since the precision λ is gamma distributed, α_0 and β_0 are determined using the relations $\mathrm{E}[\lambda] = \alpha_0/\beta_0$ and $\mathrm{Var}[\lambda] = \alpha_0/\beta_0^2$. It can be shown that the random variable μ is t-distributed with $2\alpha_0$ degrees of freedom, location parameter μ_0 and precision parameter $\kappa_0\alpha_0/\beta_0$. Thus, $\mathrm{E}[\mu] = \mu_0$ and $Var[\mu] = \beta_0/(\kappa_0(\alpha_0 - 1))$ which allows us to compute κ_0.

For simplicity, we have chosen the prior mean μ_0 just as we did in the dictionary-based model. Likewise, in the dictionary model we set the standard deviation of each Gaussian $\sigma = 0.05$, so we now set $\mathrm{E}[\lambda] = 400$ and choose $\mathrm{Var}[\lambda] = \rho\,\mathrm{E}[\lambda]$ where $\rho = 0.5$. The κ_0 parameter corresponds to the *equivalent sample size* of the prior, so rather than choosing a prior value for $\mathrm{Var}[\mu]$ it is more intuitive to set $\kappa_0 = 1$. This is akin saying that our prior carries the same weight as a single observation. The impact of these choices of priors will be addressed below. Combining the initial hyperparameters and observed phonetic event data, the posterior distribution of μ and λ is calculated using simple hyperparameter updates found in [8]. Finally, MAP estimates of μ and λ are found using numerical optimization.

How do we account for pronunciation variation using this approach? In the absence of examples, our prior consists of the dictionary model pictured in Figure 3c where each distribution is weighted by the phone confusion matrix element C_{ij}. As the number of examples increases, our model should asymptotically approach the GMM-based model in 3a. We can explain the production of phonetic events using the following generative story: a phonetic event is the result of two independent random variables, the first being a Bernoulli random variable which dictates whether the event is observed in a word, and the second a Gaussian random variable which specifies the event's position in word. This is analogous to the graphical model description of a GMM, and in our Bayesian treatment, the parameter π associated with the Bernoulli random variable *is itself a random quantity* with conjugate prior distribution $\mathrm{Beta}(\pi|a, b)$. Under this distribution, $\mathrm{E}[\pi] = a/(a+b)$ and the sum $a+b$ constitutes the *effective number of observations*. For the prior mean, we set $\mathrm{E}[\pi]$ equal to the phone confusion matrix value C_{ij}. This and our choice of the effective number of observations fully specify the values a and b. We will now describe the posterior update. If we have a total of n keyword examples in which a phonetic event for the phone we are modeling is present in m cases and absent in $l = n - m$ cases, then the posterior distribution of π will be $\mathrm{Beta}(\pi|a + m, b + l)$. Thus, we will take the mean of the posterior distribution $(a + m)/(a + b + n)$ as the posterior updated value of the parameter π. An illustration of how the

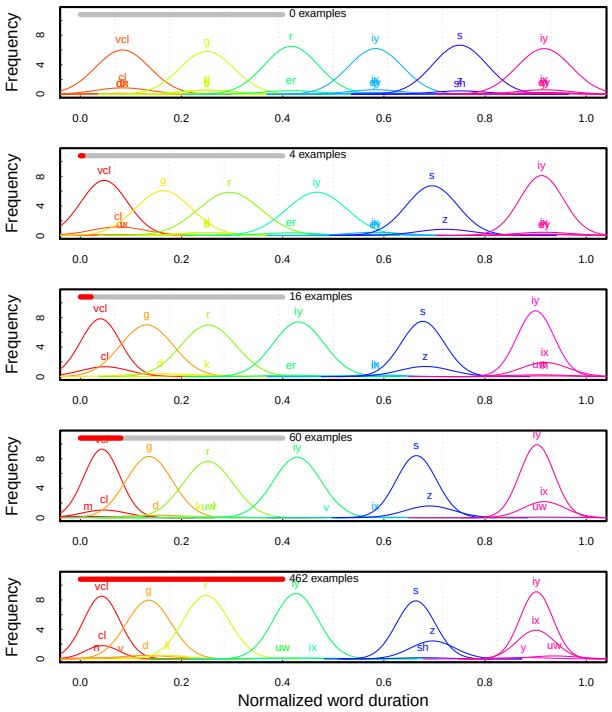

Figure 4: *Bayesian estimated phone timing models for the keyword "greasy" constructed using various numbers of examples.*

Bayesian model for "greasy" evolves as we increase the number of keyword examples is shown in Figure 4. Note that for the case $n = 0$ examples, the Bayesian model is identical to the dictionary model with phonetic confusions in Figure 3c. Likewise at $n = 462$ examples, the Bayesian model largely mirrors the GMM-based model in Figure 3a (there are minor differences such as the handling of the phone /ix/). As the number of examples increases, the mean of each Gaussian shifts rapidly toward its limiting value and the variance contracts. Additionally, we observe the development of pronunciation variation (/s/ with /z/ and final /iy/ with /ix/).

4. Experiments

We have proposed several approaches to modeling phonetic event timing distributions. In this section we apply these techniques in constructing point process models for keyword spotting, specifically in the estimation of inhomogeneous Poisson rate parameters. In previous works, MLE rate parameter estimates were derived from the counts of events in each word subdivision. Given any model of the phonetic event distribution, we can simply replace the hard counts with *expected counts* under the model. All of our keyword spotting experiments were conducted on the TIMIT database to permit easy comparison to previously published work on the point process model. The phonetic event data was derived in the same manner as described in [7]. Perceptual linear prediction (PLP) acoustic features were transformed into a phone posteriorgrams using a sparse multilayer perceptron (SMLP) based system from [6]. Posteriorgrams were then converted into phonetic events by applying phonetic matched filters to the posterior trajectories and selecting local maxima above a threshold $\delta = 0.22$.

In Figure 5 we show keyword spotting performance as a function of the number of training examples used in model construction. Performance is measured by average figure of merit

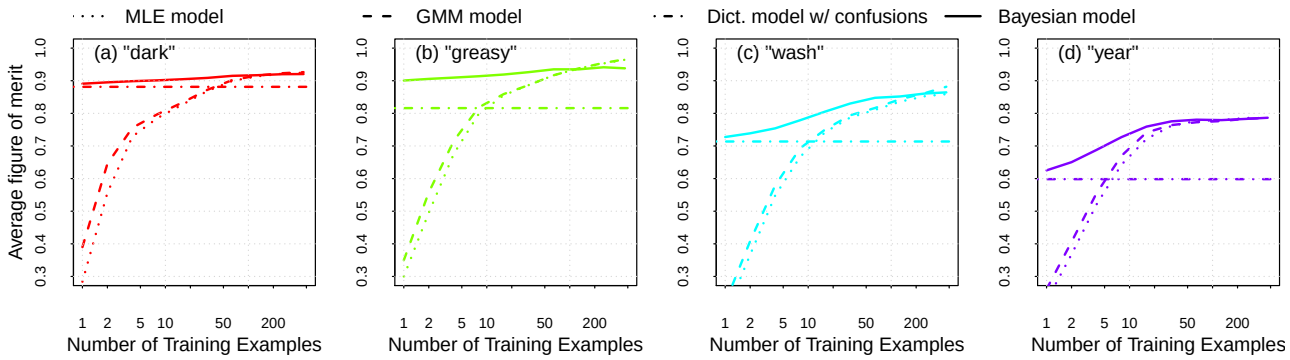

Figure 5: *Average figure of merit vs. number of examples used in model construction for various TIMIT sa1 keywords.*

(FOM), the mean detection rate given $1, 2, \ldots, 10$ false alarms per keyword per hour as the detection threshold is varied. The plot for each keyword includes results from each of the parametric event timing models as well as the original MLE-based model. Our test keywords were necessarily limited to the few sufficiently long TIMIT sa1/sa2 words. The dictionary-based model with phone confusions does not depend on any keyword examples, thus its performance is constant. The results depicted for GMM, MLE and Bayesian models represent the mean performance of many models based on random draws of examples.

In comparing the various models we first note that the dictionary-based model dramatically outperforms both the GMM and MLE models in the low example count regime. Clearly, when there are fewer than 10 examples, insufficient data exists to estimate the distributions. However, when training data is abundant the MLE and GMM models provide as much as 20% absolute increase in performance as they more accurately describe the word's phonetic events. Between the GMM and MLE models, we note that the GMM model provides a small improvement ($\sim$6%) over the original MLE model for small example counts, likely because the GMM model has fewer parameters to estimate. Finally, using the Bayesian model which incorporates both the phonetic form and evidence from keyword examples, we achieve strong performance in all regimes. Remarkably, despite using an extremely simple prior model, we have achieved a 55% relative increase in keyword spotting performance when confronted with few keyword examples.

Although we have not included error bars, the performance of MLE-based models exhibits very high variance. Bayesian models, on the other hand, are heavily constrained by prior distributions at small sample sizes which results in dramatically smaller variance in performance. In fact, for the set of words considered in this work, we observed that variance in average FOM was *reduced by a factor of 14* for cases of few ($\leq$8) training examples. In a minor change from previous work on the point process model, we have incorporated a parametric word duration model. In this work, the distribution of keyword duration is modeled by a gamma distribution whose parameters are estimated using maximum likelihood. In all of the keyword spotting results for parametric timing models in this paper, we used identical duration models based on all 462 examples of each keyword.

The estimation of the Bayesian model does require choices about prior distributions for the mean μ and precision λ. In this work we have made no effort to optimize these values. For prior estimates of precision λ, we tried $\mathrm{E}[\lambda] = 100$, 400, and 2500. Likewise for the parameter ρ relating the mean and variance of the gamma distribution for precision λ, we ran trials with $\rho = 0.25$, 0.50 and 0.75. For all cases, we observed little difference in the models or keyword spotting performance.

5. Conclusions

In previous applications of the point process model for keyword spotting, the chief limitation was the large numbers of keyword examples (>50) required to construct representative keyword models. Though simple models based solely on a word's dictionary form offer reasonable performance, they are incapable of benefiting when examples are available. We have demonstrated in this paper that the use of Bayesian estimation techniques provides a principled method of combining both prior knowledge of phonetic composition and information from keyword training examples. Furthermore, we have shown the evolution in model distributions ultimately results in an optimal interpolation of the performance gap as the number keyword examples grows. Lastly, we believe that this approach holds promise for the iterative construction of keyword models through self-supervision.

6. Acknowledgments

The research presented in this paper was partially funded by DARPA RATS program under D10PC20015.

7. References

[1] Lee, C.-H., et al.,"Word recognition using whole word and subword models, in *Proc. of ICASSP,* pp.683–686, 1989.

[2] Ostendorf, M. and Digalakis, V.V. and Kimball, O.A.,"From HMM's to segment models: a unified view of stochastic modeling for speech recognition," in *IEEE Trans. Audio, Speech and Audio Proc.,* 4(5):369–378, 1996.

[3] Cutler, A., Dahan, D., van Donselaar, W., "Prosody in the Comprehension of Spoken Language: A Literature Review," in *Language and Speech*, vol. 40, pp.141–201, 1997.

[4] Gauvain, J.-L., Lee, C.-H., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," in *IEEE Trans. Audio, Speech and Audio Proc.,* 2(2): 291–298, 1994.

[5] Jansen, A. and Niyogi, P.,"Point Process Models for Spotting Keywords in Continuous Speech", *IEEE Trans. Audio, Speech and Language Proc.,* 17(8):1457–1470, 2009.

[6] G.S.V.S. Sivaram and Hermansky, H., "Multilayer Perceptron with Sparse Hidden Outputs for Phoneme Recognition," in *Proc. of ICASSP,* 2011.

[7] Kintzley, K., Jansen A. and H. Hermansky, "Event Selection from Phone Posteriorgrams Using Matched Filters," in *Proc. of INTERSPEECH,* pp.1905–1908, 2011.

[8] DeGroot, M., *Optimal Statistical Decisions,* McGraw-Hill, 1970.