Text-to-Speech Inspired Duration Modeling for Improved Whole-Word Acoustic Models

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

In the construction of whole-word acoustic models, we have previously demonstrated substantial gains by using MAP estimation to introduce a simple prior model of phonetic timing. Based solely on the word's phonetic (dictionary) pronunciation, this simple model included no information about the individual durations of constituent phones. However, the problem of modeling segmental duration has long been studied in the text-to-speech (TTS) community. We draw upon this work to develop a classification and regression tree (CART) approach for constructing prior models of phonetic timing which considers factors such as syllable stress, syllable position, adjacent phone class and voicing. This improved prior model closes 33% of the gap in keyword spotting performance between highly supervised whole-word models and those estimated without any examples.

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

1. Introduction

A large body of evidence suggests the preeminent importance of temporal properties of the speech signal in human speech perception. Spectral cues in speech are directly related to vocal tract configuration and have long been considered principal in the identification of speech sounds. Interesting, human speech perception is remarkably tolerant of significant degradation in frequency information. In one notable study [1], when detailed frequency information was excised from speech while retaining the signal's temporal envelope, listeners were remarkably accurate in identifying consonants, vowels and words using just three wide bands of modulated noise. Furthermore, it has been demonstrated that the distortions which result in the most significant loss of intelligibility are those which affect slow temporal modulations (2-10 Hz) of speech [2]. Finally, many studies of children with language impairment have concluded that their deficiency stems from a basic temporal processing deficit [3].

In addition to intelligibility, temporal structure also plays a crucial role producing natural sounding synthetic speech. Early attempts to predict the systematic changes in the duration of phonetic segments involved defining a set of hand-designed rules based on contextual factors such as adjacent segment identity, within-word position, syllable stress, among others [4]. A more statistically grounded approach that offers greater ability to model the interaction between factors is found in the sum-of-products model presented in [5]. CART-based modeling is another widely used approach to predicting segmental duration that provides automatic selection of relevant features, accommodates both categorical and continuous features, and produces easily interpretable rules [6].

In light of the importance of temporal cues on human speech perception, we are motivated to investigate whole-word acoustic models such as those found in the point process model (PPM) for keyword spotting [7]. In the PPM framework, keywords are characterized by an inhomogeneous Poisson process, and keyword detections are derived from the relative timing of a sparse set of phonetic events. Operating on a small number of discrete events, PPM has advantages in computational simplicity and has been shown to enable very fast keyword searches without the use of an index [8]. Like other whole-word approaches that rely on training examples, data sparsity is a problem. In its original inception, PPM word model parameters were been calculated exclusively from keyword examples using maximum likelihood estimation (MLE) but the accurate estimation of Poisson rate parameters was shown to require large numbers of example keywords. In a recent work [9], it was demonstrated that strong performance could be achieved without the need for large amounts of training data by modeling phonetic timing with Gaussian distributions and estimating parameters using a Bayesian approach.

Each Gaussian in this model requires the estimation of a mean and variance (or precision), and MAP estimation of these parameters presumes the existence of reasonable prior distributions. In the initial presentation of MAP-estimated whole-word acoustic models, a very basic prior distribution was assembled from equally-spaced Gaussian means with uniform variance. This simple model suffices as an initialization point when combined with training examples. However, we would like to improve the estimation of word models for cases when no examples exist, and this naive prior ignores obvious differences in phone duration. As suggested, the problem of constructing a reasonable prior model of phonetic timing is very closely related to that of computing segmental duration for TTS synthesis. In this work we will evaluate the gains possible by enhancing priors with CART-based duration estimates.

2. Theory

The approach to keyword spotting considered here is based on modeling the underlying distributions of phonetic events within a word as inhomogeneous Poisson processes. Unlike most speech recognition systems which are based on dense, frame-by-frame estimates of phone likelihoods, the PPM approach operates on a sparse set of phonetic events. If one considers a standard frame-level phone alignment, we would mark a single phonetic event at the midpoint of each phone label (in previous work we have referred to these as *oracle* phonetic events). To obtain phonetic events from actual speech data, we employ a multilayer perceptron to estimate the posterior probability of each phone class for each frame of speech. This phone posteri-

orgram representation is then processed using phonetic matched filters as presented in [10]. Consistent with previous PPM implementations, we make the simplifying assumption that the relative timing of phonetic events within words is independent of word duration and thus consider models of word duration normalized to 1.0. This unit duration is partitioned into D (typically 10) equal subdivisions, and for each division we estimate a piecewise constant approximation of the inhomogeneous rate parameter for each phone. Given a number of keyword examples, maximum likelihood estimates for the rate parameters are computed from the total count of phonetic events observed for each phone and each subdivision.

Computing reliable MLE estimates typically requires a sizeable number (>50) of keyword training examples. However, in a previous work [9] we observed that distributions of phonetic events across duration normalized keyword examples were well described by Gaussian distributions. Based on this finding, we developed a Bayesian approach to modeling the distributions of phonetic events in which a simple prior was constructed from a word's pronunciation dictionary phonetic base form. Then, as an alternative to estimating inhomogeneous Poisson rate parameters from actual counts of phonetic events, with a parametric distribution we can instead use *expected counts* under the each distribution.

In the following sections, we consider three approaches to defining phonetic timing distributions in the *absence* of any keyword examples. We begin by reviewing the simple dictionary model of [9] and then introduce two enhanced models.

2.1. Simple dictionary-based prior model

If no training examples of a keyword are available, it is possible to construct a naive model of phonetic timing using the keyword's dictionary pronunciation. Given a normalized word duration of 1.0, we simply assign one Gaussian to each phone in the dictionary form using equally spaced means μ and a fixed standard deviation σ. An example of such a model with σ=0.05 for the word "often" is depicted in Figure 1a. Despite its simplicity, we have shown this to be a practical method of assembling a prior for subsequent MAP estimation.

We have also found that introducing phonetic variation is a critical element in obtaining reasonable keyword spotting performance with such models. The use of alternate pronunciations could account for different speaker productions. However, another significant source of differences in observed phonetic events is caused by errors which occur in our phone posteriorgrams. A reasonable means of accounting for both errors and variation is to factor in phone confusion matrix data associated with the phone detectors. If rows of the confusion matrix correspond to actual phone classes (p_i) and columns correspond to predicted phone classes (p_j), then each matrix element C_{ij} represents $\Pr(p_j|p_i)$. Here, we have obtained a phone confusion matrix from the count matrix employed in phonetic event selection presented in [10]. To incorporate likely confusions into our dictionary model, we replace the single Gaussian for phone p_i from a word's dictionary form with with multiple Gaussians for the confusable phones p_j each weighted by C_{ij} but sharing a common μ and σ as illustrated in 1a.

2.2. Monte Carlo prior using average phone durations

In the simple dictionary model, assigning Gaussian means at equal intervals corresponds to an assumption that all phones are identical in duration. The fixed standard deviation σ=0.05 was chosen empirically to produce satisfactory keyword spotting performance over many keywords. To develop a more realistic model which accounts for phone duration, we first introduce the following expression for the relative timing of phonetic events. Given a word with baseform pronunciation $p_1, p_2, \ldots, p_N$, where each p_i is drawn from the set of all phones $\mathcal{P}$, we define D_i as a random variable representing the duration of the p_i. We can then define R_i as the midpoint of p_i (after word duration normalization), which is given by

$$R_i = \frac{\sum_{j=1}^{i-1} D_j + 0.5 D_i}{D_1 + D_2 + \cdots + D_N}. \quad (1)$$

As a starting point we assume that the distribution of the phone duration D_i is derived from the duration statistics of phone p_i realized across all words in the corpus and that D_i is independent of the other phones in the word. A convenient distribution for modeling phone duration is the two-parameter gamma distribution [11]. Studies have shown that the gamma distribution provides a high-quality fit to empirical phone and word duration distributions [14].

The random variable R_i is a function of N independent, gamma-distributed random variables and there is no simple closed-form solution for its distribution. Fortunately, it is sufficient for our purposes to estimate just the mean and variance of R_i. These quantities are easily obtained from a Monte Carlo simulation as follows: (i) compute gamma parameters (α, λ) to fit all of the phones in $\mathcal{P}$ based upon examples across the entire corpus; (ii) for a particular word, independently generate N sample phone durations corresponding to each D_i; (iii) from the N duration samples D_i, compute the corresponding N values of R_i; (iv) repeat over many (10,000) iterations and compute sample mean and variance for each R_i; (v) construct a model from N Gaussian distributions using the mean and variances of each R_i.

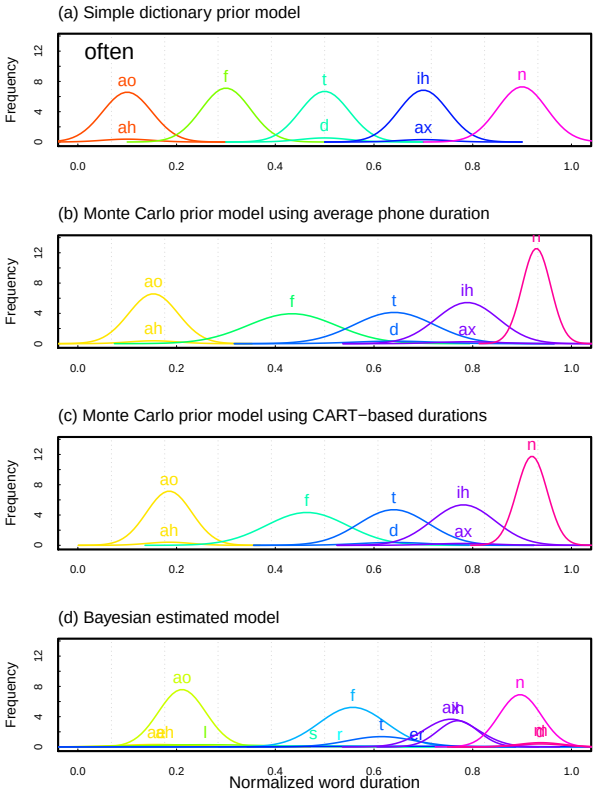

Figure 1: *Example of phone timing models for the word "often."*

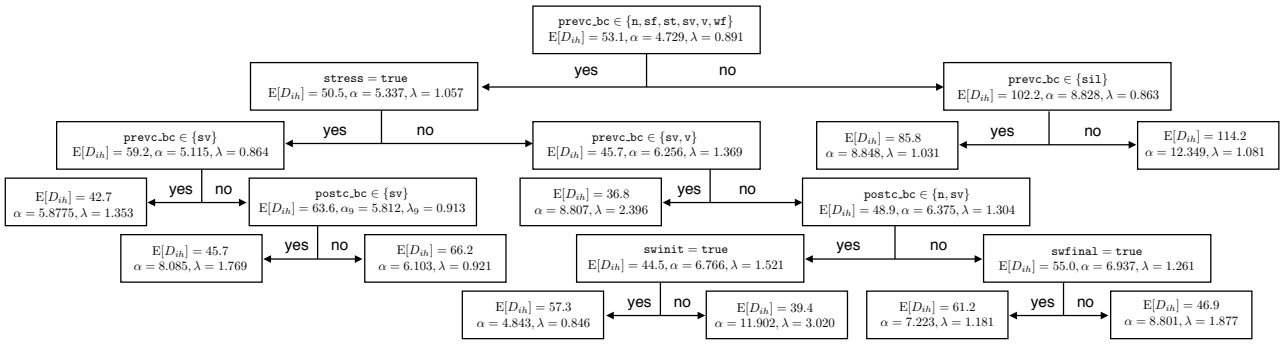

Figure 2: *CART tree for predicting duration of the phone /ih/. Each node shows the decision tree question, mean duration* $\mathrm{E}[D_{ih}]$*, and gamma distribution parameters (*α*,* λ*) for all training examples at that node.*

An example of a model computed using this approach is shown for the word "often" depicted in Figure 1b. Unlike the simple dictionary model shown in Figure 1a, we observe that the positions of the means better respect the average phone durations. Additionally, we find that the variances of the distributions are smaller for phones nearest the word beginning and ending, and larger for phones in the middle of the word. This is a natural byproduct of normalizing word durations and is evident in equation (1) and can also be seen in models estimated from many keyword examples (Figure 1d).

2.3. Monte Carlo prior using CART-based phone durations

While the incorporation of average phone duration clearly improves the fidelity of the model compared with the simple dictionary version, it is well known that segmental duration is a function of many factors such as phonetic context, stress, syllable position, and others. The text-to-speech community has developed several approaches to model segmental duration and here we adopt the method based on classification and regression trees. In order to perform CART training, we begin with a pool of example phone durations and an associated set of linguistically relevant features for each example. In speech synthesis, the prediction of duration is performed for an entire utterance, but our prediction can only consider a word in isolation. Therefore, it is not possible to consider some commonly used features such as utterance and phrase position. To compile a training set, we extracted phone durations from our corpus and generated a feature vector for each sample. The features associated with each phone and each word position were derived from the syllable and stress markings provided by the CMU dictionary [13]. We used the following set of features:

- `wsl` - word syllable count (7 levels)
- `swinit` - syllable word initial (boolean)
- `swfinal` - syllable word final (boolean)
- `sp` - syllable position (3 levels: onset, nucleus, coda)
- `stress` - stressed syllable (boolean)
- `prevoice` - previous phone voiced (boolean)
- `postvoice` - next phone voiced (boolean)
- `prev_bc` - previous phone broad class (5 categories)
- `post_bc` - next phone broad class (5 categories)

For each of the phones, we constructed a regression tree using the package `tree` in the statistical software package `R`. An example tree for the phone /ih/ is shown in Figure 2. In our example for the word "often" (/ao,f,t,ih,n/), the phone /ih/ which is preceded by a stop consonant, is not in a stressed syllable, is followed by a nasal and is not word initial, the decision tree shown would predict an expected duration D_{ih} of 39.4 ms compared with the population average of 53.1 ms.

Unlike speech synthesis where it is sufficient to predict just a duration, we need to predict the distribution of the phone /ih/ in its context. We accomplish this by using the decision tree to cluster training examples, and then estimate gamma distribution parameters (α, λ) at each node of the tree. The root node contains all examples, and its distribution represents the entire population independent of context. Each question in the tree partitions the examples into two subsets from which we compute corresponding gamma distribution parameters. Continuing to split our examples at each tree node allows us to compute distribution parameters for each context as shown in in Figure 2.

Having estimated the context-dependent gamma distribution parameters, the construction of a word model follows in a similar manner to the Monte Carlo model based on average phone durations in the previous section. However, instead of drawing samples for D_i from the distribution of the entire population, we instead use each phone's context determined by a word's dictionary form to identify the context-dependent distribution parameters contained in the decision tree.

An example of a model computed using this approach is shown for the word "often" depicted in Figure 1c. For reference, the model in Figure 1d is generated using MAP estimation with many training examples of the keyword. Note that in the progression of models from simple to more complex, the locations of the distributions better reflect the models derived from keyword data. To quantify the effect improved timing models, we computed the root mean squared error (rmse) between the mean values of R_i under these three models and the positions determined from keyword examples. We found that the Monte Carlo average model provided a 16.7% reduction in rmse relative to the simple dictionary model, and the Monte Carlo CART model yielded a 21.3% relative reduction.

3. Experiments

To measure the impact of more precise timing models, we conducted a series of keyword spotting experiments using the Wall Street Journal (WSJ0 and WSJ1) datasets. The training portions of this corpus were partitioned into two folds of 23 hours of speech. The audio data was processed into perceptual linear prediction (PLP) features and then transformed into a phone posteriorgram representation using a hierarchical MLP with 9 context frames [12]. From posteriorgram data, we then extracted phonetic events using phonetic matched filters as described in [10] with a threshold of $\delta = 0.24$. In order to eval-

uate models over a wide variety of words, we assembled a list of 1521 keywords from the WSJ corpus with minimum average duration of 200 ms, a minimum of 4 phones, and which occurred at least 10 times in each data fold.

For each keyword and each data fold, we created 4 types of keyword models: 1) simple dictionary model, 2) Monte Carlo estimated model based on average phone duration statistics, 3) Monte Carlo estimated model using CART-based phone duration statistics, and 4) Bayesian estimated models as described in [9]. Of these four phone timing model types, the first three were constructed without using keyword examples and only relied on duration statistics of their constituent phones. On the other hand, the Bayesian model uses all available keyword examples. All training and model parameter estimation (phone duration statistics, CART estimation, etc.) was performed on one data fold and evaluation was performed on the other (unseen) data fold. Performance reported represents an average per keyword over both folds.

We evaluated keyword detection performance using average figure of merit (FOM) which is defined as the mean detection rate given $1, 2, \ldots, 10$ false alarms per keyword per hour as the detection threshold is varied [15]. FOM provides a summary of the performance at the higher precision portion of the receiver operating characteristic curve. A summary of results of keyword spotting experiments is shown in Figure 3. The average performance of the various models is enumerated in Table 1. In addition to the average over all 1521 keywords, the table also shows the average over the subset of keywords which ranked below the bottom 10th percentile with the simple dictionary model.

We observe from Table 1 that improvements in estimating the prior models of phone timing distributions do result in improvements in FOM, on average. However, it would appear most of the gain in prior model performance is obtained from the Monte Carlo estimation of relative timing. The additional gain achieved through the inclusion of CART duration modeling was limited. There were several keywords for which the CART model provided improvements compared to the Monte Carlo average model, as evidenced by the significant increase in the minimal FOM value. However, we were unable to identify a systematic keyword property that accounted for these occurrences. While generating a more sophisticated prior model for phonetic timing was a logical place to look for improved performance, we observe other factors in Bayesian models which account for their superior performance. Chiefly, Bayesian models more accurately represent phonetic variation observed in keyword examples. This suggests that further improvement might come by adding alternate pronunciations in our dictionary. Additionally, more investigation may reveal systematic errors in phone posteriorgram estimates which may be predictable from context instead of using on phone confusion matrix data.

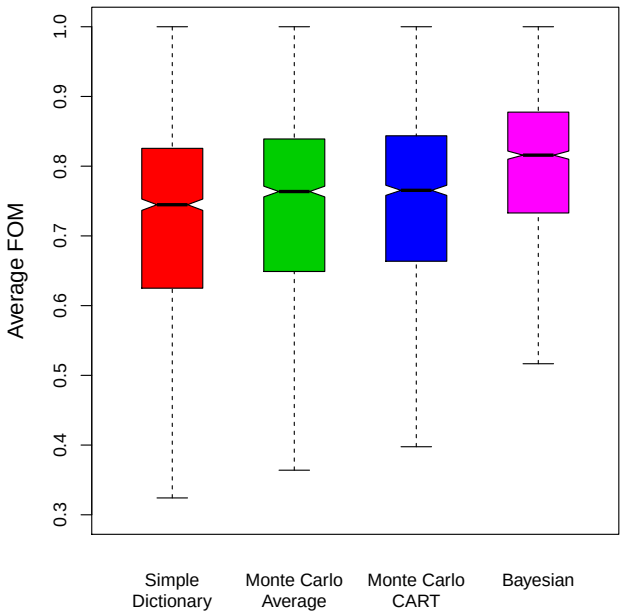

Figure 3: *Boxplots depicting average figure of merit for 1521 WSJ keywords for each model type.*

Table 1: *Comparison of figure of merit based on 1) average over all 1521 keywords, 2) average over subset of keywords which scored in the lowest 10th percentile using simple dictionary model.*

model type	mean FOM all words	mean FOM dict lowest 10%
Simple dictionary	0.7040	0.322
Monte Carlo (average)	0.7253	0.385
Monte Carlo (CART)	0.7331	0.413
Bayesian	0.7905	0.605

4. Conclusions

In our previous work on MAP estimation of whole-word acoustic models [9] we demonstrated that Bayesian approach to estimating phone timing models provided significant gains in keyword spotting performance in the case that few keyword examples are available. In that work, the prior model of phone timing used in MAP estimation was based on the simple dictionary model. The motivation for this work was to assess the gains possible by considering more sophisticated prior models. By incorporating a Monte Carlo approach to estimating phone-timing distributions, we were able to obtain a 4.2% relative improvement in average FOM compared to using a simple dictionary model. While modest in absolute terms, this gain represents 33% of the difference in performance between simple dictionary and MAP-estimated models.

5. References

[1] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M., "Speech recognition with primarily temporal cues," *Science,* 270(5234), pp. 303–304, 1995.

[2] Drullman, R., Festen, J., and Plomp, R., "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, 95(5), pp. 2670–2680, 1994.

[3] Tallal, P., Miller, S., and Fitch, R. H., "Neurobiological basis of speech: a case for the preeminence of temporal processing," *Annals of the New York Academy of Sciences*, 682(1), pp. 27–47, 1993.

[4] Klatt, D. H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," *The Journal of the Acoustical Society of America,* 59(5), pp. 1208–1221, 1976.

[5] Van Santen, J. P., "Assignment of segmental duration in text-to-speech synthesis," *Computer Speech and Language,* 8(2), pp. 95–128, 1994.

[6] Riley, M. D., "Tree-based modeling for speech synthesis," in *Proc. of The ESCA Workshop on Speech Synthesis,* pp. 229–232, 1990.

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

[8] Kintzley, K., Jansen A., Church, K., and Hermansky, H., "Inverting the Point Process Model for Fast Phonetic Keyword Search," in *Proc. of INTERSPEECH,* 2012.

[9] Kintzley, K., Jansen A., and Hermansky, H., "MAP Estimation of Whole-Word Dictionary Priors," in *Proc. of INTERSPEECH,* 2012.

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

[11] Levinson, S., "Continuously variable duration hidden Markov models for automatic speech recognition," *Computer Speech and Language,* 1(1): pp. 29–45, 1986.

[12] Pinto, J., Yegnanarayana, B., Hermansky, H., and Magimai-Doss, M., "Exploiting contextual information for improved phoneme recognition," in *Proc. of ICASSP,* pp. 4449–4452, 2008.

[13] "Carnegie-Mellon Pronouncing Dictionary," `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[14] Burshtein, D.,"Robust parametric modeling of durations in hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, 4(3), pp. 240-242, 1996.

[15] J.R. Rohlicek, W. Russell, S. Roukos, H. Gish, "Continuous hidden Markov modeling for speaker-independent word spotting," in *Proc. of ICASSP,* pp. 627-630, vol.1, 1989.