# Modeling of Vowel Duration in Malayalam Speech using Probability Distribution

*Deepa P. Gopinath, Veena S.G.*

College of Engineering, Trivandrum
University of Kerala, India
{deepapgopinath; veenasg83}@gmail.com

*Achuthsankar S. Nair*

Centre for Bioinformatics
University of Kerala, India
sankar.achuth@gmail.com

## Abstract

The first part of the work detailed in this paper, analyzes the probability distribution of vowels in Malayalam, an Indian language. Different probability distributions are fitted to the duration values of each vowel and the best fit is determined using quantile-quantile plot. The probability distribution of duration values, in accordance with the different factors affecting durations are also analyzed. In the second part, based on the results of the distribution fitting, a duration model is developed. It predicts duration as that, which maximizes the conditional probability, for the given set of factors. The model has the advantage that it can be expanded easily to include more factors, without affecting the existing structure. The performance of the model is evaluated using root mean square error (RMSE) and correlation between predicted duration and actual duration. The objective evaluation of the model gave an RMSE of 13.2 ms and a correlation of 0.80.

## 1. Introduction

Prosody is an important factor for a high quality text to speech (TTS) system. Prosody refers to the supra segmental features of natural language (such as rhythm and intonation) that are used to convey linguistic and paralinguistic information (such as emphasis, intention, attitude and emotion). To synthesize intelligible and natural sounding speech, it is essential to create good prosodic characteristics.The main prosodic features include duration of phonemes, intonation patterns and stress.The duration of speech segments vary dynamically in continuous speech, giving the rhythm and naturalness to speech. To produce natural sounding speech, text to speech synthesis system should capture the durational characteristics properly. Since prosodic characteristics are language specific, detailed analysis has to be carried out to capture the durational knowledge for each language.

In continuous speech a large number of factors affect the duration of basic units. These include phonological, positional and contextual factors. The duration of speech segments are influenced by (i) position of speech segment in the word, (ii) position of word in the sentence, (iii) phrase and sentence boundaries, (iv) preceding and succeeding speech segments etc. To predict the duration of speech segments, the effect of these factors have to be analyzed in detail.

Based on the analysis, duration models are developed to predict the duration of speech segments. The duration modeling methods can be divided into two categories-rule based and statistical. The most prevalent rule based duration model is a sequential rule based system proposed by Klatt [1]. In this model, based on a set of rules, the duration of a segment is modified. The statistical models include Sum of Product model [2], neural network model [3], CART [4] and SVM [5] etc.

In this work a duration model is developed based on the analysis of probability distribution of vowels in Malayalam, an Indian language spoken by around 35 million speakers. The first part of the work analyzes the probability distribution of vowels. The different distribution functions such as normal, lognormal, gamma and Weibull distributions are fitted on the duration values of vowels in the database and the best fit is determined using quantile-quantile (qq plot). The probability distribution of vowels depending on different factors are investigated to observe the effect of factors on duration values. In the second part, a probabilistic duration model is developed based on the analysis.

## 2. Distribution fitting

The database consists of a total of 200 sentences (2030 words and 19468 phonemes) from malayalam news bulletins.The sentences are manually segmented into words and phonemes. From the segmented database, the duration of phonemes are measured, which is used for the analysis of the factors that affect the phoneme duration.

The Malayalam language has 5 short vowels ( /a/, /i/, /u/, /e/, /o/), 5 long vowels (/aa/, /ii/, /uu/, /ee/, /oo/) and 2 diphthongs (/ai/, /au/) [6]. The duration values of each vowel are separately tabulated and the data are fitted onto different probability distributions (Normal, Lognormal,Gamma and Weibull distributions). The quantile-quantile plot (qq plot) is used to determine the best fit for the duration of phonemes. The qq plot shows the relationship between the quantiles of expected distribution and actual data. The straight line shows the quantiles of expected distribution. The quantiles of data are shown on the same plot. If the data have the same underlying distribution as the expected distribution, quantiles of the data will be quite close to the straight line. The duration data are further classified according to positional and contextual factors. The distribution of duration values according to different factors is further investigated. The standard deviation of the distribution for the whole data are compared with the standard deviations, when the data are grouped according to different factors.

## 3. Results and Discussion

### 3.1. Probability distribution of vowels

The pdf fitting of duration values of vowel /a/ considering all instances of the vowel, using normal, lognormal, gamma and weibull distribution is depicted in figure 1. Figure 2 gives the qqplots for the normal, lognormal,gamma and weibull distribution of the vowel /a/. From the qqplot it can be seen that the gamma distribution provides a best fit for the duration of

phonemes. All other vowels gave the same result. The gamma

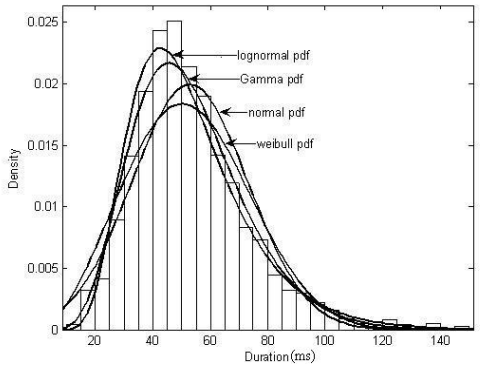
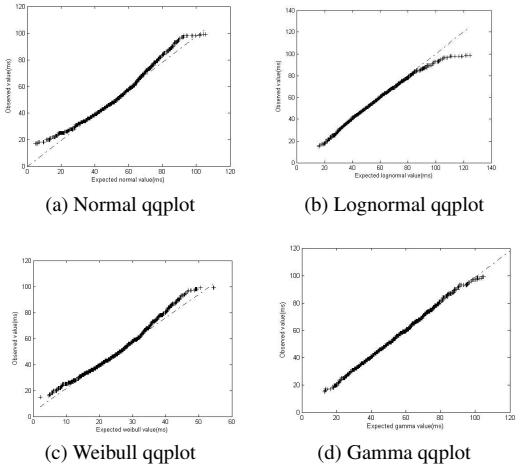



Figure 1: *Normal,Lognormal,Gamma and Weibull distributions of the duration of vowel /a/*

distribution of the short vowels /a/,/i/, /u/, /e/, /o/ are given in figure 3. The standard deviation is fairly large for all the vowels which means that duration of vowels vary significantly during continuous speech. Figure 4 shows the relationship between the duration of short vowel /a/ and long vowel /aa/. The mean duration of /a/ is 53.05ms and that of /aa/ is 98.67ms. The mean of duration of long vowels in all cases is around 1.6 to 1.85 times the duration of corresponding short vowels. Figure 5 shows the distribution of duration of vowel /a/ in different positions (begin,middle, end) in a word. The mean duration of vowels in different positions are significantly different from each other. Also when the position is fixed, the standard deviation of the vowel duration reduces. Figure 6 depicts the distribu-



(a) Normal qqplot (b) Lognormal qqplot

(c) Weibull qqplot (d) Gamma qqplot

Figure 2: *Normal,Lognormal,Weibull and Gamma qqplots for the duration of vowel /a/*

tion of duration of vowel /a/ (i) considering all instances, (ii) when grouped according to position(fourth syllable), (iii) when grouped according to position and succeeding consonant(fourth syllable, voiced labials as succeeding phoneme) and (iv) when grouped according to position,succeeding and preceding consonant(fourth syllable,voiced labials as succeeding phoneme, unvoiced velar as preceding phoneme). The standard deviation

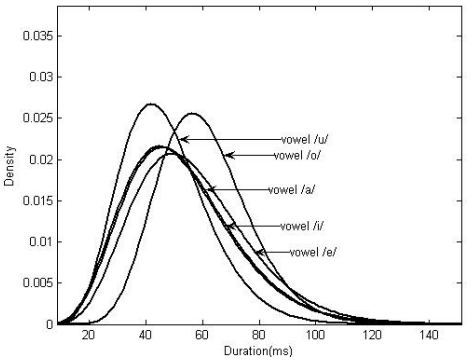



Figure 3: *Gamma distribution of vowels /a/, /i/, /u/, /e/, /o/.*

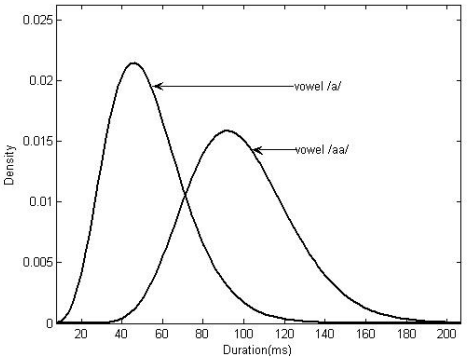



Figure 4: *Gamma distribution of short and long vowel.* $\mu/aa/ = 1.8\mu/a/$

of the above groups are plotted in figure 7. From the distribution plot and the standard deviation plot, it is evident that as the factors are jointly fixed, the standard deviation considerably reduces (from 16.80ms to 5.67ms). This means that when more factors are considered, the standard deviation of the distribution for each particular set of factors or feature vector will reduce further. Hence the mean value can be used to represent the duration of phonemes corresponding to that particular feature vector.

### 3.2. Probabilistic duration model

A duration model based on probability distribution is developed for predicting the duration of phonemes. The duration values are grouped according to different factors($f_1, f_2..f_n$) and the probability distribution function (pdf) is obtained corresponding to each feature vector, f=$f_1, f_2..f_n$. The mean of the distribution can be used to represent the duration of phonemes for that particular feature vector. Predicted duration $d_p$=E(p(d/f=f$_1, ..f_n$)).

When the duration of phonemes are predicted for each feature vector in this manner, the standard deviation corresponds to the root mean square error(RMSE), which is the objective measure used for evaluating performance of duration model. If sufficient number of factors are analysed, it is possible to predict the duration with high accuracy. But as the number of factors increase, the corresponding number of distributions which have to be evaluated also increases exponentially( as many as the num-

ber of different feature vectors). This problem can be solved by finding the conditional probability $p(d/f_1, f_2, f_3...f_n)$ from the conditional probabilities of duration with respect to each factor $(p(d/f_1), ...p(d/f_n))$ as given by equation (6).

$$p(f_i/d) = \frac{p(d/f_i)\, p(f_i)}{p(d)} \tag{1}$$

$$p(f_1 f_2/d) = \frac{p(d/f_1)\, p(f_1)\, p(d/f_2)\, p(f_2)}{p(d)^2} \tag{2}$$

$$p(f_1 f_2 ... f_n/d) = \frac{\prod_{i=1}^{n} p(d/f_i)\, p(f_i)}{p(d)^n} \tag{3}$$

$$p(d/f_1, f_2, ...f_n) = \frac{p(f_1, f_2, ...f_n/d)\, p(d)}{p(f_1, f_2 ... f_n)} \tag{4}$$

$$p(d/f_1, f_2, ...f_n) = \frac{\prod_{i=1}^{n} p(d/f_i)\, p(f_i)\, p(d)}{p(d)^n\, p(f_1, f_2 ... f_n)} \tag{5}$$

$$p(d/(f_1, f_2, f_3, ..., f_n)) = \frac{\prod_{i=1}^{n} p(d/f_i)}{[p(d)]^{n-1}} \tag{6}$$

The occurrence of factors $f_1, f_2, ..f_n$ are assumed to be independent of each other. The number of distributions to be evaluated, now reduces to number of factors times the number of different values taken by each factor (for example, if $f_n$ takes 8 values, 8 distributions are to be evaluated for that factor). This considerably reduces the complexity of the duration model.

The predicted duration is the duration value that maximizes the probability p(d/($f_1, f_2, f_3, ..., f_n$)). A small amount of gaussian noise ( $\epsilon$ ) is added to this duration value to reflect the high degree of variability in speech production [7]. Hence the predicted duration can be written as

predicted duration=duration predicted by the model + $\epsilon$.

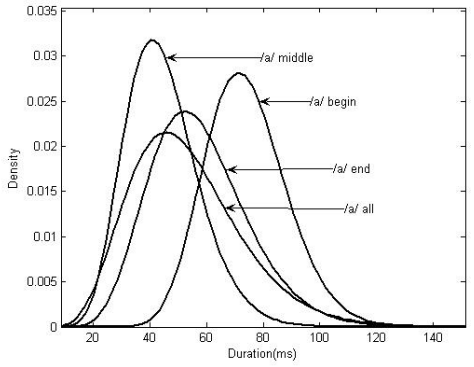



Figure 5: *Gamma distribution of duration in begin,middle,end and all position for vowel /a/. /a/ all: ($\mu$ =51.75ms, $\sigma$=16.80ms); /a/ begin: ($\mu$=76.13ms, $\sigma$=9.53ms); /a/ middle:($\mu$=44.65ms, $\sigma$=10.34ms); /a/ end:($\mu$=53.25ms, $\sigma$=13.45ms);*

Figure 8 illustrates the probabilistic duration model as a tree diagram. At level 1 only one factor is considered. The predicted duration corresponds to that factor alone. At each state,

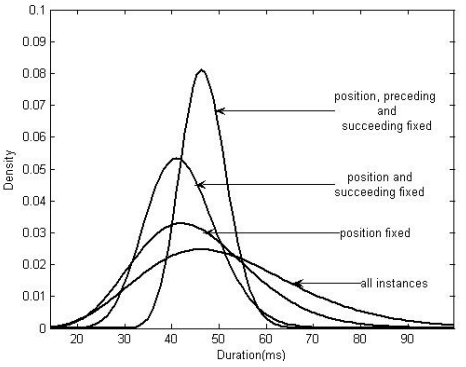



Figure 6: *Gamma distribution of vowel /a/ in fourth syllable position and in different context*

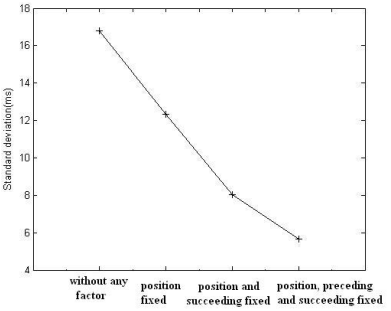



Figure 7: *$\sigma$ of vowel /a/ grouped according to positional and contextual factors.1:Duration value of /a/ in all instances of position and context. 2:Duration value of /a/ in the fourth syllable position. 3:Duration value of /a/ in the fourth syllable with succeeding fixed. 4: Duration value of vowel /a/ in the fourth syllable position with preceding and succeeding fixed*

the emitted probabilities are the conditional probabilities at that level. The different states in a level corresponds to the different values taken by the factor. As we proceed to the higher levels more factors will be taken into account. The advantage of the model is that, more factors can be added to the model easily without changing the existing structure by just adding more levels.

### 3.3. Performance evaluation

The performance of probabilistic duration model is evaluated using the root mean square error (RMSE) and the correlation

Table 1: *Correlation between actual and predicted duration*

| factors | correlation |
|---|---|
| position fixed | 0.75 |
| position and preceding fixed | 0.77 |
| position, preceding and succeeding fixed | 0.80 |

between the predicted duration and actual duration. The duration model is trained with a training data of 12648 phonemes and evaluated with test data of 6265 phonemes. Figure 8 shows

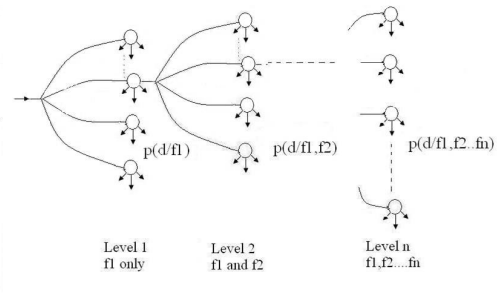



Figure 8: *Tree diagram representation of the probabilistic duration model*

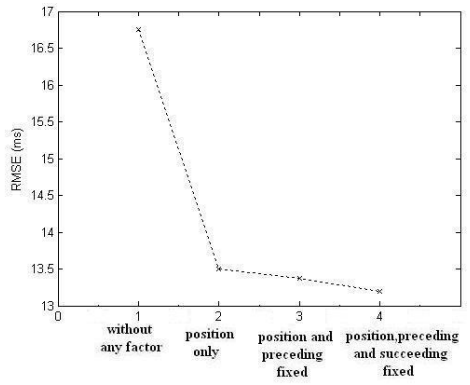



Figure 9: *RMSE of prediction with different factors fixed*

the reduction in the RMSE when the factors are fixed. Table 1 shows the correlation values with different factors fixed. The root mean squared error (RMSE) of prediction obtained is 13.2 ms and the correlation between the actual and predicted duration is 0.80.

## 4. Conclusion

The probability distributions of vowels in Malayalam speech are investigated. Different probability distribution functions were fitted onto the duration data and the best fit was determined using qq plot. It was found that the gamma distribution gives the best fit for the duration of phonemes. The effect of the different factors affecting the duration was investigated by grouping the duration data according to each factor and finding the probability distribution for each group. It was observed that as we consider more factors jointly the standard deviation of distribution reduces considerably. Hence the expected value of the distribution can be used to represent the duration of phonemes in that group.

In the second part of the paper, a probabilistic duration model is developed. The model predicts the duration of phonemes corresponding to each feature vector, by finding the duration that maximizes the conditional probability distribution of duration given that feature vector. The performance of the model is evaluated by the RMSE and the correlation between actual and predicted duration. When the factors are fixed, the RMSE of the model is reduced. The RMSE of the model can be reduced further by considering more number of factors.

The advantage of the model is that it is easy to expand the model, without affecting the structure to include more factors. Additional factors can be included by finding the conditional probability distribution of duration given that factor and adding more levels to the model structure. Another advantage is that the database required is not very large.

## 5. References

[1] D.H. Klatt., 1976. Linguistic uses of segmental duration in English. *Journal of Acoustic Society of America* vol.59, 1209-1221.

[1] D.H. Klatt., 1976. Linguistic uses of segmental duration in English. *Journal of Acoustic Society of America* vol.59, 1209-1221.

[2] Jan P. H., 1994. Assignment on segmental durattion in Text to Speech Synthesis. In *Computer speech and language*, J.Ph. van Santen (ed.).vol.8 95-128.

[3] K.Srinivasa Rao and B.Yegnanarayana., 2007. Modeling syllable duration in Indian languages using Neural Networks. *Computer Speech and Language*, vol.21 282-295.

[4] N.Sridhar Krishna and Hema a. Murthy., 2004. Duration modeling of indian languages Hindi and Telugu'*ISCA speech synthesis workshop*.

[5] K.Srinivasa Rao and B.Yegnanarayana., 2005. modeling syllable duration in indian languages using support vector machines'*In proceedings ICSLP*.

[6] R.E. Asher and T.C. Kumari, Malayalam. *routledge London and Newyork*.

[7] Michael Walsh.;Hinrich Schitze.;Berd Mobius.;Antje Scheweitzer., 2007. An exemplar theoretic account of syllable frequency effects'*ICPHS*.