Proceedings of the Institute of Acoustics

A MULTIPLE SPEAKER PHONEME DURATION MODEL

Christine M. Tuerk and Anthony J. Robinson

Cambridge University Engineering Department Trumpington Street, Cambridge, CB2 1PZ.

1. INTRODUCTION

A satisfactory segmental duration model is a very important part of many speech applications. Duration serves as one of the prosodic cues which carries information about the underlying content of a spoken phrase or sentence. Duration is an important indicator in identifying segments of an utterance and thus plays a large role in perceptual theory. Automatic speech recognition systems are improved by a good understanding of durational phenomena. A durational model is also an essential part of any speech synthesis system. An inadequate model will certainly degrade the quality of the output speech and will detract from the strengths of a synthesis application.

The importance of duration can be seen from the work of Klatt in [1]. He performed a number of experiments to aid in determining the contributions of duration to perception. Klatt stated, "it may be hypothesized that segmental timing contributes to the perception of constituent structure and phrasal and lexical stress patterns. In addition, the duration pattern reflects the speaker's mood, speaking rate, and the locations of emphasized material. Finally, duration serves as a cue to the phonetic identity of many segment types." [1] His experiments led to conclusions of duration playing a primary role as a perceptual cue in distinguishing between long and short vowels, voiced and voiceless fricatives, phrase-final and non-final syllables, voiced and voiceless postvalic consonants, stressed and unstressed vowels, and emphasis and non-emphasis.

As shown by Klatt, duration plays a very important role in perception, and thus, a good duration model is essential for many speech applications. In this paper, we seek to develop such a model. Section 2 will review prior work in this area. Section 3 will develop a model based on the TIMIT database with Section 4 reporting on results in comparison with other durational models. Section 5 will offer conclusions.

2. RELEVANT WORK

The duration of allophonic segments in spoken speech is influenced by a great number of interrelating factors. At a low level, phonetic context may effect the duration of neighbouring allophones and lexical stress may lead to longer durations. At a higher level, syntactic phenomena will contribute to prosodic boundaries within a sentence. Further, semantic variables may also influence areas of emphasis and speaking rate. In short, predicting the duration of allophonic segments requires accounting for a number of interrelating factors. Modeling all these interrelating factors is a difficult task and many different approaches have been used. The two most common approaches are parametric and non-parametric models. Parametric models include additive-multiplicative constructs while non-parametric models include tabular approaches. Although additive-multiplicative constructs give good insight into the underlying processes governing duration, they are hard to formulate due to the difficulty of separating the effects of interrelating factors. Tabular approaches are very useful for speech applications, but require large amounts of data for adequate estimation.

Speech synthesis has served as an important motivator for the development of durational models. Differentiation must be made between producing an adequate model for synthesis and producing a model which

[1] Klatt, "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence", pg.1217

accurately predicts actual durations of naturally spoken speech. A durational model that is sufficient to serve as a synthesis model will not necessarily be a good predictor of actual speech. This is due to the fact that a synthesis model hypothesises one set of durations for a given utterance, while, from natural speech, it is known that a given sentence or phrase can be spoken in many acceptable ways. However, it is reasonable to suppose that a model which accurately predicts the durations of natural speech will serve as a very good synthesis model.

An example of an early parametric durational model is given by Coker's rule based articulatory synthesis system [2]. The durational model consisted of a set of rules combined with tabular data and arose from the study of 20 minute passages provided by three different speakers. Coker showed that vowel duration is effected by stress and the identity of the consonants following the vowel. These effects are summarised in his vowel duration prediction equation

$$T = K_1 + S(K_2 + K_3C) \qquad (1)$$

where K_1, K_2, and K_3 are constants for a given vowel, S represents the effects of stress (including position of vowel in word and sentence, word prominence, sentence stress, and speech rate), C is the factor for the consonant following the vowel, and T is the estimated duration. The 20 minute passages were used to produce tables of values for each of the variables in Equation 1. The developed model gave standard error deviations for one of the speakers as ranging from 11ms to 29ms depending on the conditions. Greater detail of the vowel model and its development are found in [3]. The model's consonant durations arose from studies described in [4]. The factors found to affect consonantal duration included context, content/function difference of parent word, position in relation to pauses, lexical stress, and position within the word. An additive model for consonantal duration was developed where a base duration of a given consonant was lengthened or shortened depending on the previously listed factors. Umeda reported that the model worked well for a wide variety of consonants and conditions.

Perhaps the most well-known and successful parametric durational model comes from the work of Dennis Klatt as used in the MITalk Text-to-Speech System. Klatt's model is an additive-multiplicative model whose rules are designed to match observed durations for a single speaker reading paragraph length material. The model details are given in [5]. The following formula summarises the model.

$$DUR = MINDUR + \frac{(INHDUR - MINDUR) \times PRCNT}{100} \qquad (2)$$

The duration (DUR) is calculated from modifications to an inherent ($INHDUR$) and minimum duration ($MINDUR$) for a given allophone. A series of rules are applied which modify the $PRCNT$ value used in Equation 2. These rules are implemented in the following form

$$PRCNT = (PRCNT \times PRCNT1)/100 \qquad (3)$$

where $PRCNT1$ varies according to the rule being applied. These rules fall under the classes of clause-final lengthening, nonphrase-final shortening, nonword-final shortening, polysyllabic shortening, noninitial-consonant shortening, unstressed shortening, lengthening for emphasis, postvocalic context of vowels, shortening in clusters, and lengthening due to plosive aspiration (an additive rule). For example, an emphasised vowel is lengthened by $PRCNT1 = 140$. The developed model was used to predict segmental durations of new paragraphs from the same single speaker from which the model was developed. The standard deviation was 17ms. In [1], experiments showed that only after changing segmental durations by more than 20ms were unnatural timing patterns reported. In other words, deviations less than 20ms should have little effect on perception. Klatt used his model to produce segmental durations for synthetic sentences. The sentences were compared with synthesised sentences taking durations identical to a naturally spoken utterance. Perceptual ratings of the model-derived duration sentences were very close to those taking natural durations, thus showing that Klatt's model was very suitable for synthesis applications.

Some recent research has focused on trying to refine and improve Klatt's formulation. For example, in [6] various data analysis methods are undertaken in order to find better functional combinations of the many interrelating factors which must be accounted for in an additive-multiplicative durational model. The work has highlighted areas of predictive weaknesses in Klatt-type models. Klatt's methodology is also being adopted by pioneers of speech synthesis in languages other than English. For example, in [7], 400 sentences from a single male Moroccan speaker are analysed. The resultant model modifies the inherent duration of a segment by a percentage obtained from applying rules. The developers claim the model can predict the data set with a standard deviation of just under 16ms.

Some durational model developers have chosen to avoid the difficulty of determining how multiple factors interrelate by adopting a non-parametric, statistical approach. In this approach, statistical calculations are made and the durational model is formulated in either tabular or decision tree format. If a large enough data set is available and appropriate categorical parameters determined, then these models can be very powerful and quite useful for synthesis systems. Such a model for Japanese was developed in [8] in which a 503 sentence database was analysed using control parameters which included segment position, part-of-speech, context, accentedness and segment type. The resultant model yielded a 15.8ms standard deviation error for vowels and was used to produced natural sounding speech in a synthesis system. A similar effort for French synthesis was performed as described in [9]. Here, 150 French sentences were analysed to determine a tree structure model for durations. Various phoneme and word level features were accounted for including context, class, position in syllable and word, word nature and word length. Results of the model were used in listening tests which showed that listeners equally preferred natural durations and modeled durations in synthesised sentences.

Another important statistical study which will be used for comparison purposes is found in [10]. In this study Pitrelli and Zue examined data taken from the multiple speaker TIMIT database. The model was based on examining 2520 sentences by 504 different speakers and was developed through an automatic regression analysis of phoneme durations into a hierarchical discrete-variable tree. The regression modeling procedure was supplied with a large collection of features on which to build the tree. This collection included distinctive features of the current and immediate context phonemes, gemination, position in syllable, lexical stress, position of parent syllable in word, number of syllables in word, function/word classification, and location in relation to the following pauses or syntactic boundaries. The modeling procedure automatically selected the relevance of each feature. Performance of the resultant model on 630 test sentences was given as a standard deviation prediction error on average of 31ms for vowels and 26ms for consonants. Although these prediction errors seem large in comparison to work cited above, it must be noted that Pitrelli and Zue were working in the more difficult multiple-speaker domain.

3. PROPOSED MODEL

We propose to develop our model by studying the characteristics of the multiple speaker DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [11]. Although a model based on data from multiple speakers will contain both inter-speaker and intra-speaker variability, and hence, may have greater prediction error than single speaker counterparts, such a model can be advantageous. Single speaker models may simulate the characteristics of the given speaker but often show a decrease in performance when applied to novel speakers. A multiple speaker domain can provide a better, more robust generalisation of overall durational trends. The amount of data available for study in the TIMIT database is quite large (3696 training sentences from 432 speakers and 1344 test sentences from 168 other speakers) and we propose to develop a model through statistical studies. The proposed model will explore phonemic durations in terms of the combination of the variables context, stress, syllabic position, word position and phrase position.

Because segmental durations will vary with the speaking rate, it is desired to have a normalised duration measure for comparison purposes. Thus, a measure of speaking rate for each speaker is needed. The TIMIT corpus provides recordings of two identical sentences for each speaker. If it is assumed that the individual speakers used a fairly consistent speaking rate across all sentence recordings, then the two identical sentences should yield some indication of speaking rate. Comparison of the duration of selected carrier phrases within the 2 recordings led to the establishment of a relative speaking rate measure. This speaking rate measure was used to normalise segmental durations for each speaker. The results presented in Section 4 utilise these normalised durations and show improvement by approximately 1% over results using unnormalised durations.

The data for the durational model was obtained by analysing the training set sentences of the TIMIT database. The analysis involved performing lexical stress assignment with the use of a small rule-set and manual post-editing. Syllabic assignment was also performed through the use of letter-to-sound rules and manual post-editing with the assistance of a Webster's on-line dictionary. The results of the analysis associated a number of parameter values to each phonemic realisation within the database. These parameters are shown in Figure 1.

Binary valued variables:
1) Phrase Initial
2) Phrase Medial
3) Phrase Final
4) Word Initial
5) Word Medial
6) Word Final
7) Word Overlap*
8) Phrase Initial
9) Phrase Medial
10) Phrase Final
*As in a single [r] used in "her red..."
Trinary valued variable (0, 1, or 2):
11) Lexical Stress
Real valued variable:
12) Measured duration

Figure 1: List of parameters associated with each phonemic realisation

To simplify storage requirements, the values of the binary and trinary variables can be used to assign each phonemic realisation to one of 36 legal position/lexical stress category types. Contextual information is also stored in the form of the identity of the previous and following phonemes. 51 different phonemes (including silences) are possible. Analysis of the TIMIT database led to 132,658 allophonic training tokens (each consisting of left phonemic context, central phonemic context, right phonemic context, position/lexical stress category and normalised segmental duration measure). There exist 4,775,436 different possible combinations of the token variables (51 x 51 x 51 x 36 = 4,775,436). Although many of these combinations will rarely, if ever, be seen, the training data is still not extensive enough to realistically cover the remaining combinations. Thus, simple tabulation as a means of estimating model parameters is not wise. Further, a simple tabulation will also ignore close relationships between phonetic classes and category divisions that can be helpful in prediction. Thus, a predictive model which utilises a more complex search strategy was devised. In addition, an efficient data structure is needed.

As noted above, not all of the 4,775,436 possible combinations are actually used. Thus, a direct implementation of a look-up table would be grossly inefficient. In addition, the large amount of data necessitates a storage structure which allows for ease in retrieval. For each central phoneme, the data is stored in a set of nested binary trees as illustrated in Figure 2. A binary search through these trees allows access to the duration prediction (and the number of tokens on which the prediction is based) for the desired position/lexical stress category, right phonemic context, and left phonemic context used to reach the node. The structure is advantageous in that search time is in the order of $O(\log m) + O(\log n) + O(\log n)$ (where $m =$ the number of position/lexical stress categories and $n =$ the number of phonemic contexts). In addition, the structure saves on storage as the 132,658 training tokens lead to a set of trees with a total of only 28,783 terminal nodes.

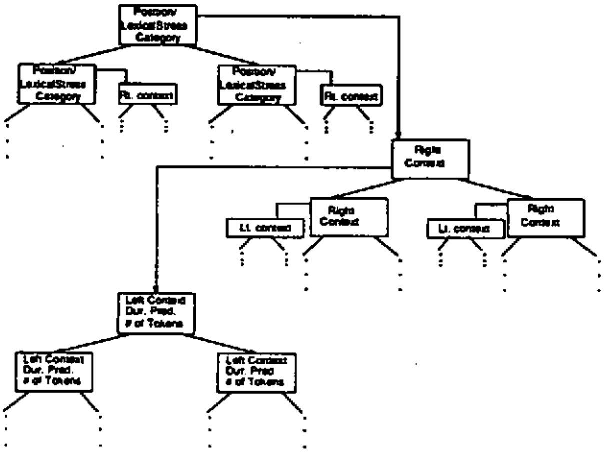
Position/ LexicalStress Category
Position/ LexicalStress Category
Rt. context
Position/ LexicalStress Category
Rt. context
Right Context
Lt. context
Right Context
Lt. context
Right Context
Left Context Dur. Pred. # of Tokens
Left Context Dur. Pred. # of Tokens
Left Context Dur. Pred # of Tokens

Figure 2: Duration Data Structure Tree

As mentioned above, a simple tabulation algorithm may ignore close relationships between phonetic classes and position/lexical stress categories. Such an algorithm may also perform poorly when faced with desired conditions which do not exactly match any item present in the data table. To improve on these weaknesses, an improved search algorithm was devised. This algorithm predicts segmental durations given a condition set (the segment identity, left and right context, and position/lexical stress category information). Prediction durations for test examples are found by searching through the data structure for conditions matching the test set. If the tree does not contain exact matching conditions, the scope of the search is broadened by exploiting the relationships in phonetic classes and position/lexical stress categories. Broadening the scope of a search involves looking at examples whose conditions are of a similar class. This requires defining what is meant by 'similar class'. In the case of the left and right phonemes, similar class refers to other phonemes which fall in the same phonemic class. The phonemic classes are: vowels, closures, bursts, nasals, fricatives, affricates, glides, liquids and silences. In the case of the position/lexical stress category, similar class is defined by comparing the Hamming distance (from the binary and trinary values given in Figure 1) between categories. Categories whose Hamming distances differ by only 1 are considered to be of the same category class.

The search procedure begins by looking for an exact match for the given condition set. If the search procedure fails to find an exact match, it will broaden the search by gathering information for examples in which 2 of the 3 parameters match exactly and third parameter is of the same class of the desired parameter. In addition, exact matches which have too few training tokens (< 5) to yield an unbiased estimate, are ignored in favour of a broader search. For example, if the test condition set for an /iy/ in the context of /n/ and /f/ being in an unstressed syllable medial position is not found in the tree, a search will be made to determine closely matching information. In this case, either the /n/ can be replaced by another nasal, the /f/ by another fricative, or the sentential position by stressed syllable medial examples. If this search yields multiple matches, a weighted average is used to obtain the predicted duration. If this wider search fails to gather enough information on which to make a prediction, an even broader search is used. Figure 3 gives the matching criteria as the scope of search is broadened. Basing a statistical prediction on only a few training samples is unwise. Thus, in all search levels, an estimate based on less than 5 training tokens will be rejected in favour of a broader search. Weighted averages of all tokens matching the conditions are used to generate the final predictions. In this way, predictions should better reflect general trends as opposed to

Search Level	Matching Criteria
1	Exact match of 3 given parameters
2	Exact match of 2 given parameters, class match of remaining parameter
3	Exact match of 1 parameter, class match of 2 other parameters
4	Class match of all 3 parameters
5	Exact match of 2 given parameters, any example for remaining parameter
6	Exact match of 1 parameter, class match of 1 parameter, any example of 1 parameter
7	Class match of 2 parameters, any example for remaining parameter
8	Exact match of 1 parameter, any example for remaining 2 parameters
9	Class match of 1 parameter, any example for remaining 2 parameters
10	Any example of 3 parameters (i.e. average for phoneme)

Figure 3: Algorithmic Search Levels

the characteristics of a few examples.

4. EXPERIMENTS AND RESULTS

The search algorithm described above was used as a segmental duration predictor on training and test sets. The training set consisted of the 132,658 tokens used to build the data tree. The test set was taken from additional TIMIT sentences spoken by different speakers than those who appeared in the training set. In addition, the texts of the sentences spoken in the test set do not appear in the training set, and thus, serve as a good measure of the predictor's ability to generalise to novel situations. Figure 4 shows the performance of the predictor on the training and test data broken down into phonetic classes. The lower prediction error on the training set reflects the bias of using the same data to both build and make predictions. The test set error is a more realistic measure of the algorithm's performance. The percentage error (percent deviation from the actual phoneme duration) for the model prediction was 31.4% on the training data and 35.1% on the test data.

In order to gauge the performance of the predictor, the results must be compared to other durational models. Performance quotes given in Section 2 are difficult to compare to because they are mostly given for much smaller tasks and for single speakers. A more valid comparison would come from looking at the results of a different predictor on the same data. With the use of the MITalk [12] synthesis system, which incorporates the Klatt durational rules, the predicted durations for a phonetic segment could be gathered. Because MITalk is implemented as a modular system, data files containing the actual spoken phonemes of the candidate sentences, and other information required by MITalk, could be prepared. These data files could then be given to the PROSOD module which predicted durations. Because MITalk is designed to model a single speaker, all results were compared with rate normalised durations. The MITalk rules predicted the training set with 40.7% error and the test set with 40.4% error, a full 5% higher than our search algorithm predicted. Figure 5 shows a comparison of prediction performances on the test sets. Each group of bars compares the performance of the predictors on a phonetic class. Within each group of bars, the white bar gives the new model performance and the horizontally hatched bar gives the Klatt performance (the solid bar will be explained in the following paragraph). In each category, the new model predictor gave significantly better segmental duration predictions than the Klatt rules.

It is also of interest to compare the performance of our new predictor model to another predictive model. Pitrelli and Zue [10] developed such a predictor with the object of applying the model to recognition tasks. Data for their hierarchical model was also taken from the TIMIT database although the division into training and test sets was not equivalent to the division used in our example. It is also unclear as to whether Pitrelli's test set contained only unseen sentence texts or whether the test set contained repeated texts but by different speakers. Results from Pitrelli were quoted as a root-mean-square prediction error of approximately 31ms for vowels and 26ms for consonants for test set prediction. Details of test set prediction of phonemic classes was given in the original paper in the form of a bar graph. Estimates from the bar graph are reproduced as the solid bars in Figure 5. Our model and the Pitrelli model perform roughly the same with vowel prediction (both approximately 31ms), however, our model performs significantly better over the remaining classes.

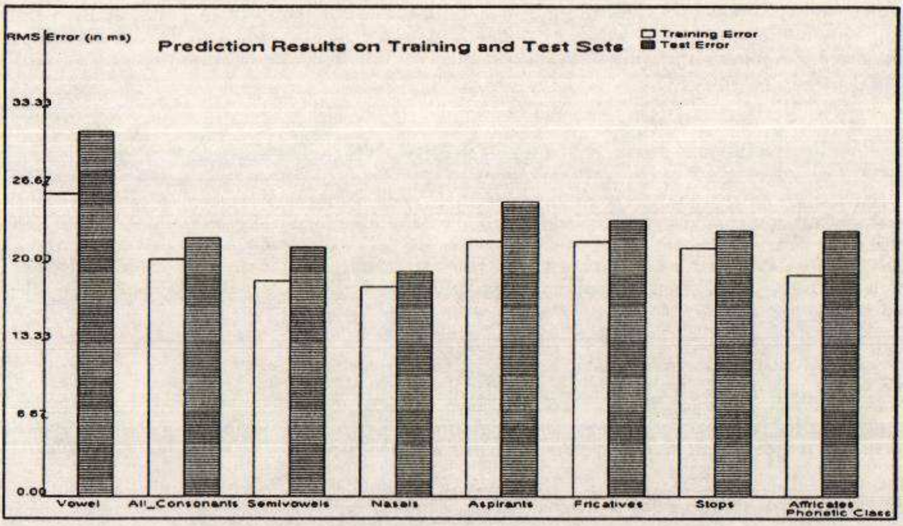

Figure 4: Results of predictor model on training and test data

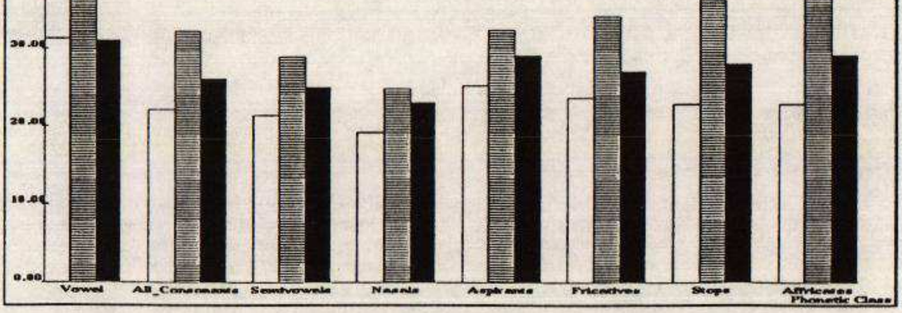

Figure 5: Comparison of tree model vs. Klatt rules vs. Pitrelli and Zue model

Over all the consonants, our model performed about 4ms better than Pitrelli's model (26ms to 22ms). Some of this improvement can be attributed to our algorithm's ability to utilise varying degrees of search scope, and thus multiple pieces of information, to obtain a prediction; this is in contrast to a hierarchical tree model which will only use information from a single terminal node to make its prediction. In addition, the new model is advantageous in that it utilises less features than the Pitrelli model and thus requires less effort in gathering the model parameters.

5. CONCLUSIONS

The extensive study of segmental durations in the multiple-speaker TIMIT database has led to the development of a new durational model. This model predicts the durations of naturally spoken speech significantly better than the Klatt durational model implemented in the MITalk synthesis system. Further, the new durational model also performs better than the Pitrelli durational predictor designed expressly for use in a multiple speaker domain. Thus, the developed model should be sufficient not only for use in synthesis applications, but also in wider ranging speech technology tasks.

REFERENCES

[1] D.H. KLATT. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, 59:1208–1221, 1976.

[2] C.H. COKER, N. UMEDA, and C.P. BROWMAN. Automatic synthesis from ordinary English text. *IEEE Transactions on Audio and Electroacoustics*, AU-21:293–298, 1973.

[3] N. UMEDA. Vowel duration in American English. *Journal of the Acoustical Society of America*, 58:434–445, 1975.

[4] N. UMEDA. Consonant duration in American English. *Journal of the Acoustical Society of America*, 61:846–858, 1977.

[5] D.H. KLATT. Synthesis by rule of segmental durations in English sentences. In B. Lindblom and S. Ohman, editors, *Frontiers of Speech Communication Research*, pages 287–300. Academic, New York, 1979.

[6] J.P.H. VAN SANTEN and J.P. OLIVE. The analysis of contextual effects on segmental duration. *Computer Speech and Language*, 4:359–390, 1990.

[7] S. BENAOUICHA, A. RAJOUANI, and M. ZYOUTE. Construction of an Arabic speech data base. Duration model of Arabic vowels. In *Eurospeech 91*, pages 541–544, 1991.

[8] N. KAIKI, K. MIMURA, and Y. SAGISAKA. Statistical modeling of segmental duration and power control for Japanese. In *Eurospeech 91*, pages 625–628, 1991.

[9] L. MORTAMET. Implementing duration expert rules into a text-to-speech synthesis system. In *Eurospeech 91*, pages 621–624, 1991.

[10] J.F. PITRELLI and V.W. ZUE. A hierarchical model for phoneme duration in American English. In *Eurospeech 89*, volume 2, pages 324–327, 1989.

[11] L.F. LAMEL, R.H. KASEL, and S. SENEFF. Speech database development: Design and analysis of the acoustic-phonetic corpus. In *Proceedings of the DARPA Speech Recognition Workshop*, pages 26–32, 1987.

[12] J. ALLEN, M.S. HUNNICUTT, and D. KLATT. *From Text to Speech: The MITalk System*. Cambridge University Press, 1987.