A HIERARCHICAL MODEL FOR PHONEME DURATION IN AMERICAN ENGLISH

John F. Pitrelli and Victor W. Zue
Department of Electrical Engineering and Computer Science and
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, U.S.A.

ABSTRACT

We are developing a model which predicts phoneme duration as a function of segmental and suprasegmental factors, with the objective of using it for speech recognition. Our goal is to account for the many duration effects, ranging from local phonetic to sentence-level, and to determine how accurately we can model segment durations for sentences drawn from a large database spoken by many speakers. Our approach is to develop a hierarchical structure of categorical distinctions based on discrete-valued variables representing attributes of a phoneme and its context. We choose this technique over additive or multiplicative models because duration effects often interact in a complex manner. In our procedure, two descendents of a parent node can be split using different variables, thus allowing us to model non-uniform interactions among factors. When tested on 630 sentences from 126 speakers not used for training, our models explain 60% of vowel duration variance and 55% of the consonant duration variance within manner classes, yielding a root-mean-square prediction error of approximately 31 ms for vowels and 26 ms for consonants.

INTRODUCTION

This paper describes an investigation into modeling segmental and suprasegmental effects on phoneme duration in continuous speech. The linguistic use of segmental duration for speech perception has been known for quite some time [3]. For example, certain vowels are inherently longer than others and the durational differences are important for distinguishing among them. Similarly, the voicing characteristics of stops and fricatives are often encoded in the temporal domain. Furthermore, the duration of one segment may cue the identity of a neighboring segment. A vowel tends to be shorter preceding an unvoiced consonant than preceding the corresponding voiced consonant. Segment duration can also potentially cue gemination, lexical stress and syntactic boundaries.

The association of duration patterns with these distinctions implies that duration cues may be useful for speech recognition. In practice, however, speech recognition systems typically distort the temporal dimension by time-warping procedures, or use only rudimentary duration models, such as one duration prediction per phoneme class. The main problem with using duration cues for recognition is that each duration pattern is obscured by influences ranging from phoneme-level to syntactic, semantic, and extra-linguistic effects. The duration patterns identified as potentially useful for speech recognition are sufficiently confounded by other factors to hinder their use.

Many studies have evaluated the factors affecting segmental duration one at a time. However, such studies may not lead to accurate modeling, because the factors affecting duration interact with each other in several complex ways. First, the degree of an effect is often contingent on values of other contextual variables. For example, the effect of obstruent voicing on the duration of the preceding vowel is enhanced in phrase-final position [3]. One effect can even reverse another; Crystal and House [1] reported that front vowels are shorter than back vowels preceding labial and dental consonants, but longer preceding velars. The existence of such interactions implies that a simple arithmetic model incorporating factors independently will not model all interaction accurately.

Second, the occurrences of effects are sometimes correlated. For example, the short duration of /b/ in the common word "the" could be attributed to the strident/non-strident distinction, the voiced/unvoiced distinction, or function-word reduction. Therefore, obtaining duration statistics independently for each of these effects in a continuous-speech database, and then compounding the measured effects into a model, will lead to inaccurate predictions. Instead, the modeling process must somehow evaluate such factors simultaneously.

Third, the set of possible values for one phoneme feature or contextual variable sometimes varies depending on the value of another. For example, if voicing and manner of the following consonant are two variables in a vowel model, a voicing distinction exists only for stops, fricatives and affricates. Another reason certain combinations of values do not exist in a continuous-speech database is sparse data. This type of interaction implies that the amount of a duration effect cannot always be apportioned individually to factors. Continuing the manner/voicing example, it is an arbitrary decision to count the difference between aspirants and nasals as a voicing difference or as a manner difference.

Past studies in comprehensive, explicit modeling of duration effects have generally used additive and/or multiplicative models for factor interactions [4,7,8]. We believe that with the availability of large amounts of data and computing resources, we may be in a position to explore the complex interactions among these linguistic factors. The goal of this study is to develop and evaluate a procedure which models...
duration effects explicitly, and also accounts for the arithmetically non-uniform interactions among the factors affecting duration. The scope of this study is limited to discrete-valued factors; thus, speaking-rate, a continuous-valued parameter, is not investigated. The study is also restricted to sentence-level and more local effects. Therefore, semantic effects, paragraph-level effects, and effects of psychological and physical state of the speaker are not studied.

EXPERIMENTAL PROCEDURE

Database

The database for this study is 3150 sentences from the TIMIT database, consisting of five sentences each spoken by 630 native speakers of American English. These sentences were drawn from a corpus of 450 sentences designed to contain all phonemes in a wide variety of phonetic environments, and the speakers were chosen to include several regional dialects. More details on this database are provided in [5].

The data by 504 (80%) of the speakers are used as training data; model development is based on only these sentences. The remaining 20% of the database has been held out as test data, for the purpose of evaluating robustness of the model.

Segmentation

Strictly speaking, the duration of a phoneme is difficult if not impossible to infer from the acoustic signal. Like many others before us, we have nevertheless adopted the operational definition of a phoneme to be the segment between major acoustic landmarks, in the belief that these acoustic segments delineate regions where the influence of one phoneme dominates. This definition is often valid for fricatives and stops, but may be questionable for semivowels, since the transitions in the latter case are often quite gradual.

The phoneme durations are obtained using the following procedure. Sentences are transcribed phonetically and aligned to the corresponding waveform using a semi-automatic alignment system developed by Leung [6]. A phonemic transcription for each word is looked up in a lexicon; these two transcriptions are aligned automatically, and the alignment is verified manually [2]. Usually, a phoneme is associated with a single phone, but in the case of stops and affricates, the closure and release portions are labeled with separate phone tokens. In these latter cases, the duration of a phoneme is taken to be the sum of the closure and release durations. In the case of geminates, on the other hand, the two underlying phonemes are both assigned to the same phone, but marked as geminates. More details on the transcription procedures for this study can be found in [9].

Model development procedure

To account for the irregular factor interactions, a hierarchical model is explored in this study. The model is in the form of a decision tree based on the variables affecting duration. An example for fricatives appears in Figure 1. The input to the modeling procedure is a data set of phonemes labeled with their durations and values for a list of discrete-valued variables representing attributes of a phoneme and its context. For each of the explanatory variables, an F-ratio is computed for the categorization of durations according to its values. The F-ratio measures inter-category differences relative to intra-category spread, thereby indicating how effectively the variable separates different duration categories. Thus, this measure can be used to infer how strongly the variable influences durations of the tokens in the data set. The data set is split based on the variable with the highest F-ratio. For the example shown in Figure 1, the most important variable used for splitting the complete data set is the feature [voicing]. Each resulting data subset is subjected to the same splitting procedure, beginning with computation of F-ratios for the remaining variables on the subset. A data set is not split when it has fewer than a prescribed number of tokens or when all variables have statistically insignificant F-ratios. In this study, the minimum size for a node to be split has been empirically chosen to be 400 tokens. The significance level used is .05.

Once a model has been generated, it can predict the duration of a given phoneme by proceeding through the tree according to the splitting rules until a terminal node is reached. The total prediction consists of a predicted duration and an indication of the reliability of this prediction, which consists of the standard deviation of training-data durations and the number of training tokens at the node.

The hierarchical modeling procedure deals effectively with all three kinds of interaction discussed earlier. In the cases in which the value of one variable influences the magnitude of another effect, once the first variable is used for a split, the resulting data subsets in general are split according to different variables. In this way, the procedure can model enhancement and reversal interactions. The model uses variables only where they are significant to duration, and so it determines which variables are important to duration and under what circumstances.

Similarly, correlation of occurrences of effects does not pose a problem for this modeling procedure. Once the variable representing an effect is used in a split, the data subsets
subordinate to that split will just have different proportions of tokens of the categories defined by the variable representing any correlated effect.

Finally, the dependence of the set of possible values of one variable on the value of another is no problem. When a variable is used for a split, the data set is simply split according to the full set of values of that variable appearing in the node’s population. Should this split be subordinate to another split in such a way as to eliminate values, these values just do not appear in the lower split. For example, a manner split in the voiced subtree subordinate to a voicing split will not have an aspirant branch.

One issue which needs to be addressed is whether the modeling procedure should be run once to generate one model for all phonemes, or separately for each phoneme, or on groups of phonemes at a time. We group the phonemes by manner class, for three reasons. First, many of the phoneme features and contextual variables relevant to duration vary among manner classes. For example, voicing is a variable for fricatives, stops and affricates, but not the other classes. Therefore, it is convenient to separate phonemes initially by manner, and then develop a model for each manner class, using the relevant set of variables. Second, the set of variables tends to be fairly constant within each manner class. Grouping the phonemes together within these classes creates a larger data set, allowing for the development of a more complex model before sacrificing robustness due to sparse data. Finally, analyzing a manner class at a time allows for the comparison of the importance of the features distinguishing phonemes within a manner class, such as voicing, to the importance of the other factors in the model, such as stress.

Another feature of this modeling procedure is that it completely specifies the model. Once the data set and a set of candidate factor variables are selected, the algorithm generates one particular tree. Furthermore, it is not necessary to pre-determine which factors will actually appear in the model. If a candidate factor supplied to the modeling procedure is never significant to the durations in any data subset, that factor will be eliminated properly by simply never being used in the model. Therefore, variables representing disputed as well as agreed-upon effects can be evaluated in the modeling procedure.

Since the modeling procedure can automatically reject irrelevant effects, one can start with a large collection of candidate variables based on our speech knowledge. Typical variable sets used in this study include distinctive features of the phoneme and the two adjacent phonemes, gemination, position of the phoneme in the syllable, lexical stress, position of the syllable in the word, number of syllables in the word, the function-word distinction, and proximity to following pauses or syntactic boundaries.

EVALUATION OF RESULTS
Accuracy and robustness of model
To score the model’s effectiveness in explaining duration data variability, we use root-mean-square (RMS) prediction error. The primary reason for this choice is that this statistic is directly comparable to standard deviation of the data set. Standard deviation of a data set can be interpreted as the RMS prediction error of a null model whose predicted duration for a data set is simply the mean duration in the set. RMS prediction error can be viewed as a model standard deviation, and model effectiveness can be described in terms of a reduction in standard deviation.

For the sake of comparison, the RMS error is also computed on a phoneme-based model, in which each phoneme’s duration is modeled by the phoneme’s average duration in the training data set. Robustness can be evaluated for a model by comparing its performance on the training and test data sets. Figure 2 shows RMS error for the training and test data sets, for phoneme-based models and for the hierarchical models. The hierarchical model performs consistently better than the phoneme model, which is to be expected because the variables representing phoneme identity are just a subset of the many variables available to the modeling procedure. The model is quite robust; RMS error on test data is not much larger than on training data.

An add-on procedure was used to determine how many variables are needed to approach the performance of the full model. It was determined that in general, approximately 90% as much duration variance within a manner class could be explained by the first five variables accepted by the add-on procedure as was explained by the full model.

Irregularity of factor interaction
As discussed above, the choice of our modeling procedure is based on the belief that some factor interactions are not

![Figure 2: Accuracy of models. The black bars show “total standard deviation” within the manner class, defined to be the RMS difference between token duration and the manner-class training-data mean. The other bars show RMS error for the phoneme-based model and RMS error for the hierarchical model.](image-url)
arbitrarily “regular.” Arithmetically regular interactions, such as additive or multiplicative interactions, imply that the importance of one factor is independent of the values of other factor variables. The format of the models produced in this study suggests a simple estimate of interaction “irregularity” for discrete-valued explanatory variables.

Consider the root node of the model depicted in Figure 1, for example. The fact that its subnodes are split according to two different variables, prepausal and sentence-final, indicates that the ranking of importance of these two variables depends on the value of the voicing variable. Therefore, an enhancement interaction exists between voicing and either prepausal or sentence-final lengthening. If no such enhancement existed, then the ranking of importance of variables would be alike in the voiced and unvoiced subnodes, and so they would be split using the same variable.

By comparing the number of nodes at which the most important variables interact regularly to the number of nodes with subnodes split according to different variables, we can estimate the prevalence of irregular interactions. This comparison can be performed by examining the nodes with multiple non-terminal children. In the eleven models developed, a total of 128 nodes have multiple non-terminal children, and in 106 (83%) of these cases, the children were split according to different variables. This result indicates that a large majority of interactions among the most important factors at nodes include enhancement to the extent that the value of the most important variable affects which is the next most important. This indicates strongly that an additive or multiplicative model would have to incorporate multiple-factor terms to correct for inaccuracies which would result from simply combining single-factor terms arithmetically.

CONCLUSIONS

A hierarchical duration model can take into account a variety of segmental and suprasegmental variables, while imposing no assumptions that their interactions follow any arithmetical pattern. Our model explains considerably more variance than a simple one-value-per-phoneme model. The model is robust, as its performance on speech data not included in the model development procedure is approximately as good as the training data performance. A further advantage of this hierarchical approach is that it does not require pre-selection of which variables appear in the model; these choices are made automatically. Given a set of variables, the modeling procedure is well-defined and generates one particular tree.

Analysis of nodes with multiple non-terminal children confirms that enhancement interactions are sufficiently frequent to warrant the choice of a model which does not impose any assumptions of independence among the factors.

While many variables are used by the model, the add-on analysis shows that the vast majority of the explanatory power of the models can be achieved by a relatively small subset of the variables. This fact is due to correlations among the factor variables and the differing importance of the factors.

Intended future elaborations on this work include expanding the model format to incorporate continuous-valued explanatory variables, such as speaking rate. We also intend to test the usefulness of the model for recognition by creating a simple duration-only classifier using the model to make some of the distinctions for which duration cues should be useful, such as fricative voicing.

ACKNOWLEDGEMENT

The development of hierarchical models was greatly facilitated by the SEARCH program created by Mark Randolph. Our discussions with him on issues in hierarchical modeling have played a significant role in setting the direction of this research. This research was supported by DARPA under Contract N00014-82-K-0727, monitored through the Office of Naval Research.

REFERENCES


