Hierarchical Duration Modelling for a Speech Recognition System

by

Grace Chung

B.S., University of New South Wales (1993)
B.Eng., University of New South Wales (1995)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1997

Author .....
Department of Electrical Engineering and Computer Science
May 9, 1997

Certified by .....
Stephanie Seneff
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Hierarchical Duration Modelling for a Speech Recognition System

by

Grace Chung

Submitted to the Department of Electrical Engineering and Computer Science
on May 9, 1997, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Durational patterns of phonetic segments and pauses convey information about the linguistic content of an utterance. Most speech recognition systems grossly underutilize the knowledge provided by durational cues due to the vast array of factors that influence speech timing and the complexity with which they interact. In this thesis, we introduce a duration model based on the ANGIE framework. ANGIE is a paradigm which captures morpho-phonemic and phonological phenomena under a unified hierarchical structure. Sublexical parse trees provided by ANGIE are well-suited for constructing complex statistical models to account for durational patterns that are functions of effects at various linguistic levels. By constructing models for all the sublexical nodes of a parse tree, we implicitly model duration phenomena at these linguistic levels simultaneously, and subsequently account for a vast array of contextual variables affecting duration from the phone level up to the word level.

This thesis will describe our development of a durational model, and will demonstrate its utility in a series of experiments conducted in the ATIS domain. The aim is to characterize phenomena such as speaking rate variability and prepausal lengthening in a quantitative manner. The duration model has been incorporated into a phonetic recognizer and a wordspotting system. We will report on the resulting improvement in performance.

In this duration model, a strategy has been formulated in which node durations in upper layers are successively normalized by their respective realizations in the layers below; that is, given a nonterminal node, individual probability distributions, corresponding with each different realization in the layer immediately below, are all scaled to have the same mean. This reduces the variance at each node, and enables the sharing of statistical distributions. Upon normalization, a set of relative duration models is constructed by measuring the percentage duration of nodes occupied with respect to their parent nodes. Under this normalization scheme, the normalized duration of a word node is independent of the inherent durations of its descendents and hence is an indicator of speaking rate. A speaking rate parameter can be defined as a ratio of the normalized word duration over the global average normalized word duration. This speaking rate parameter is then used to construct absolute duration models that are normalized by speaking rate. This is done by scaling either absolute phone or phoneme duration by the above parameter. By combining the hierarchical normalization and speaking rate normalization, the average standard deviation for phoneme duration was reduced from 50ms to 33ms.

Using the hierarchical structure, we have conducted a series of experiments investigating speech timing phenomena. We are specifically interested in the (1) examining secondary effects of speaking rate, (2) characterizing the effects of prepausal lengthening and (3) detecting other word boundary effects associated with duration such as gemination. For example, we have found, with statistical significance, that a suffix within a word is affected far more by speaking rate than is a prefix. It is also observed that prepausal lengthening affects the various sublexical units non-uniformly. For example, the stressed nucleus in the syllable tends to be lengthened more than the onset position.

The final duration model has been implemented into the ANGIE phonetic recognizer. In addition to contextual effects captured by the model at various sublexical levels, the scoring mechanism also accounts explicitly for two inter-word level phenomena, namely, prepausal lengthening and gemination. Our experiments have been conducted under increasing levels of linguistic constraint with

correspondingly different baseline performances. The improved performance is obtained by providing implicit lexical knowledge during recognition. When maximal linguistic contraint is imposed, the incorporation of the relative and speaking rate normalized absolute phoneme duration scores reduced the phonetic error rate from 29.7% to 27.4%, a relative reduction of 7.7%. These gains are over and above any gains realized from standard phone duration models present in the baseline system.

As a first step towards demonstrating the benefit of duration modelling for full word recognition, we have conducted a preliminary study using duration as a post-processor in a word-spotting task. We have simplified the task of spotting city names in the ATIS domain by choosing a pair of highly confusable keywords, "New York" and "Newark." All tokens initially spotted as "New York" are passed to a post-processor, which reconsiders those words and makes a final decision, with the duration component incorporated. For this task, the duration post-processor reduced the number of confusions from 60 to 19 tokens out of a total of 323 tokens, a 68% reduction of error.

In another experiment, the duration model is fully integrated into an ANGIE-based wordspotting system. As in our phonetic recognition experiments, results were obtained at varying degrees of linguistic contraint. Here, when maximum constraint is imposed, the duration model improved performance from 89.3 to 91.6 (FOM), a relative improvement of 21.5%. This research has demonstrated success in employing a complex statistical duration model in order to improve speech recognition performance. It has shown that duration can play an important role in aiding word recognition and promises to offer greater gains for continuous word recognition.

Thesis Supervisor: Stephanie Seneff
Title: Principal Research Scientist

Hierarchical Duration Modelling for a Speech Recognition System

by

Grace Chung

Submitted to the Department of Electrical Engineering and Computer Science
on May 9, 1997, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

Durational patterns of phonetic segments and pauses convey information about the linguistic content of an utterance. Most speech recognition systems grossly underutilize the knowledge provided by durational cues due to the vast array of factors that influence speech timing and the complexity with which they interact. In this thesis, we introduce a duration model based on the ANGIE framework. ANGIE is a paradigm which captures morpho-phonemic and phonological phenomena under a unified hierarchical structure. Sublexical parse trees provided by ANGIE are well-suited for constructing complex statistical models to account for durational patterns that are functions of effects at various linguistic levels. By constructing models for all the sublexical nodes of a parse tree, we implicitly model duration phenomena at these linguistic levels simultaneously, and subsequently account for a vast array of contextual variables affecting duration from the phone level up to the word level.

This thesis will describe our development of a durational model, and will demonstrate its utility in a series of experiments conducted in the ATIS domain. The aim is to characterize phenomena such as speaking rate variability and prepausal lengthening in a quantitative manner. The duration model has been incorporated into a phonetic recognizer and a wordspotting system. We will report on the resulting improvement in performance.

In this duration model, a strategy has been formulated in which node durations in upper layers are successively normalized by their respective realizations in the layers below; that is, given a nonterminal node, individual probability distributions, corresponding with each different realization in the layer immediately below, are all scaled to have the same mean. This reduces the variance at each node, and enables the sharing of statistical distributions. Upon normalization, a set of relative duration models is constructed by measuring the percentage duration of nodes occupied with respect to their parent nodes. Under this normalization scheme, the normalized duration of a word node is independent of the inherent durations of its descendents and hence is an indicator of speaking rate. A speaking rate parameter can be defined as a ratio of the normalized word duration over the global average normalized word duration. This speaking rate parameter is then used to construct absolute duration models that are normalized by speaking rate. This is done by scaling either absolute phone or phoneme duration by the above parameter. By combining the hierarchical normalization and speaking rate normalization, the average standard deviation for phoneme duration was reduced from 50ms to 33ms.

Using the hierarchical structure, we have conducted a series of experiments investigating speech timing phenomena. We are specifically interested in the (1) examining secondary effects of speaking rate, (2) characterizing the effects of prepausal lengthening and (3) detecting other word boundary effects associated with duration such as gemination. For example, we have found, with statistical significance, that a suffix within a word is affected far more by speaking rate than is a prefix. It is also observed that prepausal lengthening affects the various sublexical units non-uniformly. For example, the stressed nucleus in the syllable tends to be lengthened more than the onset position.

The final duration model has been implemented into the ANGIE phonetic recognizer. In addition to contextual effects captured by the model at various sublexical levels, the scoring mechanism also accounts explicitly for two inter-word level phenomena, namely, prepausal lengthening and gemination. Our experiments have been conducted under increasing levels of linguistic constraint with

correspondingly different baseline performances. The improved performance is obtained by providing implicit lexical knowledge during recognition. When maximal linguistic contraint is imposed, the incorporation of the relative and speaking rate normalized absolute phoneme duration scores reduced the phonetic error rate from 29.7% to 27.4%, a relative reduction of 7.7%. These gains are over and above any gains realized from standard phone duration models present in the baseline system.

As a first step towards demonstrating the benefit of duration modelling for full word recognition, we have conducted a preliminary study using duration as a post-processor in a word-spotting task. We have simplified the task of spotting city names in the ATIS domain by choosing a pair of highly confusable keywords, "New York" and "Newark." All tokens initially spotted as "New York" are passed to a post-processor, which reconsiders those words and makes a final decision, with the duration component incorporated. For this task, the duration post-processor reduced the number of confusions from 60 to 19 tokens out of a total of 323 tokens, a 68% reduction of error.

In another experiment, the duration model is fully integrated into an ANGIE-based wordspotting system. As in our phonetic recognition experiments, results were obtained at varying degrees of linguistic contraint. Here, when maximum constraint is imposed, the duration model improved performance from 89.3 to 91.6 (FOM), a relative improvement of 21.5%. This research has demonstrated success in employing a complex statistical duration model in order to improve speech recognition performance. It has shown that duration can play an important role in aiding word recognition and promises to offer greater gains for continuous word recognition.

Thesis Supervisor: Stephanie Seneff
Title: Principal Research Scientist

Acknowledgments

First and foremost, I would like to thank my thesis supervisor, Stephanie Seneff for her guidance and support. Her wealth of knowledge and creative ideas have been a source of inspiration and I have benefited vastly from our many fruitful hours of brainstorming. It has truly been a great pleasure to work with her.

There are numerous others who have helped bring this work into fruition.

I am deeply indebted to Victor Zue for providing me with the opportunity to work in this rewarding research environment. I would like to thank my very talented colleagues, Ray Lau, Tim Hazen and Jim Huginin for having the answers to almost all conceivable technical questions. I have Ray Lau to thank for conducting the wordspotting experiments and Joshua Koppelman who was originally responsible for this project and whose computer programs I inherited.

I would also like to express my most sincere gratitude to Warren Lam. His warm friendship, boundless support and thoughtful advice have made all the difference to my life at MIT and beyond.

Finally, I dedicate this thesis to my father, C.M., and my sister, Mary, and Dennis Lee who have given me invaluable emotional support and encouragement. Their belief in me have inspired me to pursue my goals.

Contents

1 Introduction **15**

1.1 Background 15

1.2 Factors Which Affect Speech Timing 16

1.2.1 Phonological Component 16

Inherent Duration 16

Contextual Effects 16

1.2.2 Lexical Component 17

1.2.3 Syntactic Structure 18

1.2.4 Speaking Rate and Rhythm 18

1.2.5 Discourse Level 19

1.2.6 Summary 19

1.3 History of Duration Modelling Experiments 19

1.4 Goals and Overview of Research 24

2 Hierarchical Duration Modelling **27**

2.1 The ANGIE Framework 27

2.2 The Duration Model 29

2.2.1 Hierarchical Normalization Scheme 31

2.2.2 Relative Duration Model 32

Computing a Duration Score 32

Advantages of Relative Duration Model 33

2.2.3 Speaking Rate Parameter 33

2.2.4 Absolute Duration Model 35

2.3 Experimental Conditions 35

2.3.1 ATIS Corpus 35

2.3.2 Forced Alignment 36

3 Analysis of Speech Timing and Speaking Rate Variability **37**
3.1 Variance Reduction . . . 38
3.1.1 Hierarchical Normalization . . . 38
3.1.2 Speaking Rate Normalized Absolute Duration . . . 39
Non-uniform Rate Normalization . . . 44
Absolute Duration Models . . . 47
3.2 Speaking Rate Experiments . . . 49
3.2.1 Rate Dependence of Relative Duration Model . . . 49
3.2.2 Secondary Effects of Speaking Rate . . . 50
3.2.3 Variability of Speaking Rate . . . 55
3.2.4 Analysis of Slow Words . . . 55
3.3 Studies Characterizing Prepausal Lengthening . . . 56
3.3.1 Speaking Rate and Prepausal Lengthening . . . 57
3.3.2 Pause Duration . . . 58
3.3.3 Prepausal Model . . . 59
3.3.4 Secondary Effects of Prepausal Lengthening . . . 62
3.4 Word Boundary Effects of Duration . . . 64
3.4.1 Gemination . . . 65
3.4.2 Word-Final Stop Closures . . . 65
3.5 Summary . . . 66

4 Phonetic Recognition Experiments **68**
4.1 Implementation Issues . . . 69
4.1.1 Integration with the ANGIE Recognizer . . . 69
4.1.2 Computation of Duration Scores . . . 69
Model for Prepausal Data . . . 70
Model for Geminate Data . . . 71
4.1.3 Combining Duration Scores . . . 71
4.2 Details of Experiment . . . 72
4.2.1 Training . . . 72
4.2.2 Linguistic Constraint . . . 72
4.3 Results . . . 73
4.4 Analysis . . . 73

5 Wordspotting Experiments **77**
5.1 Preliminary Investigations . . . 78
5.2 Duration as a Post-processor . . . 78

5.2.1 Details of Experimental Procedure 81
5.2.2 Results and Analysis 81
5.3 A Wordspotting System with Fully Integrated Duration Component 85
5.3.1 Results and Analysis 86
5.4 Summary 87

6 Conclusion and Future Work **89**
6.1 Future Work 90

A ANGIE Categories **92**

B Tables **94**

List of Figures

2-1 *Sample parse tree for the phrase "I'm interested...".* 28

2-2 *Hierarchical Duration Modelling Scheme* 30

2-3 *Phoneme /d*ed/ realized as ix followed by dx.* 31

3-1 *Reduction of Standard Deviation due to Hierarchical Normalization for the* WORD *Node: Mean duration is 331ms. Standard deviation is reduced from 180ms to 109ms.* 40

3-2 *Reduction of Standard Deviation due to Hierarchical Normalization for the* DSUF *Node at the Morph Layer: Mean duration is 200ms. Standard deviation is reduced from 96ms to 67ms.* 40

3-3 *Reduction of Standard Deviation due to Hierarchical Normalization for the* DNUC *Node at the Syllable Layer: Mean duration is 110ms. Standard deviation is reduced from 64ms to 44ms.* 41

3-4 *Reduction of Standard Deviation due to Hierarchical Normalization for the Special Phoneme /ra/ in Function Word "from": Mean duration is 87ms. Standard deviation is reduced from 62ms to 51ms.* 41

3-5 *Speaking Rate Normalization for the Phone dh: Standard deviation is reduced from 23ms to 13ms. Mean duration is 40ms before and after normalization.* 45

3-6 *Speaking Rate Normalization for the Phone ow: Standard deviation is reduced from 53ms to 38ms. Mean duration is 124ms prior to normalization and 127ms after normalization.* 45

3-7 *Speaking Rate Normalization for the Phoneme /uw/ in the Function Word "to": Standard deviation is reduced from 68ms to 22ms. Mean duration is 75ms prior to normalization and 73ms after normalization.* 46

3-8 *Speaking Rate Normalization for the Phoneme /t/ in the Non-Onset Position: Standard deviation is reduced from 41ms to 25ms. Mean duration is 117ms prior to normalization and 120 after normalization.* 47

3-9 *Statistical Distribution and Gamma Model for Phoneme /uw/ in the Function Word "to": The statistical distribution, based on the hierarchical and rate normalized duration, is computed from 2298 tokens. Mean duration is 117ms. Standard deviation is 35ms. For this Gamma model,* $\lambda = 0.1$ *and* $\alpha = 11.7$. 48

3-10 *Hierarchical Word Score as a Function of Speaking Rate: All words (43,467) are divided into subsets of 1000 tokens according to their speaking rate. The average word score (*μ*), calculated from the relative duration model, is plotted against the average speaking rate for each subset.* $\mu + \sigma$ *and* $\mu - \sigma$ *represent the boundaries one standard deviation from the mean. The total mean score is 1.54*. 49

3-11 *Relative Duration for Parts of a Word Corresponding with 3 Speaking Rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example of this word pattern is in the word "December".* 51

3-12 *Relative Duration for Parts of a Morph Unit Corresponding with 3 Speaking Rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. Two contrasting examples of (*UONSET NUC*) are given. Examples are the prefix and unstressed root in the word "tomorrow".* 51

3-13 *Relative Duration for Parts of a Phoneme Units Corresponding with 3 Speaking Rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. Examples are function-word specific diphones in the words "from" and "and".* 52

3-14 *Relative Duration for Parts of a Phoneme Units Corresponding with 3 Speaking Rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. Examples of special diphthongs are given. The phoneme /ehr+/ may appear in the word "air" and the phoneme /aol+/ may appear in the word "all".* 53

3-15 *Behaviour of Stops in Onset Position with respect to Speaking Rate: Each of the six stops are divided into five equal subsets of data in accordance with speaking rate of the word they occur in. Average percentage duration occupied by stop closure within a stop phoneme is plotted as a function of average speaking rate for each subset, represented by a "*$\times$*" on the plot.* 54

3-16 *Percentage of Prepausal Words amoung All Words: All words are partitioned into subsets according to their measured speaking rate. The size of each subset is determined empirically and corresponds approximately as equal bins on a logarithmic display axis. For all words in each subset, the percentage of prepausal words is plotted.* 58

3-17 *Average Speaking Rate of Prepausal Words vs Duration of Pause: All sentence-internal prepausal words are divided into equal bins of 400 tokens according to duration of corresponding pause. In each bin, the average speaking rate of the prepausal words is plotted.* 59

3-18 *Statistical Distributions of Speaking Rate for Non-Prepausal Words, Sentence-Internal Prepausal Words and Sentence-Final Words. The top histogram indicates the speaking rate distribution of non-prepausal data. For the middle and bottom plot, the histograms depict speaking rate distributions for data tagged as "irregular" while the dotted lines depict distributions for data tagged as "regular".* 60

3-19 *Relative Duration for Normal, Regular Prepausal and Irregular Prepausal Speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example of this word pattern is in the word "flights".* 62

3-20 *Relative Duration for Normal, Regular Prepausal and Irregular Prepausal Speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example of this word pattern is in the word "travel".* 63

3-21 *Relative Duration for Normal, Regular Prepausal and Irregular Prepausal Speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example where this pattern occurs is in the word "slip"* 63

3-22 *Relative Duration for Normal, Regular Prepausal and Irregular Prepausal Speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example where this pattern occurs is in the final syllable of the word "tomorrow"* 64

5-1 *Probability Distributions of Relative Duration Scores for Hits and False Alarms.* 79

5-2 *Probability Distributions of Absolute Duration Scores for Phonemes for Hits and False Alarms.* 79

5-3 *Probability Distributions of Absolute Duration Scores for Phones for Hits and False Alarms.* 80

5-4 *Probability Distributions of Speaking Rates for Hits and False Alarms.* 80

5-5 *Number of Errors versus Weight Using Relative Duration Model. The total number of errors is the sum of the number of misses and false alarms. 288 tokens are scored in total.* 83

5-6 *Number of Errors versus Weight Using Absolute Duration Model for Phonemes. The total number of errors is the sum of the number of misses and false alarms. 288 tokens are scored in total.* 83
5-7 *Number of Errors versus Weight Using Absolute Duration Model for Phones. The total number of errors is the sum of the number of misses and false alarms. 288 tokens are scored in total.* 84

List of Tables

1.1 *History of Duration Modelling Development. † This is the only experiment where spontaneous speech was used. All other experiments were conducted over corpora of read speech.* 25

3.1 *Hierarchical Normalization: reduction in standard deviation for each sublexical layer. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. $\Delta\%$: Percentage reduction of variance.* 39

3.2 *Speaking Rate Normalization: reduction in standard deviation for three sublexical layers. μ_1: Normalized mean duration of all tokens. σ_1: Standard deviation after rate normalization of all tokens. μ_2: Normalized mean duration of all tokens except those which are normalized deterministically. σ_2: Standard deviation after rate normalization, discarding deterministic nodes. $\Delta\%$: Percentage reduction of the standard deviation over mean ratio for respective normalization scheme.* 43

3.3 *Comparing the Characteristics of Regular and Irregular Prepausals and Non-prepausal Words.* 61

3.4 *Geminate Phones: comparing the mean duration for phones under geminate and non-geminate conditions. μ_1: Mean of the non-geminate phone. μ_2: Mean of the geminate counterpart.* 65

3.5 *Lengthening of Word-Final Closures: Mean duration (ms) and counts for word-final /tcl/ and /kcl/ when followed, in word-initial position, by the six stop releases compared with all other phones.* 66

4.1 *Results of Phonetic Recognition Experiment Using the* Angie *Parse Tree with No Additional Constraint. The percentage error rate with their component substitutions, deletions and insertions are given. Δ represents the percentage error reduction from error rate using no duration.* 74

4.2 *Results of Phonetic Recognition Experiment Using Morph Constraints. The percentage error rate with their component substitutions, deletions and insertions are given. Δ represents the percentage error reduction from error rate using no duration.* 74

4.3 *Results of Phonetic Recognition Experiment Using Morphs with Word Constraints. The percentage error rate with their component substitutions, deletions and insertions are given. Δ represents the percentage error reduction from error rate using no duration.* 75

5.1 *Results of using Duration Processor to Rescore Hypothesized "New York"s with Optimized Duration Weights. 323 waveforms were processed in total.* 85

5.2 *Results of Wordspotting Experiments.* 87

B.1 *Hierarchical Normalization of Morphological Units: reduction in standard deviation. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.* 94

B.2 *Speaking Rate Normalization of Morphological Units: reduction in standard deviation. μ Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discarded. Δ% : Percentage reduction of standard deviation to mean ratio.* . . 94

B.3 *Hierarchical Normalization of Syllablic Units: reduction in standard deviation. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.* 95

B.4 *Hierarchical Normalization of Phonemic Units: reduction in standard deviation for vowels. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.* 96

B.5 *Speaking Rate Normalization of Phonemic Units: reduction in standard deviation for vowels. μ: Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discared.. Δ% : Percentage reduction of standard deviation to mean ratio.* 97

B.6 *Hierarchical Normalization of Phonemic Units: reduction in standard deviation for function word specific phonemes. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.* 98

B.7 *Speaking Rate Normalization of Phonemic Units: reduction in standard deviation for function word specific phonemes. μ Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discared.. Δ% : Percentage reduction of standard deviation to mean ratio.* 98

B.8 *Hierarchical Normalization of Phonemic Units: reduction in standard deviation for affricates, stops and fricatives. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.* . . 99

B.9 *Speaking Rate Normalization of Phonemic Units: reduction in standard deviation for affricates, stops and fricatives. μ: Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discared.. $\Delta\%$: Percentage reduction of standard deviation to mean ratio.* 100

B.10 *Hierarchical Normalization of Phonemic Units: reduction in standard deviation for nasals, semivowels and aspirants. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. $\Delta\%$: Percentage reduction of variance.* 101

B.11 *Speaking Rate Normalization of Phonemic Units: reduction in standard deviation for nasals, semivowels and aspirants. μ Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discared.. $\Delta\%$: Percentage reduction of standard deviation to mean ratio.* 101

Chapter 1

Introduction

1.1 Background

It is generally well-known that the durational patterns of phonetic segments and pauses convey information about the linguistic content of an utterance. Listeners make linguistic decisions on the basis of durational cues which can serve to distinguish, for example, between inherently long versus short vowels, voiced versus unvoiced fricatives, phrase-final versus non-final syllables and stressed versus unstressed vowels. Duration is also used to detect the presence or absence of emphasis.

If duration information is of perceptual importance to the human listener, then it possibly holds significant potential for improving speech recognition performance. However, our current understanding of durational patterns and the many sources of variability which affect them, is still sparse. To date, most speech recognition systems only have rudimentary duration models and have yet to incorporate comprehensive models that fully utilize the knowledge provided by durational cues. This is attributed to a vast array of factors that influence speech timing and the complexity with which they interact. In fact, speech timing is modified at various linguistic levels of the sentence generation process. Because of this abundance of factors that coexist at multiple levels, their interactions obscure each other's manifestations, rendering their presence difficult to detect or analyze.

This thesis presents a complex duration model designed to aid speech recognition[1]. Our framework, called ANGIE, is a hierarchical representation composed of *sublexical* or *subword* units. Such a framework enables us to characterize and model speech timing phenomena at multiple linguistic levels simultaneously. It is the hope that this duration model will ultimately enhance the performance of a large vocabulary continuous speech recognizer and as a consequence, demonstrate the important role of durational cues. In the following, we will begin in Section 1.2 by considering the

[1] The primary focus of this thesis is on duration patterns for English although our modelling framework is applicable to other languages as well.

durational effects at each level of the phonetic hierarchy, that is, from the phonological component to the discourse level. Next, in Section 1.3, we will review and compare some previous approaches to duration modelling and the difficulties encountered. Finally, the remainder of this chapter will provide an overview of our goals.

1.2 Factors Which Affect Speech Timing

1.2.1 Phonological Component

Inherent Duration

Each phonetic segment has its own intrinsic or inherent phonological duration. Hence features of a phoneme's identity may be perceptually cued by duration. Some of these effects are listed below.

- In general, vowels can be divided into contrastive long-short pairs or tense-lax opposition. Schwas such as */ix, ax, axr/* [2] are found to be shorter than other vowels such as */iy, ae, aa/* [10, 22]. Also, vowel durations tend to vary inversely with vocalic height [20].
- For fricatives, those which are voiceless, */s, sh, f, th/* are about 40ms longer than their voiced counterparts, */z, zh, v, dh/* [13].
- For voiced stops */b, g, d/*, the duration of releases as well as the voice onset time (VOT) are inherently shorter than those of voiceless stops */p, k, t/*. Labial stop closures are generally longer than alveolar closures [13].

Contextual Effects

The intrinsic duration of a phonetic segment is often perturbed under varying contextual conditions; that is, the identity of neighbouring segments exerts some influence on the phone duration. Many observations of these context dependent effects on segmental durations have been documented, although not all studies are consistent with each other. These are described below.

- House [10] claimed that the primary influence on vowel duration is attributed to the voicing characteristic of the following consonant. Numerous others have demonstrated that vowels are shorter when followed by voiceless consonants than when followed by voiced consonants [13, 20, 22]. Van Santen found differences of 100ms for vowel durations when followed by voiced and voiceless consonants for a single speaker [33]. In fact, vowel duration is perceptually important for humans to detect the presence of voicing in post-vocalic consonants [16]. One study [22] found the influence of the initial consonant on the duration of the syllable nucleus

[2] Throughout this thesis, we will use the ARPAbet nomenclature for phonetic units.

to be negligible, while another [6] reports that vowels are longer following stops than following other consonants. It is generally agreed that the manner of production (nasal, fricative, etc.) of the post-vocalic consonant shows a smaller effect than does voicing, although [22] reports that some vowels are 20–25% longer when followed by a fricative than when followed by a stop, controlling for vowel identity, consonant place and voicing.

- Consonants also undergo durational changes depending on the phonetic environment. Umeda, [32], found that the identity of adjacent vowels has negligible effects on consonantal duration, but adjacent consonants both within words and across word boundaries have significant influence on consonant duration. For example, /*p*/ and /*s*/ are shortened in an /*sp*/ cluster and /*r*/ is about 30ms longer in consonant clusters with aspirated /*p, t, k*/ than in clusters with /*b, d, g*/.

- The distinctive features of phonemes can also be cued by the duration of an adjacent phoneme. For example, in a vowel-nasal-obstruent sequence, voicing of the obstruent is associated with lengthening of both the vowel and the nasal [6, 25].

- The correlation between the number of syllables in a word and segmental duration is not well understood. It has been observed that, in carrier-phrase mode, vowel and consonant duration can decrease as the number of syllables in a word increases [9, 12, 25] but this effect becomes obscure in more natural speech settings.

- *Gemination* is the phenomenon where two identical phonemes such as nasals or fricatives occur adjacent to each other [23]. Because their acoustic realization often exhibits a minimal transition or no spectral change at all, they can be viewed as a single phonetic unit representing two phonemes. Geminates in general are longer than one single phoneme but shorter than the sum of two. Duration is an important cue for their detection.

Further details of contextual effects can be found in [6, 31, 32].

1.2.2 Lexical Component

Segmental duration is affected by position in the word, and duration is used perceptually to distinguish two words which differ by lexical stress patterns. Stressed vowels are known to be longer than unstressed vowels [26] whereas consonants in pre-stressed positions are often longer than their unstressed and post-stressed counterparts [13]. Word-final lengthening has been observed by some but not others [9].

1.2.3 Syntactic Structure

Duration is influenced by syntactic structure, such as phrasal and clausal patterns. Vowels in syllables preceding phrase boundaries have been found to be twice as long as vowels in non-phrase-final syllables. This lengthening occurs at phrase or clause boundaries even when there is no physical pause in the acoustic signal, and vowels exhibit more lengthening than consonants [13, 31, 32].

Prepausal lengthening is the effect where segments followed by pauses are lengthened in duration. Studies report a 60–200ms increase in syllable duration with most of the durational increment restricted to the vowel and any postvocalic sonorant or fricative consonants [13]. Under prepausal conditions, the variances of vowel durations are larger [20, 31] and effects at sentence-internal pauses are more pronounced than at the end of sentences [20].

Many studies have examined the nature of prepausal lengthening and found that lengthening in the syllable can imply lengthening of some phonemes more than others [2, 8, 37]. Wightman et al. [37] reported a study of segmental lengthening in the vicinity of prosodic boundaries and found that it is restricted to the rime of the syllable preceding the boundary. Campbell [2] found that whereas segments in prepausal sentence-final syllables undergo greater lengthening in the rime than in the onset, segments in sentence-internal syllables are lengthened or compressed (due to stress, rhythmic or other factors) more uniformly across the syllable.

Segmental duration is also dependent upon whether the context is a content word or function word. Content words usually carry information concerning the content of the message, whereas function words are easily predicted and therefore pronounced with minimum effort. Umeda [32] found that the duration of consonants varies according to the function/content word distinction. However, this distinction is not clear-cut. For example, while words in some classes such as prepositions, articles and conjunctions are generally agreed to be function words, other frequently occurring words such as pronouns may or may not be classified as function words, even though they are reduced.

1.2.4 Speaking Rate and Rhythm

Variation in speaking rate, both within one speaker and among several speakers, is an important factor when accounting for variation in duration. However it is difficult to quantify speaking rate because it is a continuous variable and a reliable measure has not been found. Its influence on segment durations is also not well understood.

In general, pauses make up about 20% of the time in fluent reading and 50% in conversation. Crystal [5] found that slowing down in speech can be attributed to the introduction of new pauses (54%), increased duration of existing pauses (27%) and increased duration of speech segments (19%), tempo being measured by the total elapsed time.

Increases in speaking rate also accompany phonological and phonetic simplifications as well

as differential shortening of vowels and consonants. Moore and Zue [18] found that palatization, gemination, flapping and schwa devoicing become more frequent as speaking rate increases, and the incidences of pause insertion and glottal stop insertion increase as speaking rate slows. Many studies have shown that the duration of unstressed syllables is proportionally more affected by rate changes than that of stressed syllables [5, 22, 25].

In addition to speaking rate, it has been postulated that speakers aim to maintain a certain rhythmicity although there is no conclusive evidence about the influence of rhythm and the constraints that it imposes on segmental duration [8].

1.2.5 Discourse Level

Little is known about duration at a discourse level, although it may appear that speakers slow down at the end of conceptual units while emphasis and contrastive stress tend to increase duration by 10–20% [13]. Umeda [31] has shown that semantic novelty has an influence on segmental durations in that an unusual word is longest the first time it appears in a connected discourse and has lower stress and therefore shorter duration in subsequent occurrences in a passage.

1.2.6 Summary

As we have shown above, there is an abundance of factors which influence the behaviour of durational patterns, and this renders duration modelling particularly difficult. These durational effects operate in concert at multiple levels, ranging from the detailed phonological effects to paragraph phenomena. Such interactions are complex and poorly understood, and therefore difficult to incorporate into a comprehensive model. For example, unstressed vowels which are often reduced may occur in word-final positions. In this case, these reduced vowels are also subject to lengthening. Front vowels are shorter than back vowels when preceding labial or dental consonants but are longer when preceding velars [6]. A large source of variability in segmental duration can be attributed to speaker differences. Such differences may be larger than the contextual constraints observed for the speech of a single speaker [5]. Wang et al. [24, 35] found that much more variation is introduced when examining data from a multi-speaker corpus such as TIMIT. For example, the durational distributions of stressed vowels followed by voiced and unvoiced plosives in the same syllable are very similar, contrary to similar previous studies which used only one speaker.

1.3 History of Duration Modelling Experiments

Researchers have attempted to mathematically model durational patterns and predict segmental duration since the seventies. At first, most duration modelling experiments were directed towards

speech synthesis applications. Synthesis systems require duration models to provide a single duration estimate for each linguistic unit given its contextual environment, in order to produce natural sounding speech. Complex statistical and heuristic models were developed to account for the multiple factors that influence speech timing. This research was predominantly conducted over corpora based on a small number of speakers with read speech or isolated words. Over the past decade, researchers have progressively begun to consider duration models for speech recognition. This regime differs from the synthesis domain in that it is more desirable here to use corpora based on continuous spontaneous speech rather than read speech or isolated words, such that real conditions of use can be better emulated, particularly for large vocabulary continuous speech recognizers. In addition, for speaker-independent recognizers, duration models should be constructed from a multi-speaker corpus. At this nascent stage, most work has concentrated on small vocabularies, isolated words and read speech, with the most recent research progressing towards large vocabulary read speech with a greater number of speakers. Thus, even the most sophisticated duration models that were driven by synthesis, are not necessarily applicable for recognition models. The presence of multiple speakers adds greater variability and acts to confound or obscure duration phenomena observed for single-speaker corpora. These difficulties and our incomplete understanding of durational phenomena have prevented researchers in recognition from developing complex duration models that address contextual issues at various linguistic levels. Instead, most recognition systems have resorted to simple context-independent duration estimates for phones. However, researchers have recently begun to develop simple ways of estimating speaking rate and consequently normalizing input speech by this estimated rate. Since the vast linguistic information encoded in duration remains largely untapped, performance improvements have thus far been modest, despite the potential for greater gains. This section will detail the history of duration modelling experiments and elaborate on some of the problems encountered.

Van Santen [34] has identified four categories in duration modelling approaches: sequential rule systems, equation systems, table-lookup systems and binary trees. Most models predict segmental duration at the lowest phoneme-sized level by taking into account explicitly some of the observed contextual effects and their interactions.

The first duration model to appear in the literature [13] was a sequential rule-based system which addresses one or two factors at a time but fails to capture the more complex interactions. Consequently, researchers have employed more sophisticated techniques such as classification and regression trees (CART) [23, 29] and complex equation models [33]. More recently, researchers in recognition have attempted to solve the problem of speech rate variability by normalizing segmental duration [1, 11, 19]. However, to date, models which extensively incorporate duration knowledge have not been employed in recognition.

The following details some of these approaches.

Umeda (1975) [31, 32] studied rules which explain the behaviour of segmental durations. For vowels, she developed multiplicative models whose scaling factors are dependent upon vowel identity, the identity of the consonant following the vowel and suprasegmental variables such as position in word and sentence, word prominence, sentence stress and speech rate. In the consonant model, separate initial values are computed for the case of the word-initial intervocalic consonant and the word-final intervocalic consonant. Consonant durations are then modified according to different conditions of phonetic context, stress and pause position.

Klatt (1976) [12, 13] developed a model which specifies a generative theory of segmental duration in which phonetic segments are assumed to have some inherent duration and vowels were strongly incompressible beyond a certain amount of shortening. Thus, there exists a minimum duration that is about 45% of the inherent duration for a given vowel. A sequence of ordered rules, either multiplicative or additive, can be applied to modify the portion of the inherent duration exceeding a specified minimum as a function of phonetic and phrasal environment. He also incorporated a speaking rate parameter based on the number of words per minute. This model explained 84% of the variance for new paragraphs for the speaker on which the rules were developed.

O'Shaughnessy (1984) [20] presented a generative model of French durations for synthesis-by-rule. Baseline durations are specified by manner class and effects such as function-word reduction and voicing. These durations are then modified under conditions of phonetic context and word and sentence position. He postulated that a duration model is likely to be useful in recognition in confirming and rejecting hypotheses proposed by the speech recognizer.

While the above experiments were motivated by synthesis, Port et al. (1988) [26] aimed to extract properties in speech suitable for distinguishing words from a small vocabulary, a first step in using duration for speech recognition. They sought to capture linguistically relevant information in timing at the syllable level by examining words produced by different speakers at different speech rates. The authors demonstrated some success in differentiating words on the basis of timing when words differed dramatically in terms of stress pattern and consonant voicing. They also found that changes in overall speech rate alter segmental durations non-uniformly over a number of neighbouring segment types and this reduces the effectiveness of uniform scaling to eliminate tempo variation.

Crystal and House (1988) [6] developed duration models which were incorporated into a hidden Markov model (HMM) speech recognition system. They analyzed segment duration data for two 300-word passages read by each of six speakers of American English and then computed statistics for context-dependent phoneme segments. The model utilized Gamma functions as distributions for gross categories.

In the early nineties, researchers experimented with using new hierarchical paradigms for modelling duration. These are algorithms for generating statistical decision trees using an automatic procedure. The trees reduced durational variance given a set of contextual variables. Riley (1992) [28,

29] used CART for predicting segment duration for speech synthesis. He used 1500 utterances from a single speaker to build decision trees and his model accounted for segment identity and two surrounding segments and higher level phenomena such as lexical stress, word frequency, word and sentence position and dialect. Additionally, speaking rate was calibrated by the duration of two sentences which every speaker was asked to produce. Riley's model produced predictions with residuals of 23ms standard deviation.

Pitrelli (1990) [23] used a hierarchical model based on phoneme duration. He found that a substantial portion of duration data variance in a large, multi-speaker corpus can be explained by duration modelling. He conducted recognition experiments that indicated a duration post-processor using his model can yield a statistically significant improvement in performance for a limited vocabulary isolated-word speech recognizer. He chose a task in which 50 town names were spoken over long distance telephone lines, and a duration post-processor rescored the transcriptions proposed by the speech recognizer. The addition of a duration component reduced the error rate from 15.9% to 12.9%. This approach has the advantage that the modelling procedure automatically generates one particular model given a corpus and a set of candidate models. But it also has several disadvantages. Sparse data problems occur when data are partitioned at tree nodes successively down the tree. Secondly, the model is not suitable for modelling continuous parameters such as speaking rate.

Van Santen (1992) [33, 34] studied the speech of two speakers, generating a database of 18000 and 6000 vowel segments respectively, and measured the effects on vowel duration of several contextual factors, including those of syllabic stress, pitch accent, identities of adjacent segments, syllabic structure of a word, and proximity to a syntactic boundary. The research was motivated by the need to characterize durational effects and their interactions, in order to derive rules for natural speech synthesis. He argued that any set of durational rules has to address at least eight factors and their interactions, and so the degree of complexity lends itself to the use of mathematical equations to specify durations as opposed to the use of sequential rule systems. A sums-of-products model, consisting of sums of products of factor scales, was formulated to describe such factor interactions. Furthermore, Van Santen developed a methodology for analytically fitting such models to data, using analysis of ordinal patterns to determine some functional form and then analytically finding factor scales.

Underlying the work of Campbell (1992) [2, 3] is the concept that timing in speech can best be represented from the higher levels of the phrase, foot and syllable, and it is only finally realized at the level of the phonetic segment as a result of an interaction with the effects at higher levels. He argued that segment durations can, to a large extent, be predicted by a process of accommodation into a syllable-level timing framework. He developed a two-layer model of timing in which syllable duration is calculated to reflect the rhythmic and structural organization of an utterance while segment durations are calculated at a secondary stage of the process. A three-layer neural net

was trained by back-propagation to predict the overall syllable duration based on factors such as number of segments in the syllable, phrasal position, stress and grammatical category of the word in which the syllable occurs. Syllable duration was predicted in a log-transformed domain to map distributions closer to a Gaussian shape. Duration of segments within the syllable were derived by way of accommodating into this syllable time frame. Campbell conceived an "elasticity hypothesis" of segment duration which states that each segment in a particular syllable will have a durational value that reflects the same number of standard deviations about the mean for each segment. This implies that all segments in a given syllable fall at the same place in their respective distributions. For any given syllable, there is a number k of standard deviations such that the length of every segment in the syllable is equal to $\mu_{\text{seg}} + k\sigma_{\text{seg}}$, where μ_{seg} and σ_{seg} are the mean and standard deviation respectively of durations of the particular segment type. For instance, a vowel with a high variance such as a tense vowel that shows a large difference in absolute duration to its mean, is said be in the same relative state of expansion or compression as one with a much smaller variance that changes less in absolute terms. Therefore, duration of phonemes within a syllable are found by computing one value which can be applied to modify the mean duration of each phoneme in a syllable, in terms of its standard deviation, such that the results sum to the desired duration for the syllable as a whole. Campbell's experiments used two corpora of speech—one based on spontaneous radio broadcast speech and the second from readings of 200 phonetically balanced sentences. The implementation of this model accounted for 76% of syllable duration variance. These experiments suggest that while syllable duration can be predicted to a large extent, there is greater freedom in durational specification at the phonetic level.

In more recent research, duration information has been used to aid speech recognition and some experiments have demonstrated statistically significant improvement. Most HMM systems incorporate a minimal duration model by duplicating or adding states in the model. It is rather difficult to incorporate explicit duration models into the HMM itself, and, as a consequence, researchers [1, 11, 24, 36] have attempted to integrate a durational component as a post-processor, yielding some success. While there has not been the use of comprehensive duration models, there have been a few experiments conducted which demonstrate that speaking rate can be measured at recognition time and its variability can be taken into account by the recognizer. This is important because speakers with unusually fast or slow rates have been known to cause increased word error rates. Pols [24] cites evidence that speakers with a high speaking rate (in terms of number of words per minute) almost unanimously showed a higher word error rate.

Osaka et al. (1994) [19] described a spoken word recognition system which adapts to the speaking rate. Phoneme duration is used to estimate speech rate. A procedure effectively normalized phoneme duration by the average vowel duration and by the average duration of each phoneme class to reduce the variance of phoneme duration. An estimate of the duration of each phoneme in the input

speech is given in a first-order linear regression equation as a function of the average vowel duration. Experiments were computed from a 212-word vocabulary using five male and five female speakers. This model resulted in a word accuracy increase of 1.6% in a 212-word vocabulary to 97.3%.

Jones (1993) [11] and Anastasakos (1995) [1] both used a duration model as a post-processor for an HMM-based recognition system to rescore the N best hypotheses. Duration is modelled explicitly outside the framework of HMMs after the N best algorithm has been used to provide a list of likely hypotheses. It rescores and so reorders the N best list for a new sentence hypothesis. In both cases, the duration likelihood is given an empirically determined weight. Jones calculated speech rate by an average normalized phone duration, and the relative speaking rate of an utterance is indicated by the average normalized phone duration in that utterance. Anastasakos computed rate based on the observation of a small number of phoneme segments around a given phoneme. The phoneme duration models also incorporate lexical stress information and context dependency. Both experiments clustered the training data into different sets corresponding with slow and fast speakers and developed separate models for each set. During recognition, the measured speech rate is used to select the appropriate model. Both experiments also attempted to explicitly normalize phone duration with respect to the rate. Jones carried out his experiments using the TIMIT database with a vocabulary of 1794 words. The best result was a 10% reduction in error rate from a baseline word error rate of 13.62%. Anastasakos conducted experiments using the 5000-word WSJ corpus and, using the clustered models, reduced the error rate by 10% from 7.7% to 7%.

Thus, current research in duration modelling is moving towards developing more sophisticated models which can be employed for large vocabulary continuous speech recognizers. It has been shown that even simple models which attempt to normalize speaking rate can demonstrate modest improvements. Duration knowledge may become more important in scenarios where the acoustic signal is degraded but the relative timing is still preserved or when the input is highly spontaneous speech with large changes in speaking rate. A chronological summary of the development of duration modelling is provided in Table 1.1.

1.4 Goals and Overview of Research

This thesis has three primary goals:

- To develop a computational duration model based on a hierarchical framework which captures morphology, syllabification and phonology in the form of sublexical parse trees. This framework captures contextual effects from the phone level ranging up to the word level. Duration phenomena above the word level are beyond the scope of this thesis but could conceivably be added using the same framework.[3]

[3] However, our duration model also attempts to capture speaking rate effects which can arguably be considered

Name	Motivation	Speakers	Vocabulary	Model/Experiment
Umeda (1975)	Synthesis	3	10-20 minutes isolated words	Multiplicative models
Klatt (1976)	Synthesis	3	80 isolated words	Multiplicative/additive sequential rule system
O'Shaughnessy (1984)	Synthesis	29	111 isolated words	Multiplicative/additive sequential rule system
Port (1988)	Recognition	6	8 isolated words	Use of timing to discriminate between words
Crystal et al. (1988)	Recognition	6	300 words continuous speech	HMM using Gamma probability functions
Pitrelli (1990)	Recognition	many	50 isolated words	Hierarchical decision trees
Riley (1992)	Synthesis	1	1500 utterances continuous speech	CART algorithm
Van Santen (1992)	Synthesis	2	24,000 isolated words	Sums of products model
Campbell (1992)	Synthesis	2	200 sentences and 20 min radio broadcast† continuous speech	Elasticity hypothesis
Jones (1993)	Recognition	many	1794 words continuous speech	HMM based speaking rate normalization
Osaka (1994)	Recognition	10	212 isolated words	HMM based speaking rate normalization
Anastasakos (1995)	Recognition	many	36,000 words continuous speech	HMM based speaking rate normalization

Table 1.1: *History of Duration Modelling Development.* † *This is the only experiment where spontaneous speech was used. All other experiments were conducted over corpora of read speech.*

- To conduct a series of experiments to examine temporal phenomena in speech based on such a framework. The aim is to characterize phenomena such as prepausal lengthening, gemination and speaking rate variability in a quantitative manner.
- To implement the final duration model, which extensively incorporates knowledge of durational effects, into a speech recognizer in order to improve recognition results and demonstrate the value of durational cues.

In contrast with aforementioned approaches, this work is highly focused on applying duration knowledge for large vocabulary speech recognition, and it attempts to integrate duration closely within the core of the recognition algorithm. Experiments are trained on a database of large vocabulary multi-speaker spontaneous, continuous speech, which, again, has not been a focus of previous research. The only other instance of using spontaneous speech was for synthesis applications [3]. In fact, our work is one of the first known instances of a complex duration model for speech recognition

above word level.

developed from continuous spontaneous speech.

The novelty of the duration model stems from its basic paradigm, ANGIE, which naturally accommodates for the hierarchical nature of speech timing effects. Its power lies in its ability to extract and model context-dependent durational information at morphological, syllabic and phonological levels simultaneously. This is made possible by a unique normalization process which corrects successively for various contextual effects at the different hierarchical stages. In addition, its ability to quantify speaking rate [4] provides a valuable tool for investigating temporal phenomena in speech. We will demonstrate the effectiveness of this duration model by applying it in a phonetic speech recognizer and a word-spotting system which will be indicative of its potential success for a word recognition system.

In the next chapter, we introduce the ANGIE structure and detail the fundamental components of our duration model. In Chapter 3, we present a series of experiments which take advantage of ANGIE to analyze speech timing effects. Then, in Chapter 4, we consider incorporating the duration model into a phonetic recognizer and discuss implementation issues concerning this, while Chapter 5 describes a wordspotting task augmented by our duration model. Finally, Chapter 6 draws conclusions and points to future directions for research.

[4] This will be elaborated in following chapters.

Chapter 2

Hierarchical Duration Modelling

This chapter introduces the ANGIE structure in Section 2.1 and explains how this novel framework is used for capturing durational phenomena. In Section 2.2, we embark on explaining the fundamental basis for our duration model. We begin by describing the relative duration model and its normalization scheme in Section 2.2.1. Then, we introduce in Section 2.2.3, the relative speaking rate parameter which is a natural progression from our normalization stategy. We will then consider the use of this speaking rate parameter to build speaking rate normalized absolute duration models in Section 2.2.4. Finally, Section 2.3 will give an overview of the experimental approach and describe the corpus we have chosen to use.

2.1 The ANGIE Framework

ANGIE is a paradigm which captures morpho-phonemic and phonological phenomena under a hierarchical structure. It incorporates multiple sublexical linguistic phenomena into a single framework for representing speech and language. Together with a trainable probabilistic parser, this framework has been adopted in multiple tasks such as speech recognition and letter-to-sound/sound-to-letter generation. These are described in [30]. Here, we develop a duration model based on the subword parse trees provided by ANGIE with the intention of integrating the duration component with the ANGIE speech recognizer. As we shall see, the subword parse trees, provided by ANGIE, are well-suited for constructing complex statistical models to account for durational patterns that are functions of effects at the various linguistic levels.

Context-free rules are written by hand to generate a hierarchical tree representation which is then used to train the probabilities of the grammar used for various applications. A typical ANGIE parse structure, shown in Figure 2-1, consists of five layers below the root SENTENCE node. Each word in the sentence is represented by a WORD node in the second layer, and the remaining layers represent morphology, syllabification, phonemics and phonetics respectively. Thus far, the only purpose of

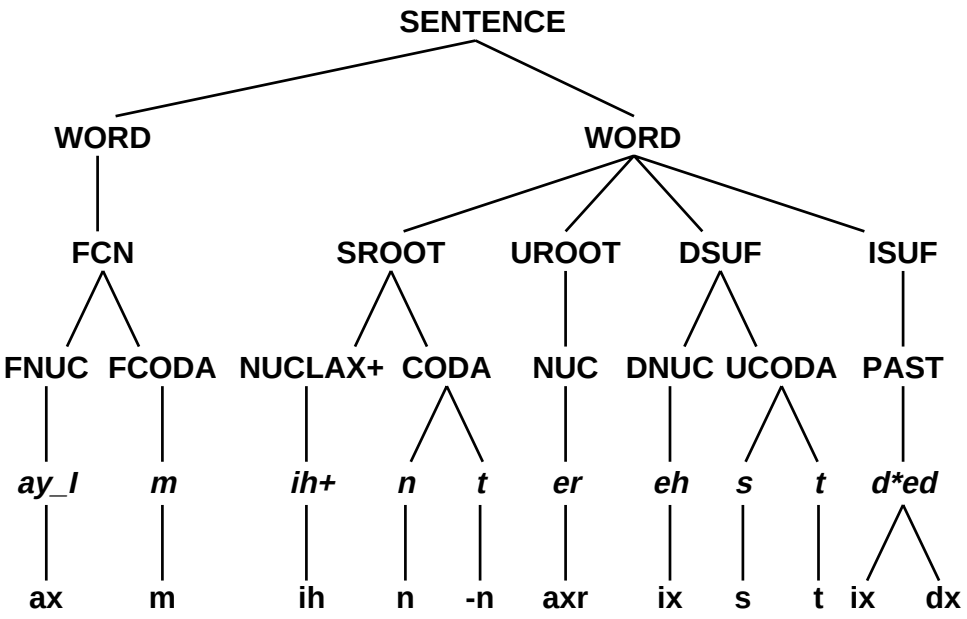

Figure 2-1: *Sample parse tree for the phrase "I'm interested...".*

the WORD layer is to delimit word boundaries, but it could in principle become, for example, a syntax layer. For the layers below, there exists a lexicon of sublexical units associated with each layer. Linguistic categories are chosen in order to optimize probability modelling and do not strictly adhere to those defined in generative phonology. The particulars of the grammar are subject to change, but we give here some examples from the current grammar to elucidate the framework. A comprehensive list of all sublexical categories used in our experiments and a brief description of them are provided in Appendix A. By morphology, we refer to nodes such as SROOT for stressed root, UROOT for unstressed root and DSUF for derivational suffix. The syllabic layer is represented by parts of the syllable, such as NUC+ for stressed nucleus and CODA. These are mostly generic but there are also over twenty special inflexional suffixes. The lexicon is organized in terms of phonemic baseforms for each word. There are approximately 100 unique preterminal phonemes. Lexical stress information is explicitly conveyed throughout the morphological, syllabic and phonemic layers. For instance, both stressed and unstressed roots, and both stressed and unstressed nuclei are represented by distinct nodes. Within the set of phonemes, vowels are marked for stress, indicated by "+", and consonants are explicitly marked for the onset position, indicated by "!". In the morphological and syllabic layers, nodes which appear in the context of function words are labelled accordingly. For example, in Figure 2-1, FNUC and FCODA denote the nucleus and coda which appear exclusively within function words. In addition, we also define special phonemes that are unique to certain morphemes (e.g., */d*ed/* for the past tense morpheme "ed" and */s*pl/* for the plural morpheme) or to particular function words (e.g., */uw_to/* for the */uw/* in "to"). Some special phonemes are pseudo-dipthongs such as */aar, aol/* while others are diphones such as */ra_from/* in "from". The

terminal layer is composed of 65 unique phones which are more generic and are not labelled with regards to morpheme or word context. The phone set is constantly evolving and has been chosen empirically by examining phonemic-to-phonetic alignments. Three distinct schwas are currently allowed – retroflex (*axr*), front (*ix*) and back (*ax*). Gemination and deletions are explicitly marked by a "-". For example, the "-" preceding the phone *n*, in Figure 2-1, indicates that its corresponding parent phoneme *t* has been deleted. The *n* marks the phone prior to the deleted phoneme. In the event of gemination, the phone in word-initial position is labelled with a preceding "-" to indicate that the previous phone in word-final position associated with the previous word is identical, and the two can be regarded as one phonetic unit.

A *parse* proceeds bottom-up and left-to-right. Each column is built from bottom to top based on spacio-temporal trigram probabilities. The terminal category is first predicted based on the entire left column. The prediction of a parent is conditioned on its child and the parent's immediate left sibling, without regard to the column above the left sibling. The linguistic score for a full-column advance is the sum of the log probability for the terminal phone and the log probabilities for the bottom up prediction scores for each column node up to the point where the parse tree merges with the left column. Phonological rules are written without specifying context explicitly. Contexts for which the rules apply are learned, along with corresponding probabilities, from 10,000 utterances from the ATIS corpus. For more details consult [30].

2.2 The Duration Model

Linguistic information at various levels of the phonetic hierarchy is encoded in the durational relationships of subword units. However, in order to extract this information, one must be able to identify and account for all the linguistic factors that operate simultaneously on a segment. For instance, stressed syllables are in general longer in duration than other unstressed parts of a word. But the relative duration occupied by the stressed syllable is also contingent upon other factors such as the number and identity of phonemes within the stressed and unstressed syllables. For example, the two syllable words "cheapest" and "Lima" both consist of a stressed syllable followed by an unstressed syllable. Yet they do not have the same ratio of duration between the stressed and following unstressed syllables, mainly because "cheapest" has a consonant cluster in the second unstressed syllable. We expect that the duration of the unstressed part will be somewhat more lengthened than usual, while the schwa in "Lima" is expected to be very short. Therefore, in order to model duration patterns at a syllable level, it is necessary to compensate for effects operating at lower linguistic levels such as that of the phoneme. If we correct for the lengthening effect of the consonant cluster in "cheapest" by shortening it with some scaling factor, and correspondingly, the vowel reduction in "Lima" by lengthening it with another scaling factor, then these words can be modelled by the

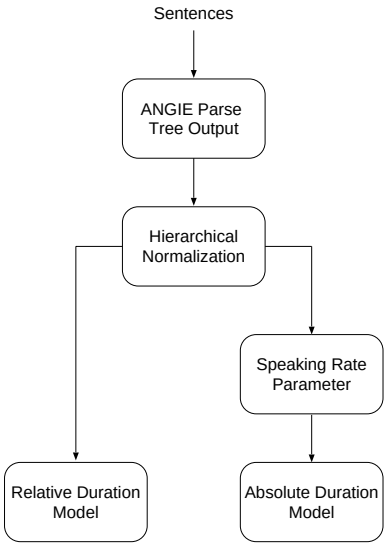

Figure 2-2: *Hierarchical Duration Modelling Scheme*

same probability distribution of two syllable words with an initial stressed syllable and a second unstressed syllable. Essentially, statistical distributions can be collapsed together because sublexical units such as syllables have been normalized in terms of their phonemic realization. Similarly, in order to model phonemic duration, phonemes can be normalized by their phonetic variations. For example, a /*t*/ can be realized as either a flap *(dx)* or a closure followed by a release *(tcl, t)* which tends to be considerably longer in duration than a flap. Both these effects can be corrected for by different scaling factors, so that all instances of the phoneme /*t*/ can be compared against each other in the same distribution. This normalization process can be used successively throughout the linguistic hierarchy and forms the basis of our duration model. Incidentally, when all morphological, syllabic, phonological realizations of a word have been compensated for, the resulting duration, in principle, is one which has been entirely normalized by the variabilities pertaining to linguistic levels below that of the word. And it can be argued that the remaining major source of variability is speaking rate, and therefore the normalized duration is an indicator for speaking rate. We will elaborate on this idea in later sections. The following will elucidate the details of our framework.

The hierarchical framework is utilized to derive two sets of statistical models—one based on relative duration and another based on speaking-rate-normalized absolute duration. To produce these models, there is a two-pass strategy in which, initially, statistics for all nodes are gathered in order to perform a hierarchical normalization. After normalization, we are now ready to collect statistics based on relative duration to formulate the relative duration model. The hierarchical normalization also yields a measurement of speaking rate which in turn is used to build rate normalized absolute duration models. Our overall scheme is illustrated in Figure 2-2.

2.2.1 Hierarchical Normalization Scheme

Within the ANGIE framework, we have formulated a normalization scheme which reduces the model variance at each node and overcomes sparse data problems. Prior to normalization, the duration of each node is given by the total duration of its child nodes while the durations of terminal phone units are obtained from some given time alignment. Our strategy involves a simple scaling of node durations, based on their respective realizations represented by their child nodes, and is propagated from the bottom nodes to the top node in the parse tree.

Basically, given a nonterminal node, its normalized duration is equivalent to the sum duration of normalized durations of its child nodes in the layer immediately below, multiplied by a scaling factor which is predetermined from training data. This factor is a ratio of the mean duration of all instances of the parent node divided by the mean duration of instances of the parent node, when it is realized by the corresponding child nodes. An example is given in Figure 2-3. Here, an instance of the phoneme */d*ed/* is phonetically realized by a schwa *ix* followed by a flap *dx*. Hence, as illustrated by Eqn 2.1 below, its normalized duration is equivalent to the sum of the *ix* and *dx* durations, derived from their time alignment, and scaled thereafter by a ratio, where this ratio is given by the overall mean duration of */d*ed/* over the mean duration of */d*ed/*, conditioned exclusively upon instances where it is realized as a *ix* followed by a *dx*.

$$\mathrm{DUR}_i(\ /d{*}ed/\ ) \triangleq (\mathrm{DUR}_i(ix) + \mathrm{DUR}_i(dx)) \times \frac{\mu_{\mathrm{DUR}}(\ /d{*}ed/\ )}{\mu_{\mathrm{DUR}}(\ /d{*}ed/\ |\ ix,dx)} \tag{2.1}$$

As we adjust all node durations in one layer, we continue upwards to adjust node durations in the immediate layer above, so that this normalization scheme is propagated successively throughout the parse tree.

The advantages of this normalization scheme are twofold:

- Firstly, according to this strategy, individual probability distributions, corresponding to different realizations of the one parent node, are all effectively scaled to have the same global mean. By merging these distributions together, we can construct models which account for various contextual information without the need to split training data, thereby enabling us to

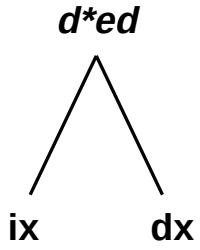

Figure 2-3: *Phoneme /d*ed/ realized as ix followed by dx.*

overcome sparse data problems.

- Secondly, normalization acts to reduce the model variance for each node. In particular, certain realizations of a parent node may have characteristically short or long durations compared with other realizations of the same node. For example, the phoneme /t/ has several allophonic realizations. A specific instance may be realized as a flap dx which would have an inherently shorter duration. Although a particular instance may not be faster than an average dx, it may be considered fast among the total pool of /t/ tokens. In multiplying by a scaling factor and shifting the distributions, we correct for these factors and, as we shall see in Chapter 3, the overall variance is reduced substantially, especially in the upper layers of the parse tree. Furthermore, we hypothesize that each node now has a duration that is independent of the intrinsic durations of its descendent nodes, and we have therefore significantly reduced the number of sources of variability.

2.2.2 Relative Duration Model

Having completed the hierarchical normalization, we are now ready to build statistical models at each node throughout the ANGIE parse tree. The duration of each node is calculated as a percentage of the total duration of its corresponding parent node. Therefore, when a sublexical unit is given by a particular realization represented by its child nodes, we model the relative distribution of duration, occupied among its children. All statistical models utilize Gaussian probability distributions.

Computing a Duration Score

Although we are building models for all nodes in the parse tree, for the purpose of integration with the recognizer (refer to Chapter 4), it is more reasonable that a total duration probability score is proposed at the WORD level. Therefore we must find a way to combine probabilities produced from these submodels based on each sublexical node. During training, we collect statistics for all two-level subtrees with distinct patterns. Because we are dealing with relative duration, only subtrees with more than one child node are modelled. At each subtree, a probability is computed for each child node based on its statistical distribution, and so, we have N probability scores for each of N child nodes. Subsequently, these are averaged together, under the assumption of statistical independence, to yield a score for the entire subtree. In order to derive a duration score for each word in the sentence, it is necessary to combine probabilities of all two-level subtrees within the ANGIE parse tree. This can be done in a number of ways and we will further explore this issue in Chapter 4.

It is of interest to highlight that the relative duration model is limited only to nodes with more than one child node and, for each parse tree, the number of duration scores is based on the number of nodes with more than one child node. All parse trees which only consist of a single branch do not

yield a duration score at all and so cannot be modelled.

Advantages of Relative Duration Model

Our novel paradigm lends itself to many advantages over previous approaches to duration modelling. Firstly, by constructing models at each and every node and subsequently combining them, we implicitly model duration phenomena at multiple linguistic levels simultaneously, and, in so doing, account for various contextual factors that interact and prevail at these hierarchical levels. While, no one linguistic unit can completely depict durational information, researchers in the past [23] have been faced with the problem of choosing a single appropriate linguistic representation to best capture durational phenomena. Here, by using the ANGIE structure, we have eliminated that problem. Also previous attempts to model the vast amount of contextual effects have met with the need to split training data into increasingly smaller cells and consequently, the detail and scope of models were limited by the issue of sparse data.

As our models are probabilistic by nature, we do not impose any arithmetic or multiplicative relationships, or any inherent linearity by way of empirical rules. Nor is it necessary to predetermine which effects or interactions are significant, or which factors actually manifest in the model. These have constituted shortcomings in previous approaches.

Relative duration models are founded upon the hypothesis that proportionate relationships between sublexical units are more important than absolute relationships. We expect relative durations to be mostly preserved in the event of variabilities such as speaking rate changes, in comparison with unnormalized absolute durations. On the other hand, models based on raw durations will penalize particularly slow and fast speakers. This will be further discussed in Chapter 3.

2.2.3 Speaking Rate Parameter

Variations in speaking rate are particularly difficult to deal with for speech recognizers and our work is motivated by the need to account for the natural variations among speakers and for any one speaker within the same sentence. Duration models which do not account for speaking rate variability tend to penalize, incorrectly, speech that deviates from the average speaking rate. The problem stems from three factors:

1. As yet, there is no consensus on a reliable measure for speaking rate.
2. Our knowledge of how rate affects segmental duration is sparse.
3. Rate is difficult to incorporate into a speech recognizer.

The latter two problems will be addressed throughout this thesis.

In speech recognition, the necessary criteria for a speaking rate measure are text-independence and realtime computability. In previous research, speaking rate usually refers to the number of words spoken per minute for a specific passage. This is not feasible for recognition applications because it imposes the restriction of speaker-enrollment. In most cases, the average duration of some linguistic unit over time is taken as a rate parameter. But this also poses a problem. If the size of the linguistic unit is too large, such as a sentence or paragraph, say, then we cannot account for rate variations within the unit. By contrast, a linguistic unit such as a phoneme has an inherent duration that is variable, in spite of a constant speaking rate.

In the following, we propose a parameter that hypothesizes a speaking rate at the end of each word based on our duration model. After we have propogated our normalization up to the WORD node, the resulting normalized duration of this WORD node is expected to be independent of inherent durations of its descendents, and thus is an indicator of speaking rate. Henceforth, we define a word-level speaking rate parameter as the ratio of the normalized WORD duration over the global average normalized WORD duration:

$$\text{Speaking Rate}_i \triangleq \frac{\text{DUR}_i(\text{WORD})}{\mu_{\text{DUR}}(\text{WORD})} \tag{2.2}$$

Effectively, this is a measure of *relative* speaking rate. According to our definition then, a speaking rate measure of 1 indicates an average speaking rate based on training data for a particular word, a measure of greater than 1 indicates slower than average speaking rate and a measure of less than 1 indicates faster than average speaking rate. Note that according to this scale, slow speaking rate corresponds with a large value and fast speaking rate corresponds with a small value.

This speaking rate parameter has the following advantages:

- Computed at a word level, it has the ability to capture speaking rate variations within a sentence.
- By the nature of our normalization scheme, it is independent of all inherent durations of component sublexical units.
- By producing a speaking rate at the end of each word, it is compatible with a speech recognizer such as ANGIE which proposes words one at a time.
- It fulfills the criterion of being causal and text-independent, suitable for recognition.

In fact, armed with this powerful speaking rate measure, we are able to investigate many secondary effects due to speaking rate variability. These form the basis of Chapter 3.

2.2.4 Absolute Duration Model

In the above section, we describe a relative duration model, the core of which exploits the proportionate distribution of total duration among sublexical units within the ANGIE parse tree. In doing so, we have disregarded absolute duration information which may also be useful.

In the following, we propose normalizing absolute durations of sublexical units by dividing by our speaking rate parameter:

$$\text{NDUR}_i(\text{NODE}) \triangleq \frac{\text{DUR}_i(\text{NODE})}{\text{Speaking Rate}_i} \tag{2.3}$$

In the above equation, DUR denotes unnormalized absolute duration of a node. This is simply the sum duration of its child nodes. NDUR denotes the normalized absolute duration which has been scaled by speaking rate. In effect, a rate normalized word has a total duration that corresponds with the average word duration, and this is accomplished via expanding or compressing sublexical durations with one uniform multiplicative factor throughout. Our goal is to compensate for the effects of speaking rate on the node duration and subsequently construct absolute duration models that are rate normalized. This formula is based on the assumption that speaking rate acts linearly and uniformly on each sublexical unit and does not account for any evidence that some nodes are more inelastic to rate changes than others. We will address this issue in Chapter 3.

In developing these models, we expect that absolute duration information will be highly correlated between levels of the tree and it is therefore decided that we will only consider one layer at a time for a set of models. We have selected the terminal phone layer and the preterminal phoneme layer in our recognition experiments, to be discussed in Chapter 4. Statistical distributions are modelled by two-parameter Gamma functions (Equation 3.1) on which we will further elaborate in Chapter 3. To gauge the success of these models, we will investigate the reduction of variance gained from speaking rate normalization in Section 3.1. A large reduction of variance will indicate reliable models. Absolute duration models provide an alternative to the original relative duration model, and it is our hope that their incorporation will augment the final duration model. We will discuss how they will be combined together in Chapter 4.

2.3 Experimental Conditions

2.3.1 ATIS Corpus

Our experiments are conducted on data from the *Air Travel Information System* (ATIS) corpus [38]. The speech data consist of user enquiries related to air travel planning to solve specific scenarios presented to the user. Approximately 5000 utterances, from the ATIS-3 corpus, are used as training data for our model and for analysis in our experiments. This subset consists of 88 speakers and

about 44,000 words in total. There are over 220,000 phones. We have chosen the ATIS December '93 test set of about 1000 utterances and 27 speakers as test data in our recognition experiments.

The reason for selecting the ATIS domain is twofold. Firstly, this corpus consists of spontaneous and continuous speech which is compatible with our goal of developing duration models based on naturally spoken speech, for application to large vocabulary continuous speech recognizers. Most previous research found in the literature has studied read speech or speech produced by a small set of speakers. Experiments undertaken in this domain will provide added insights into the properties of segmental duration for this speaking mode. Secondly, our intention is to incorporate duration into both phonetic recognition and wordspotting using the ANGIE recognizer. For the purpose of evaluation, baseline performances, which have been conducted in this domain, are readily available for comparison with our experimental results. Moreover, ATIS contains a set of city names which is particularly suitable for wordspotting experiments.

2.3.2 Forced Alignment

As training data for our models and for analysis, we have obtained time alignments and phonetic realizations automatically through forced alignment output of the ANGIE system. That is, ANGIE proposes the identity and temporal boundaries of each phone, given the orthographic transcription of an utterance as input. This is necessary because hand-labelled phonetic transcriptions are not available for ATIS. Although alignments produced by the system may conflict with phonetic transcriptions produced by a human expert, we argue that it is more reasonable to extract durational information from alignments generated using the system, if, ultimately, our models are to aid the recognizer when it proposes alignments during real recognition. The system is unlikely to be able to recognize and reproduce segment duration as accurately as that of hand-labelled data and therefore cannot take advantage of all the regularities derived from training on hand-labelled alignments. By training on alignments generated by the system, model variances are better tuned to variability associated with and peculiar to the limitations of the segmentation algorithm of the recognizer.

Chapter 3

Analysis of Speech Timing and Speaking Rate Variability

The previous chapter established the basic paradigm for our hierarchical duration model. In this chapter, we will address several issues relevant to model development, before proceeding to incorporate our model into the speech recognizer in Chapter 4. The purpose of this chapter is to present a series of experiments performed to gauge the effectiveness of the hierarchical duration model at a preliminary level and to investigate speech timing phenomena using the model. We would like to confirm effects which have previously been documented in the literature and also discover new timing relationships using the framework available. These studies will increase our understanding of segmental duration, and the results of our investigations will be incorporated into a final comprehensive duration model.

Initially, this chapter will investigate, in Section 3.1, the effectiveness of our model by computing the reduction in variance gained from normalization. As was described in Chapter 2, our model encompasses two separate normalization procedures, (1) hierarchical, and (2) speaking rate, and both of these contribute to a reduction of model variance which reflects the robustness of the statistical model.

Next, through a series of experiments, we investigate speech timing phenomena and try to characterize them via our model. Specifically, we conduct experiments in three areas:

1. speaking rate,
2. prepausal lengthening,
3. gemination and word-final stop closures.

In order to shed light on the influence of speaking rate on segmental duration, the previously defined speaking rate parameter is applied in several studies in Section 3.2. These consist of a study

of the effect of speaking rate on the relative duration model, an analysis of secondary effects of rate on durational relationships of sublexical units, the variability of speaking rate in our training corpus and a detailed examination of the properties of particularly slow words as defined by our parameter.

Section 3.3 describes studies which aim to characterize phenomena associated with prepausal speech by examining the speaking rate of prepausal words and the durations of sublexical units within prepausal words. It is our hope to identify unique characteristics of prepausal speech, useful for defining a separate model for prepausal data. Finally, we address two contextual effects which take place across word boundaries: gemination and the lengthening of word-final stop closures.

These experiments are motivated not only by the possibility of expanding our understanding of the complex interactions exhibited by speech timing but also by the potential to further incorporate durational knowledge into our model and provide any other added constraints which are not already addressed explicitly by the ANGIE parse tree. For instance, our initial model does not incorporate information that extends beyond the word level, such as prepausal lengthening and gemination. Ultimately, we would like to enhance model performance with these additions.

3.1 Variance Reduction

Model variance is an indicator of how well a model fits the data. Model effectiveness can be described in terms of a reduction in standard deviation. Given that the duration score is based on a Gaussian probability, a smaller variance corresponds with greater constraint and reliability for a given model. Alternatively speaking, a smaller variance is equivalent to greater certainty or confidence in a prediction or estimate provided by a model. In the following, we demonstrate variance reduction achieved for both hierarchical normalization and speaking rate normalization.

3.1.1 Hierarchical Normalization

We consider the amount of variance reduction attributed to hierarchical normalization as described in Section 2.2.1. Normalization is performed for all tokens in the training corpus and for each item in the lexicon of subword units, comparisons between the standard deviation before and after normalization are made and the resulting percentage reduction is computed. The full results are presented in Tables B.1, B.3, B.4, B.6, B.8 and B.10. In order to assess improvement gains at each individual layer, we compare an average standard deviation before and after normalization. This average standard deviation is weighted by the number of tokens at each category at the respective layer. A summary of the results is tabulated in Table 3.1.

The following observations can be made from our results:

- It is evident that variance reduction is more pronounced for nodes higher in the hierarchical structure. This can be attributed to the nature of normalization, which is propagated from the

Table 3.1: *Hierarchical Normalization: reduction in standard deviation for each sublexical layer. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. $\Delta\%$: Percentage reduction of variance.*

Sublexical Layer	Count	μ (ms)	σ_1 (ms)	σ_2 (ms)	$\Delta\%$
Word	43,467	331	180	109	39%
Morphological	62,851	230	105	77	27%
Syllablic	135,841	106	60	48	20%
Phonemic	146,430	99	50	45	10%

bottom up. Naturally, nodes higher up in the hierarchy have a greater number of realizable configurations and so it follows that unnormalizated durations at these nodes have larger variances. Upon normalization, probability distributions pertaining to each of these configurations are combined together, thereby, reducing the variance.

- Examples of probability distributions, before and after normalization for each layer, are included in Figures 3-1, 3-2, 3-3 and 3-4. It is evident that unnormalized distributions are characteristically multimodal in appearance. In fact, they correspond with separate probability distributions associated with different realizations of the particular sublexical node. This characteristic is, by and large, eliminated by hierarchical normalization, through merging these distributions together. The resultant distributions are smoother and therefore, better fitted by the Gaussian function.

- While phonemes which have a greater number of distinct phonetic realizations have higher variances, hierarchical normalization naturally reduces their variances more dramatically by collapsing all these distributions together. This is apparent in function word specific phonemes such as */ra/* in word "from". Here */ra/* can be realized with an initial *r* preceding a choice of vowels: *ah, ax*, or simply a retroflexed vowel *axr*.

Our hierarchical normalization procedure has achieved a large reduction of variance, thereby demonstrating its success. This reduction supports our claim that duration of sublexical units, at various linguistic levels, is better modelled when effects due to linguistic realization at the lower levels are compensated for.

3.1.2 Speaking Rate Normalized Absolute Duration

The absolute duration of sublexical units is scaled by the relative speaking rate parameter defined in Section 2.2.3 to attain rate-normalized duration. In order to assess this normalization, we have collected the statistics for absolute duration for each sublexical unit in three layers: morphological, phonemic and phonetic layers. In the first two cases, speaking-rate-normalization is applied over

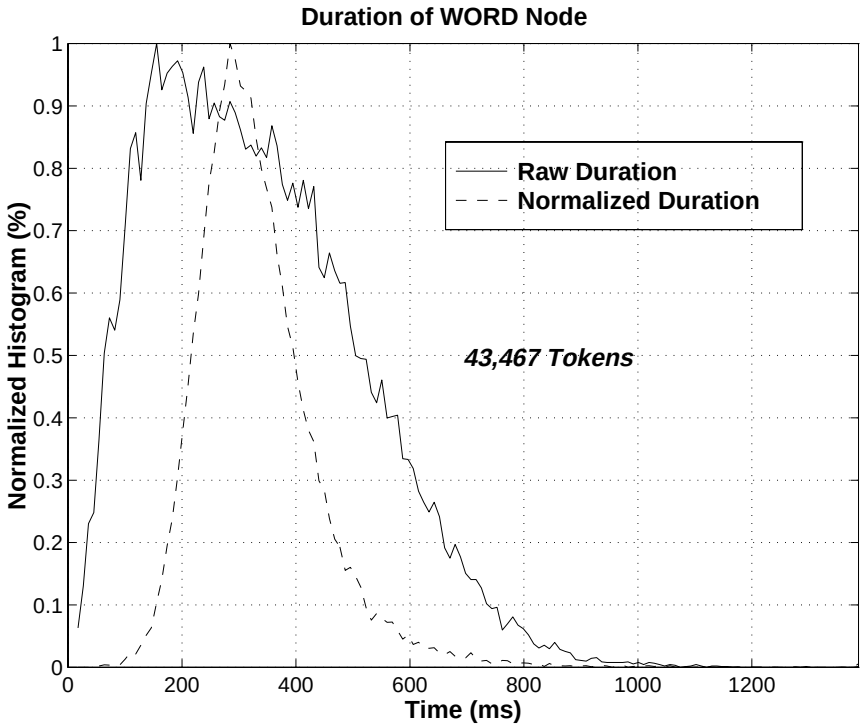

Figure 3-1: *Reduction of Standard Deviation due to Hierarchical Normalization for the* WORD *Node: Mean duration is 331ms. Standard deviation is reduced from 180ms to 109ms.*

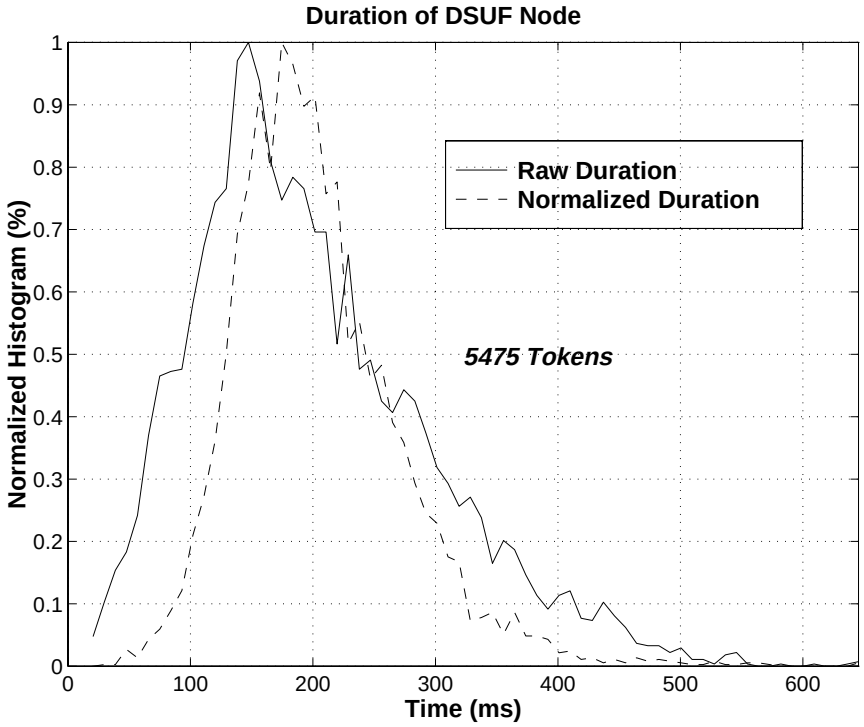

Figure 3-2: *Reduction of Standard Deviation due to Hierarchical Normalization for the* DSUF *Node at the Morph Layer: Mean duration is 200ms. Standard deviation is reduced from 96ms to 67ms.*

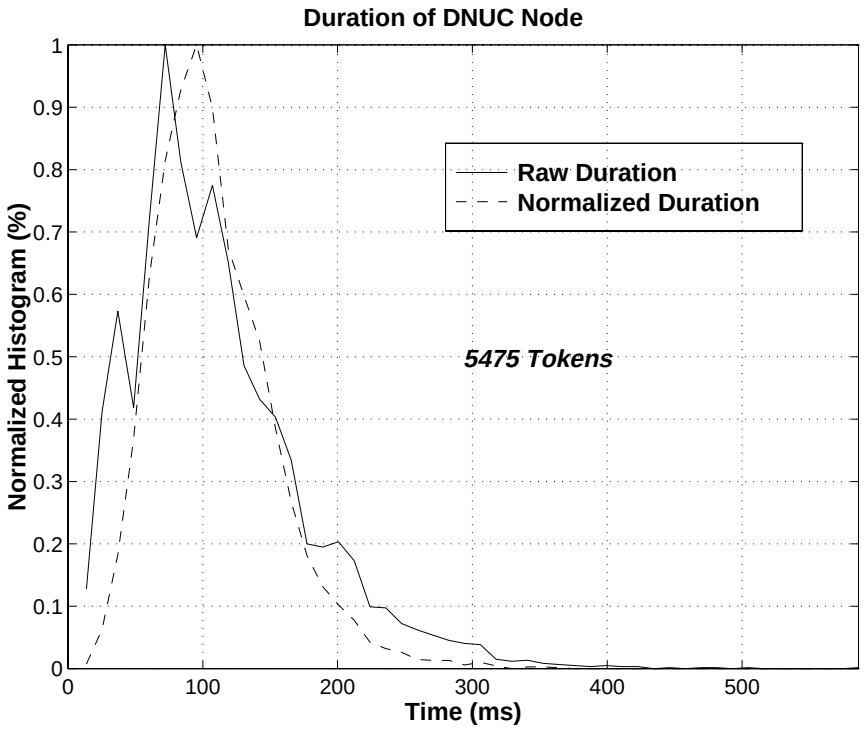

Figure 3-3: *Reduction of Standard Deviation due to Hierarchical Normalization for the* DNUC *Node at the Syllable Layer: Mean duration is 110ms. Standard deviation is reduced from 64ms to 44ms.*

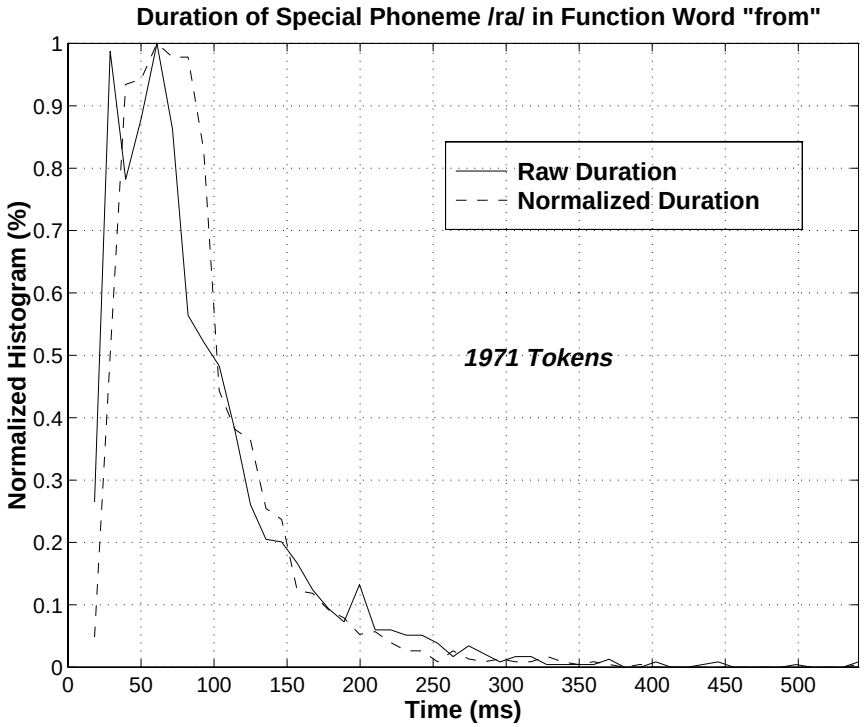

Figure 3-4: *Reduction of Standard Deviation due to Hierarchical Normalization for the Special Phoneme /ra/ in Function Word "from": Mean duration is 87ms. Standard deviation is reduced from 62ms to 51ms.*

and above hierarchical normalization while for the phone layer, rate normalization is applied to raw absolute durations.

One peculiarity of this normalization scheme is that a greater reduction of variance is achieved for nodes which are more likely to have less branching at the layers above them. It follows that nodes higher in the hierarchy must gain substantially greater decreases in variance because they fall under branches which have undergone fewer splits. It also follows that any node which belongs to a single stem tree with no branching at all, will be mapped via normalization to one single duration, that is, the WORD duration associated with average speaking rate. To illustrate this further, if a WORD node is realized by a single SROOT node, the SROOT is deterministically normalized to the average WORD duration. Similarly, if the SROOT node is realized by a single vowel nucleus, then, that child node will also be deterministically mapped to one duration. The first variability is introduced at the first node, as we traverse downwards, with more than a single child node. It is apparent that nodes which are likely to appear in trees with fewer branches will perform better in terms of variance reduction. Whenever a tree contains only one single branch, all durations are mapped to the average WORD duration. More specifically, this problem has two circumstances: (1) For nodes which exclusively occur in single branch situations, the standard deviation is reduced to zero by normalizing. This has the effect of artificially eliminating all uncertainty. (2) For nodes where many instances are mapped to one particular duration, standard deviation is again dramatically reduced with large amounts of data at this one duration. This artificially distorts the statistical distribution and the effectiveness of our probability model is degraded. This problem is most pervasive for function-word specific phonemes and stressed and unstressed vowels in one syllable word contexts. For example, all FCN nodes and phonemes /*ey*/ in the word "a" map to the average word duration. This is because the FCN is always a singleton below the WORD node and /*ey*/, being one syllable, always occurs at the end of a single stem. Many examples of /*ay*/ in "I" also occur in single branch trees and normalize to the same duration. This phenomenon is analogous to a similar shortcoming of the relative duration model whereby subtrees with only single children and therefore single branch trees are entirely excluded from scoring. An absolute duration model with zero standard deviation contains no meaningful information because any duration estimate yields a perfect score. If this perfect score is utilized in the final duration score, it will mistakenly declare greater confidence to the duration estimate. Therefore, this score cannot be used to evaluate a hypothesis and should be discarded.

For the three layers discussed, we have computed the means and standard deviations of each sublexical item after normalization. Because normalization alters slightly the mean duration, it is more informative to speak of a reduction in the ratio of standard deviation over mean. To provide a clearer perspective, this calculation is performed for (1) all nodes in the training data and for (2) only nodes which have at least one split in their ancestral nodes. In the second case, we have

Table 3.2: *Speaking Rate Normalization: reduction in standard deviation for three sublexical layers. μ_1: Normalized mean duration of all tokens. σ_1: Standard deviation after rate normalization of all tokens. μ_2: Normalized mean duration of all tokens except those which are normalized deterministically. σ_2: Standard deviation after rate normalization, discarding deterministic nodes. Δ%: Percentage reduction of the standard deviation over mean ratio for respective normalization scheme.*

Sublexical Layer	Count	μ_1 (ms)	σ_1 (ms)	Δ%	Count	μ_2 (ms)	σ_2 (ms)	Δ%
Morphological	62,851	230	30	71%	35,170	219	45	55%
Phonemic	146,430	99	32	36%	143,410	98	33	33%
Phonetic	184,123	78	32	22%	182,141	78	32	22%

discarded all nodes which map deterministically to the average WORD duration, thereby containing no useful information based on our framework. The inclusion of these nodes may yield misleading results due to large variance decreases contributed by certain nodes. Table 3.2 contains the total reduction in standard deviation over mean ratio for morphological, phoneme and phone layers. It can be seen that a more realistic figure for evaluating the benefit of normalization, at the morph-level, is when deterministic nodes have been discarded even though all function words, for example, have been disposed of. However, this only creates a small difference at the phoneme and phone layers, demonstrating that the incidence of single branch nodes does not pose a significant problem. Note that these are cumulative gains of combined normalization, and comparing them with Table 3.1, it is evident that rate normalization provides substantial additional improvement.

Results for each sublexical item in the morpheme and phoneme layers, in the case where deterministic nodes are discarded, are tabulated in Tables B.2, B.5, B.7, B.9 and B.11.

In addition, the following general observations can be made from our analysis:

- A large reduction of variation is not accompanied by large changes in mean duration. In fact, mean duration for all nodes remains fairly constant. This suggests that our normalization is a reasonable stategy.

- We have implemented a linear speaking rate normalization which underlyingly assumes a linearly proportionate influence of speaking rate on durational segments. However, the resultant gain in variance reduction differs for each category of sublexical units, suggesting that speaking rate affects different nodes to a varying extent. It is observed that gains are mostly positive, indicating that most segments do compress and expand as we expect for fast and slow speech, although they do so in varying degrees.

- At the morph-level, discarding the deterministic nodes implies eliminating all the singleton SROOT nodes and FCN nodes. Even precluding these singleton nodes, SROOT bears the greatest gain in variance reduction.

- Figures 3-5, 3-6, 3-7 and 3-8 plot the normalized histograms of data for some examples of phones and phonemes before and after the combined normalization procedures. Normalization achieves greater smoothness in the statistical distribution and reduces the skewed shape exhibited by the original unnormalized distribution.

- At the phonemic level, the greatest gains in reduction are found, in descending order, for function word specific phonemes, stressed and unstressed vowels, voiced fricatives and nasals. Phonemes such as /*sh*, *l*, *w*, *b*/ yield relatively smaller gains from normalization. They suggest that some phonemes are more inelastic to speaking rate changes. Similar trends are found at the phone layer.

In general, our results indicate that variance reduction due to our combined normalization scheme is large and comparable to the results of previous research. Variance reduction, in the past, has been achieved by explicitly accounting for the large array of contextual factors. For example, Pitrelli [23] derived phoneme standard deviations of 20–30ms in a multiple-speaker corpus. Other results of a comparable nature were derived from using fewer speakers in the corpus [20] and so it can be argued that the inherent variabilities were fewer. On the other hand, attempts to reduce variance through speaking rate adjustment have, generally, failed [33]. Our results have served to demonstrate the potential usefulness of rate-normalized absolute duration models.

Non-uniform Rate Normalization

We have observed that at each layer, gains in speaking rate normalization differ for each sublexical item. While our normalization imposes the assumption that speaking rate influences each sublexical unit uniformly, it is apparent that this is not true and that some sublexical units are more susceptible to speaking rate changes than others. The challenge is then to somehow replace our linear normalization with one which gives greater weight to nodes that, say, expand proportionately more during slow speech and smaller weight to nodes which do not expand or compress as a function of speaking rate.

Our problem is to (1) identify these nodes and (2) determine the relative weights with which we can perform rate normalization. The first attempt has been to use, as a relative weight, the standard deviation to mean ratio, at the phone and phoneme layers, as an indicator for rate dependence. This seems reasonable because nodes with small ratios such as stop releases are mostly inelastic to variabilities such as speaking rate. It is also decided that this ratio is a superior indicator than standard deviation alone because the latter is inherently proportionate with mean duration, that is, longer segments intrinsically have larger standard deviations. This relationship has been established in previous research [23] and appears independent of speaking rate.

Thus, in the event of slow or fast speech, nodes with large standard deviation to mean values

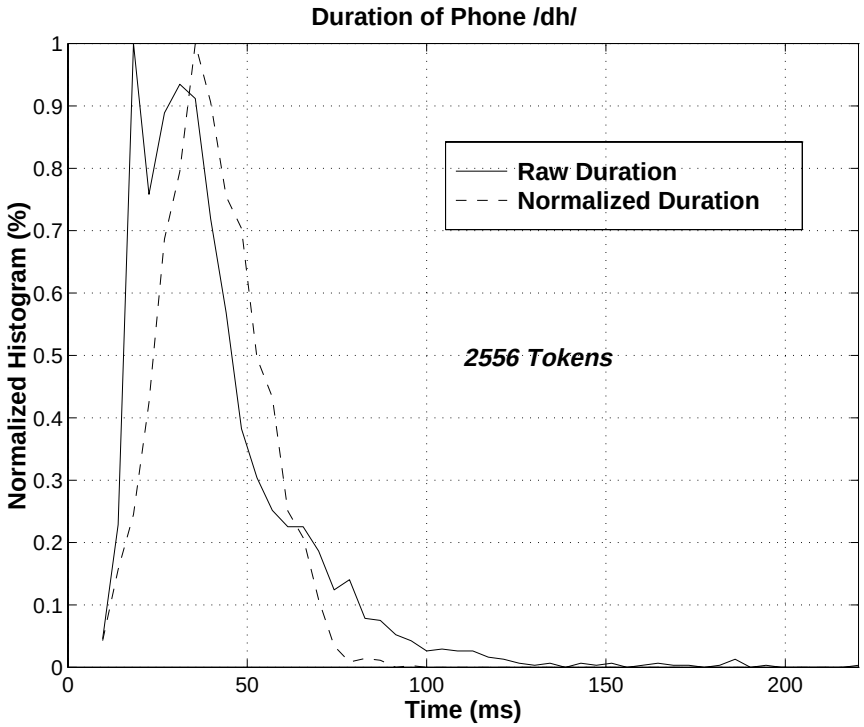

Figure 3-5: *Speaking Rate Normalization for the Phone dh: Standard deviation is reduced from 23ms to 13ms. Mean duration is 40ms before and after normalization.*

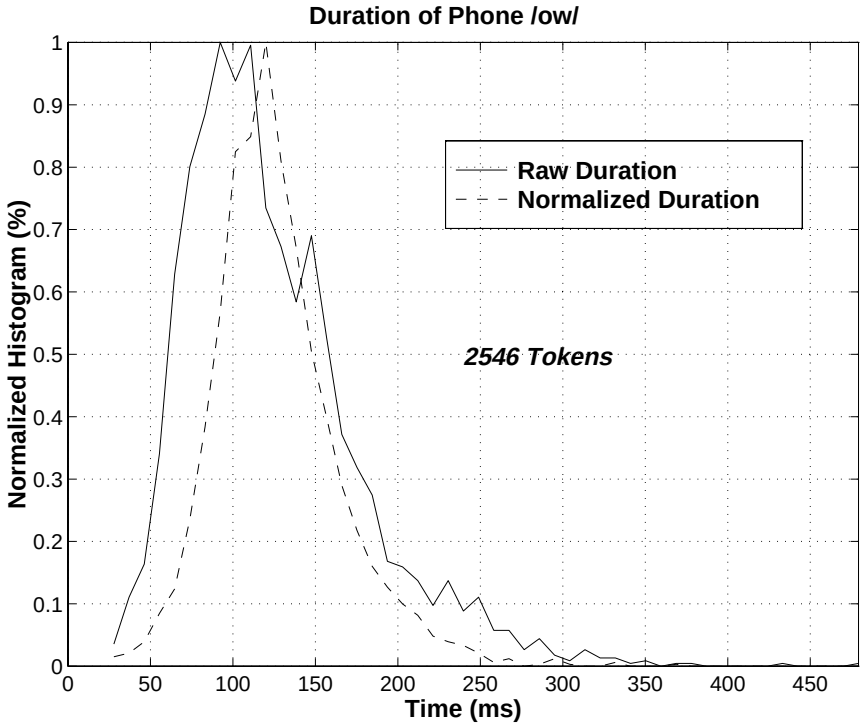

Figure 3-6: *Speaking Rate Normalization for the Phone ow: Standard deviation is reduced from 53ms to 38ms. Mean duration is 124ms prior to normalization and 127ms after normalization.*

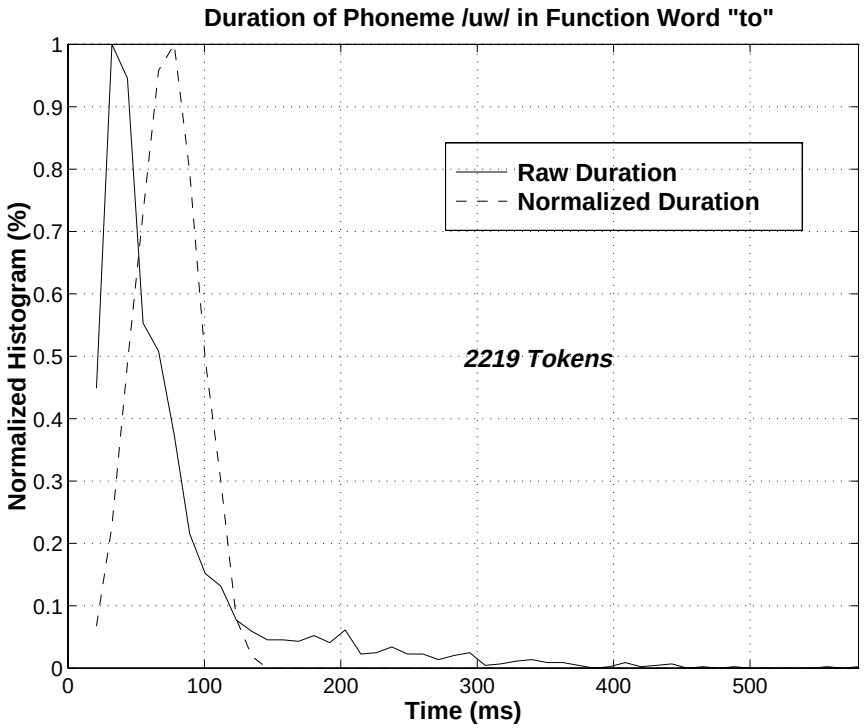

Figure 3-7: *Speaking Rate Normalization for the Phoneme /uw/ in the Function Word "to": Standard deviation is reduced from 68ms to 22ms. Mean duration is 75ms prior to normalization and 73ms after normalization.*

are compensated comparatively more than nodes with small ratio values. As in the linear scheme, the normalized absolute duration of a WORD node has been corrected to correspond to a speaking rate of 1. To produce this, all nodes are compensated for by some scaling to some extent and the relative amount with which a sublexical unit is apportioned depends on its standard deviation to mean ratio compared to that of other sublexical units in that word. By using this ratio, we carry the assumption that all or most of the variability, associated with absolute duration, is attributable to speaking rate.

The results made negligible improvement to overall variance reduction although they did not produce any increase in variance. The lack of success may be explained by a flaw in the underlying assumption of our non-uniform normalization, that is, standard deviation over mean is a good indicator for rate dependence. Variability is contingent upon many factors such as speaker differences. There may exist other mathematical methods more suitable for modelling rate dependence. For example, linear regression can be used to produce best-fitting curves that describe the relationship between absolute duration and speaking rate. In turn, having obtained a mathematical relationship, it is possible to devise a method to correct for the rate effects. However, it must be pointed out that, by normalizing speaking rate directly at the phone or phoneme level without regard for the syllable or morphological context of the phone or phoneme in question, we have deliberately

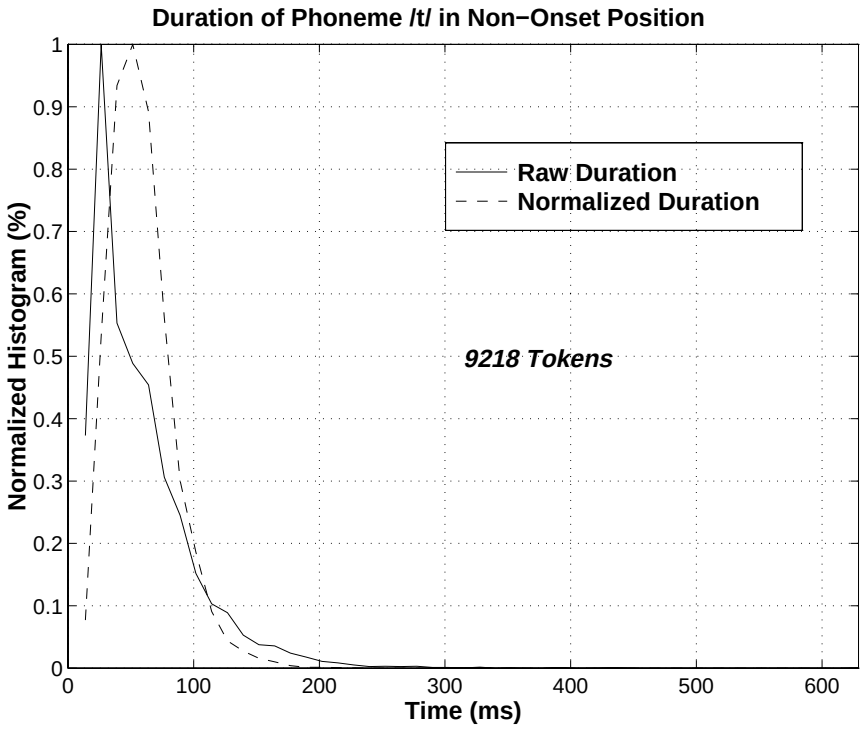

Figure 3-8: *Speaking Rate Normalization for the Phoneme /t/ in the Non-Onset Position: Standard deviation is reduced from 41ms to 25ms. Mean duration is 117ms prior to normalization and 120 after normalization.*

omitted contextual information that may play a role on speaking rate dependence. It may also be beneficial to characterize these nonlinear effects in terms of broad classes or place or manner of articulation. Nonetheless, the knowledge of how rate affects duration nonlinearly can potentially provide significant additional gains to our model. We will further probe this issue in Section 3.2.

Absolute Duration Models

Despite large gains in variance reduction at the morphological layer, absolute duration models are constructed only at the phonemic and phonetic layers. We believe it is most meaningful to use absolute duration at these layers because the lexicon of words is represented by sequences of phonemes or phones. When modelling phonemes, absolute duration models have been corrected for their phonetic realization by hierarchical normlization as well as speaking rate.

The two-parameter Gamma probability distribution function (pdf), given in Equation 3.1 below, is selected as the mathematical model to fit the normalized data.

$$\mathcal{F}_X(x;\alpha,\lambda) = \begin{cases} \dfrac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{,(\alpha)} & \text{if } x > 0 \\ \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

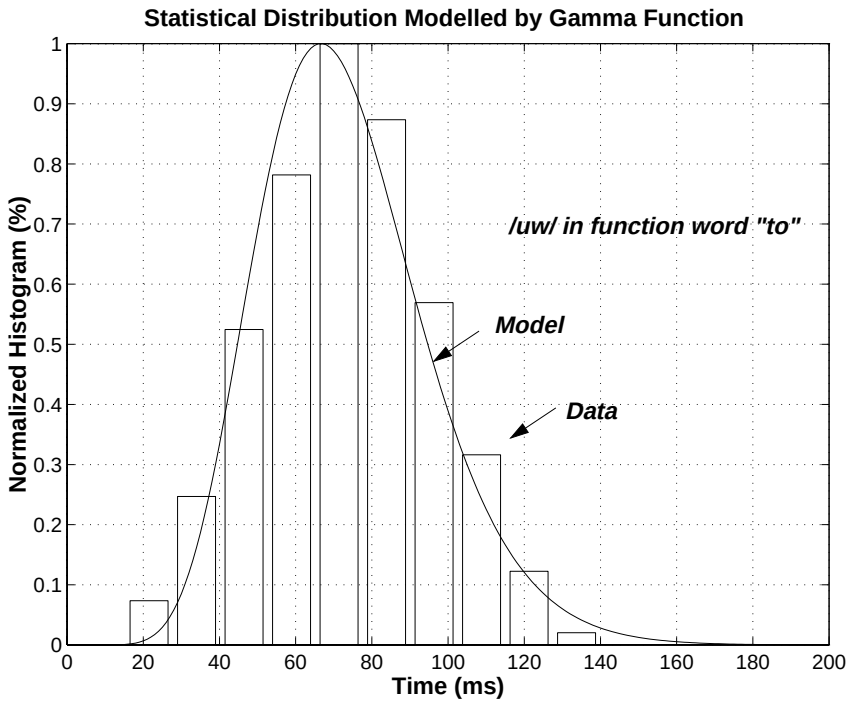

Figure 3-9: *Statistical Distribution and Gamma Model for Phoneme /uw/ in the Function Word "to": The statistical distribution, based on the hierarchical and rate normalized duration, is computed from 2298 tokens. Mean duration is 117ms. Standard deviation is 35ms. For this Gamma model,* $\lambda = 0.1$ *and* $\alpha = 11.7$.

where , (z) is the Gamma function defined as

$$, (z) \triangleq \int_0^\infty x^{z-1} e^{-x} dx \quad z > 0,$$

x is the duration variable, and λ and α are paramaters, adjusted to fit the data. These can be computed from the mean (μ) and standard deviation (σ) of the training data as follows:

$$\alpha = \mu^2/\sigma^2, \quad \lambda = \mu/\sigma^2 \tag{3.2}$$

The Gamma pdf has been used in past duration modelling experiments [6]. Its properties are well-suited for modelling duration. Unlike the Gaussian pdf, the Gamma pdf is not symmetic, with the independent variable being strictly non-negative, and displaying a longer tail that extends towards infinity, as exhibited by many duration histograms. Figure 3-9 overlays the Gamma model on the actual histogram of the phoneme $/uw/$ in the function word "to".

3.2 Speaking Rate Experiments

This section describes a series of speaking rate experiments performed utilizing our speaking rate measure. As previously mentioned, this work is driven by the need to quantify and characterize speaking rate effects. More importantly, in the interest of incorporating more knowledge in the duration model, we are concerned in both the secondary effects of rate and the patterns of rate variability.

Below we initially investigate the influence of rate on our relative duration model in two ways: (1) the effect on the final relative duration score and (2) the effect on relative duration at individual subtrees of sublexical units. Next, we examine in detail the types of words which are manifested particularly slowly, and finally we discuss the consequences of rate variability in a corpus.

3.2.1 Rate Dependence of Relative Duration Model

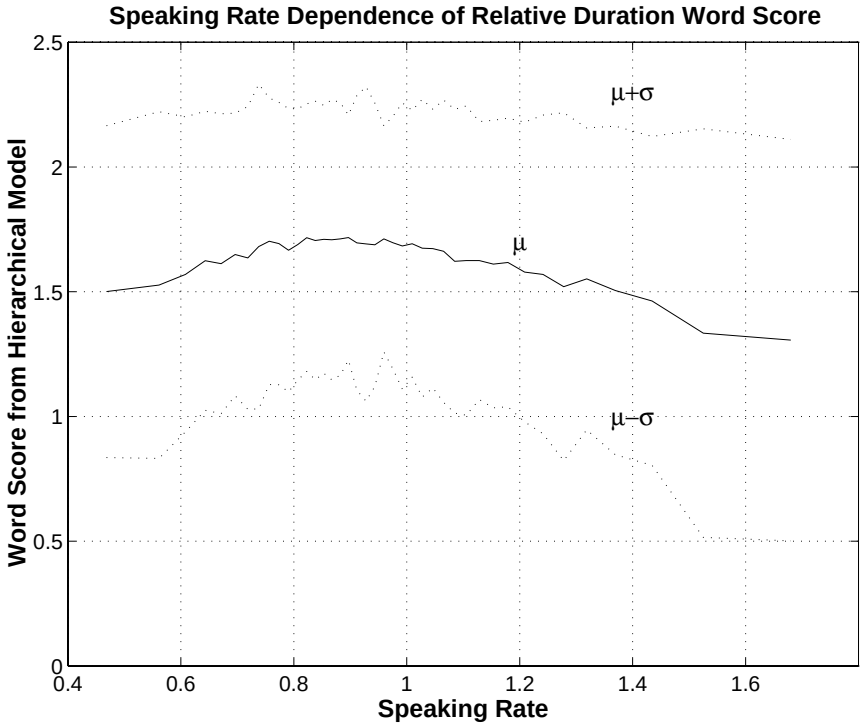

Figure 3-10: *Hierarchical Word Score as a Function of Speaking Rate: All words (43,467) are divided into subsets of 1000 tokens according to their speaking rate. The average word score (μ), calculated from the relative duration model, is plotted against the average speaking rate for each subset. $\mu + \sigma$ and $\mu - \sigma$ represent the boundaries one standard deviation from the mean. The total mean score is 1.54.*

It is an implicit hypothesis that relative duration, unlike absolute duration, is mostly preserved during large speaking rate variations. Here, we test this hypothesis by studying the speaking rate and total duration score for all words in the training corpus. For each word, the total duration

score is computed by averaging all the scores available from each subtree throughout the ANGIE parse tree. If the word is only a single branch, then no relative duration score is available and the token is discarded. Scores are based on log probabilities with an additive offset of $\log 2\pi$. Words (43,467 tokens) are arranged in order of speaking rate and partitioned into subsets of 1000 tokens. The mean score and standard deviation, and average speaking rate are derived for each subset and subsequently plotted in Figure 3-10. This plot reveals the trend in duration scores as a function of relative speaking rate. It indicates that score is relatively stable with respect to rate changes and slow words are not penalized with catastrophically poor word scores. On the other hand, average score is clearly higher where speaking rate is close to 1 and both very fast and slow speech yield slightly lower scores. Moreover the standard deviation is also higher for slower words. This experiment has furnished us with two facts:

1. Our relative duration model is sufficiently reliable under various speaking rate conditions and it achieves some degree of independence with respect to speaking rate.

2. However, there may be some conditions where durational relationships are altered significantly by speaking rate and it would be of interest to discover and characterize these. We investigate this in the next experiment.

3.2.2 Secondary Effects of Speaking Rate

In order to detect trends of relative duration with respect to speaking rate, we examine the relative distribution of duration among sublexical units for slow, medium and fast rates of speech. To do this, the pool of training data is partitioned into three equal sized subsets of slow, medium and fast words, from a total of 43,367. The average speaking rates for these subsets are 1.35, 0.95 and 0.70 respectively. We examine all the subtrees with more than one child node that are common in all three sets, comparing any changes in relationships as speech rate is manifested from fast to slow. For each two-level subtree, the statistics for relative duration among the child nodes, in each subset, are evaluated and the total absolute duration for the parent unit at each subset is computed. Durations of child nodes have been normalized under the hierarchical framework.[1] Subtrees with fewer than 20 tokens in the training corpus are not considered due to the lack of reliability of sparse data. A vast amount of data is generated from this experiment and a handful has been selected pertaining to each level of the ANGIE parse tree, for analysis. Examples are chosen to provide an overview of the nature of effects induced by speaking rate changes. The following is a summary of our observations. We will provide some qualitative discussion as well as present some quantitative analysis.

[1] The absolute duration of the parent unit is the sum duration of its hierarchically normalized child nodes, prior to any further hierarchical normalization that is usually performed on that node.

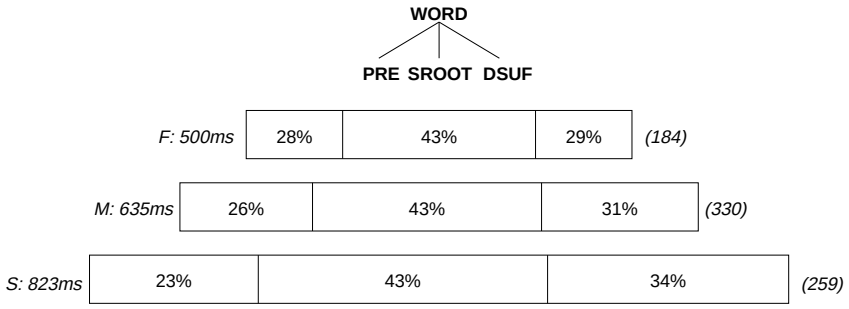

Figure 3-11: *Relative Duration for Parts of a Word Corresponding with 3 Speaking Rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example of this word pattern is in the word "December".*

Word Level: It is found that whenever a WORD is realized by the sequence (PRE SROOT DSUF), the DSUF node progressively occupies proportionately more of the WORD duration as the WORD duration lengthens, that is, as the word speaking rate slows. This is illustrated by the diagram in Figure 3-11. On the contrary, the PRE node occupies proportionately less of the total WORD duration. Its absolute duration only expands slightly. The percentage duration of which an SROOT node occupies its parent remains constant. Therefore, the stressed root must change exactly linearly with respect to speaking rate. These results are confirmed statistically significant at the level $p = 0.01$. Similar evidence of nonlinearity is found for WORD nodes that are realized by sequences such as (PRE SROOT ISUF), although behaviour is not identical. Here, the ISUF also expands proportionately more as speaking rate slows but, in contrast, the SROOT occupies proportionately less in slow words while the PRE occupies proportionately more. It can be inferred from these two cases that there is a tendency to lengthen comparatively more the suffix part of a word in slow speech.

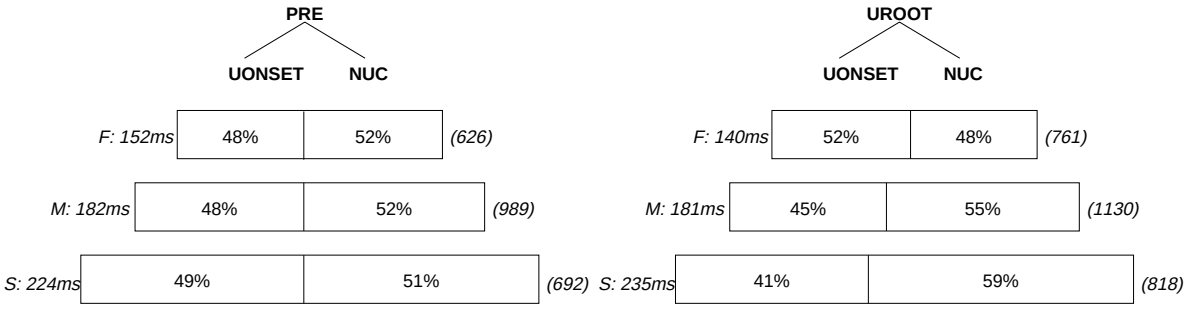

Figure 3-12: *Relative Duration for Parts of a Morph Unit Corresponding with 3 Speaking Rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. Two contrasting examples of (*UONSET NUC*) are given. Examples are the prefix and unstressed root in the word "tomorrow".*

Morph Level: It is found that a subtree with child nodes given by the sequence (UONSET NUC) exhibits different behaviour depending on the category of the parent node. In other words, the

unstressed syllable represented by an onset consonant followed by a nucleus vowel displays different behaviour depending on its position in the word. In Figure 3-12, the relationships remain constant for the case of a PRE node but in a UROOT node, the nucleus absorbs most of the expansion as speech slows down. Statistical significance is found at $p = 0.01$. In addition, constant proportions are maintained when the equivalent lexical units appear under a FCN node while the unstressed nucleus again expands proportionately more when it is within the context of a DSUF node ($p = 0.01$). In the case of the sequence (ONSET NUC+) under an SROOT node, the proportions remain constant for all speech rates. In conclusion, the consonant vowel sequence exhibits different behaviour under various linguistic contexts such as position in word and lexical stress. In some cases, linearity, meaning a uniform effect due to speaking rate is found while in many others it is violated and proportionate relationships are not preserved upon rate changes.

Syllable Level: Parts of the syllable manifest varying degrees of non-uniform behaviour for varying speech rates. For example, in two-consonant clusters containing the phoneme /*s*/, this /*s*/ expands proportionately more as speech slows. For consonant clusters in the onset position, some relationships are constant (e.g. /*p, l*/), and in others, the first consonant expands proportionately more, (e.g. /*f, l*/, /*t, r*/, /*t, w*/).

Phoneme Level: It is found that all stop phonemes consistently exhibit the same phenomenon: the closure expands proportionately more as speech rate slows. This implies that stop releases are relatively inelastic to speaking rate changes and their absolute durations change only slightly while the closure expands and compresses more responsively according to speech rate. This is depicted by Figure 3-15. For the purposes of display, all stop phonemes in the onset position are partitioned into 5 equal subsets of data in order according to speaking rate and the percentage duration occupied by the closure is plotted. This shows a clear and consistent trend, though the degree of change varies for each stop.

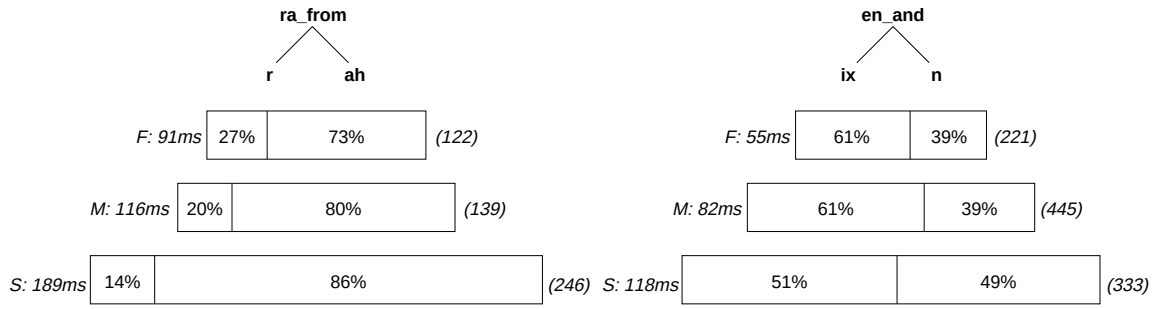

Figure 3-13: *Relative Duration for Parts of a Phoneme Units Corresponding with 3 Speaking Rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. Examples are function-word specific diphones in the words "from" and "and".*

Some word specific phonemes that are diphones or triphones are found to indicate strong signs

of nonlinearity with respect to speaking rate. For example, in Figure 3-13, in the sequence /*r, ah*/, the vowel virtually absorbs all the expansion as speech slows down. For all phonemes /*en*/ in the function word "and", the nasal tends to expand dramatically more than the preceding vowel as speech slows. All results are statistically significant at level $p = 0.01$. In addition, the special diphthongs, in stressed and unstressed contexts, also exhibit nonlinear trends. For phonemes which may be realized as a vowel-semivowel sequence, (e.g. /*aol*/, /*ehr*/ and /*ey*/), the second phoneme consistently expands proportionately more as speech rate slows, as exemplified by depictions in Figure 3-14.

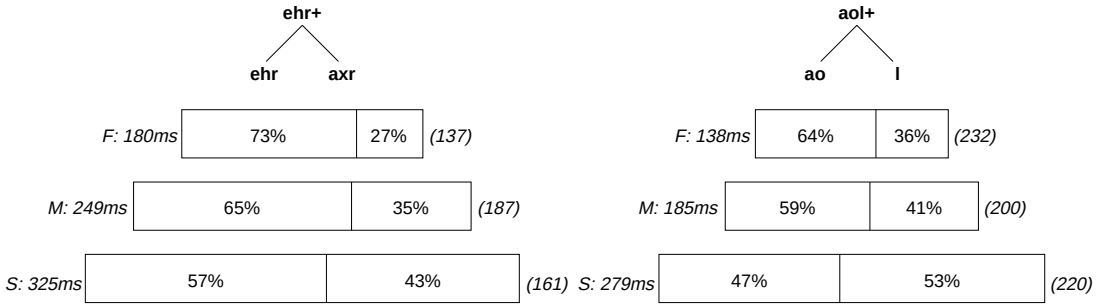

Figure 3-14: *Relative Duration for Parts of a Phoneme Units Corresponding with 3 Speaking Rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. Examples of special diphthongs are given. The phoneme /ehr+/ may appear in the word "air" and the phoneme /aol+/ may appear in the word "all".*

The above results have provided some evidence of the non-uniform effect of speaking rate on sublexical units and some insights on which linguistic units are more susceptible or more resistent to changes. There exist many examples of sublexical patterns whose absolute durations do not change uniformly with speaking rate and hence, their proportionate relationships do not remain constant. Not only is this non-uniformity complex but it is also confounded by many contextual factors which are not yet known and possibly not observed due to limitations in our model. For example, our normalization procedure corrects for effects among layers below the current linguistic level and omits factors which may be a result of the contextual environment at higher levels. Also sparse data prevents us from investigating large numbers of examples of sublexical patterns among the multitude of combinations possible, and therefore it is difficult to find effects that can be easily explained and occur systematically throughout. Hence it is difficult to draw general conclusions or make predictions about the nature of these nonlinear effects, especially without further detailed examination. Ideally, these phenomena should be incorporated into a comprehensive duration model.

We conclude that the relative duration model is quite successful at eliminating first order effects of speaking rate although clearly, it is only an approximation to reality. In fact, this is similar to the absolute duration model where in spite of a clearly nonlinear component, imposing an assumption of linearity in speaking-rate-normalization has reduced speaking rate effects. This leads us to the

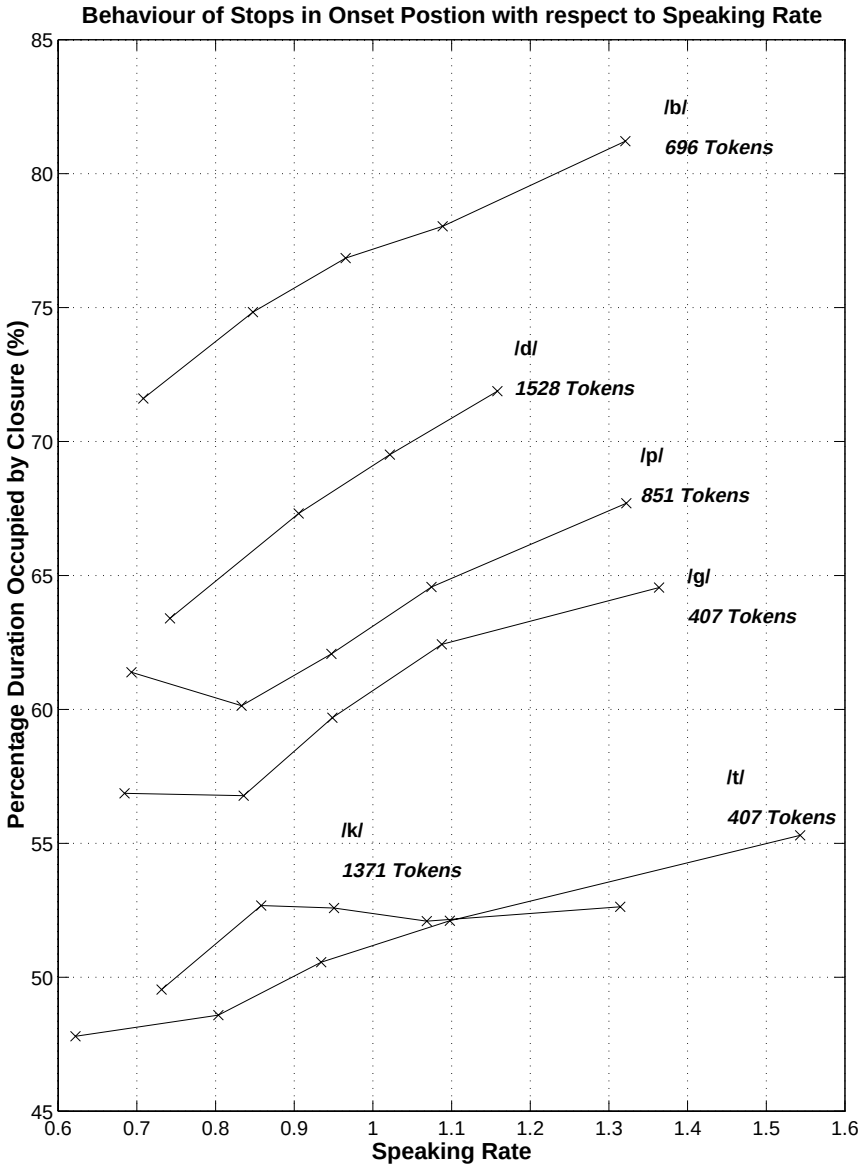

Figure 3-15: *Behaviour of Stops in Onset Position with respect to Speaking Rate: Each of the six stops are divided into five equal subsets of data in accordance with speaking rate of the word they occur in. Average percentage duration occupied by stop closure within a stop phoneme is plotted as a function of average speaking rate for each subset, represented by a "×" on the plot.*

discussion of better ways of modelling these nonlinear effects induced by speaking rate, not accounted for by the current hierarchical paradigm. It is possible, for example, to derive different models for different rates of speech by partitioning training data. The effectiveness of this approach may be limited by sparse data issues and it is unclear if gains will be significant by effectively quantizing rate which is an inherently continuous parameter. In contrast, it may be worthwhile to derive parameters by which relative duration can normalize according to speaking rate. This is consistent with the philosophy that various speech timing effects can be corrected for prior to modelling. This is not a simple task because of our lack of knowledge of the relationship between relative duration and speaking rate, and, again, the problem of sparse data in deriving such parameters.

3.2.3 Variability of Speaking Rate

It is possible that the knowledge of speaking rate can benefit speech recognition performance directly if we are able to utilize constraints regarding rate variability and average speaking rate values. It is clear that extremely slow or fast speech should not be penalized. However, it may be of interest to consider rate variability within utterances. The word speaking rate parameter enables a measurement for rate variability within a sentence to be obtained.

For each utterance in the training corpus, the standard deviation and mean speaking rate and their ratio are calculated and a total average for the entire corpus is computed. This average has been weighted by the number of words in each sentence and is evaluated to be 26.0%. Similarly, the standard deviation of rate, computed globally over all words, over the mean rate of 1.0, is evaluated to be 32.9%. The same calculations are performed for the training data with both sentence-internal and sentence-final prepausal words omitted and the results are evaluated to be 24.6% and 32.0% for the within sentence ratio and global ratio respectively.

These preliminary calculations suggest that rate variability may offer the potential for additional constraints to the duration model. For example, sentences proposed by the recognizer with pathologically large variances in rate may be penalized by the duration component. Further work can probe into the possibility of predicting rate at recognition time and taking advantage of correlations of rate between adjacent words. It is possible that there exist trends such as the tendency to slow down within utterances. Such conjectures require more future experimentation which is not conducted in this thesis due to the constraints of time.

3.2.4 Analysis of Slow Words

In another experiment, a detailed analysis is performed to chart the properties of anomalously slow words. We are driven by the potential for some semantic or syntactic regularities to emerge for words of this pathalogical nature. All words which have a rate greater than 2.5 are tagged and the actual words corresponding with the rates are retrieved. Among this group of words, any time alignments

considered by the author to be inaccurate or incorrect, are omitted from analysis. The result is 102 correctly aligned words with rate greater than 2.5 in the training corpus. The maximum speaking rate is 5. The following sets out the list of these words and their frequency of occurrence.

Function Words				Other Words			
"to"	22	"of"	3	"all"	1	"u"	1
"on"	22	"me"	2	"now"	1	"y"	1
"the"	14	"I"	2	"leave"	1	"o"	3
"from"	6	"for"	1	"four"	1	"which"	2
"is"	3			"two"	3	"nonstop"	1
"are"	3			"newark"	1		

55 (54%) of the 102 words are prepausals with only 2 of them being sentence-final. Most words are function words and only 2 words of all 102 have more than one syllable. These words are characterized by having a relatively small number of phones; most are realized with no more than two phones.

These results indicate that one syllable words have the largest variance in duration. This is the reason that upon normalization, tokens which lie at the tail or the end of the slowest speaking rate are predominantly one syllable words. The behaviour of function words possibly deserves greater study. By nature, they have a variety of phonetic realizations, many of which are reduced to schwas and flaps. On the other hand, their distributions are much broader and as shown here, some tokens are much longer in duration than their average duration. Also on a semantic level, it is interesting to note that speakers do not slow down the most at key content words in a sentence but possibly at the function words just preceding them, often punctuating these words with pauses in between. These conjectures should be confirmed with further analysis.

Information gained here is of value to recognition because it may be used to add syntactic constraint for the duration model. For example, if during recognition a proposed word is anomalously slow, it may be penalized unless it is recognized as a function word or one syllable word. Also any word which has a speaking rate of greater than approximately 5 can be considered to be entirely misaligned.

3.3 Studies Characterizing Prepausal Lengthening

In previous reasearch, prepausal lengthening, in both sentence-final and sentence-internal words, has been found to cause significant duration effects [23, 20, 31]. However, our understanding of the nature of lengthening is sparse, and little research has been directed towards quantifying lengthening and its characteristics, particularly for recognition purposes.

The aim of this research is to study the effects of lengthening on segmental duration and determine

the characteristics which signify the presence of this phenomenon. The ultimate goal is to generate a separate model which accommodates for the unique properties of prepausal data and consequently incorporate this model into the speech recognizer.

We hypothesize that when segments are followed by pauses, lengthening is optional. This is based on the observation that, under many circumstances, lengthening is either notably absent from prepausal segments, or in other incidences lengthening is distinctly identifiable. We suspect that prepausal effects are intrinsically bimodal in that a speaker is likely to either speak normally or exhibit some form of lengthening, before taking a pause. In this case, it is inappropriate to simply construct a model from all tokens that are prepausal but rather, it is necessary to determine some criteria by which we can automatically determine whether a particular segment has undergone lengthening or, more generally, contains durational characteristics peculiar to prepausal segments. Once these criteria have been determined, they can be used to detect such phenomena and we can further examine and quantify their properties.

The following sections will consider factors which may be important when considering prepausal effects. Initially, the speaking rate of prepausal data is examined. Next, we consider the duration of the pause and its bearing on the speaking rate of prepausal data. Then, we will proceed to define some criteria for detecting prepausal effects which may be used later in our prepausal model. Finally, using these criteria, we will investigate durational relationships of sublexical units, pertaining to data which are detected as prepausally lengthened.

3.3.1 Speaking Rate and Prepausal Lengthening

The first experiment explores the relationship between speaking rate and prepausal data. We expect that lengthening of subword units will translate into a slower word speaking rate, detectable by our speaking rate parameter. All words in the training corpus have been ordered with respect to speaking rate and partitioned into 6 subsets. In the forced aligned output, words which are followed by an "inter-word trash" or /*iwt*/ are marked as sentence-internal prepausal words and words which appear at the end of an utterance are marked as sentence-final words. For this experiment, the two types of prepausal words are considered indistinguishable. In Figure 3-16, the percentage of words that are prepausal in each subset is plotted on a logarithmically scaled axis. The size of each subset is approximately equal along this logarithmic axis. There is a clear trend that the fraction of words that are prepausal as speaking rate slows, steadily increases. This implies that it is more likely that prepausal tokens occur among slow words. Alternatively, the likelihood that slower words will be followed by pauses is greater. We can infer from this that a speaker is likely to punctuate slow speech with pauses. In any case, slow speech and prepausal effects are interrelated and speaking rate can be used to signify the presence of lengthening.

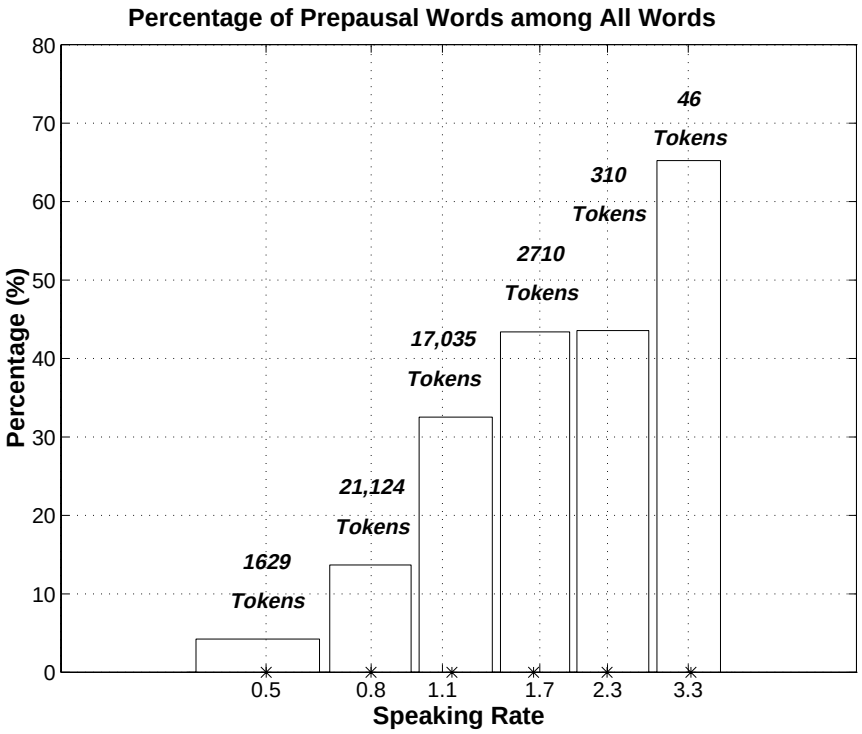

Figure 3-16: *Percentage of Prepausal Words amoung All Words: All words are partitioned into subsets according to their measured speaking rate. The size of each subset is determined empirically and corresponds approximately as equal bins on a logarithmic display axis. For all words in each subset, the percentage of prepausal words is plotted.*

3.3.2 Pause Duration

Having established that slow speech is one characteristic associated with prepausal effects, we will continue to examine prepausal data by considering the duration of the following pause. As the phonetic alignments are generated automatically and no minimum duration is imposed for an /*iwt*/ phone, some /*iwt*/ segments are very short and it can be argued that they do not constitute real pauses for a speaker. In light of this, we suspect that prepausal effects do not take place or exist to a significantly lesser degree when pauses are below a certain duration. And the aim of this experiment is to find some duration below which we can discount prepausal data.

All sentence-internal prepausal words are ordered according to the duration of the following pause. These words are then subdivided into sets of 400 tokens and the average speaking rate for each set is computed. In Figure 3-17, the average speaking rate for each set is plotted against the duration of the pause. This plot shows a dramatic increase in average speaking rate when the pause duration is below about 150ms. It indicates that the prepausal data for which pause duration is very short behaves much like normal data, with close to average speaking rate. For all sets where the pause duration exceeds 200ms, the average speaking rate is above 1.2 and is significantly higher when the pause is extremely long ($> 700ms$). This suggests that prepausal data which are followed

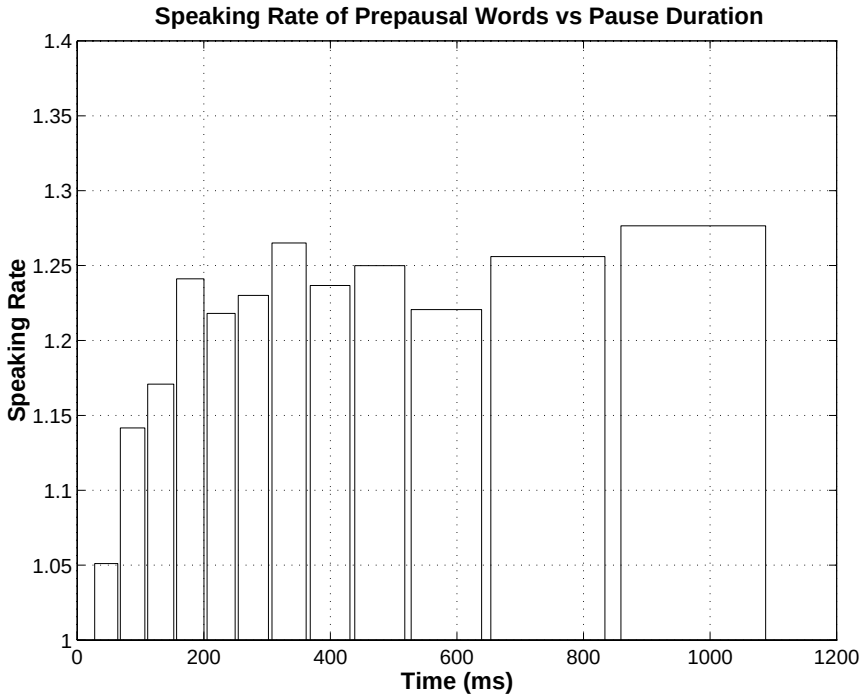

Figure 3-17: *Average Speaking Rate of Prepausal Words vs Duration of Pause: All sentence-internal prepausal words are divided into equal bins of 400 tokens according to duration of corresponding pause. In each bin, the average speaking rate of the prepausal words is plotted.*

by pauses of less than 100–150ms should be discounted from prepausal data and instead, included in the pool of non-prepausal data.

3.3.3 Prepausal Model

We are now ready to define some criteria with which we can associate the manifestation of prepausal effects. The criteria involved must allow the automatic generation of a prepausal model based on the given statistical, hierarchical framework which is capable of separating out prepausal data which behave somewhat differently. Thus, tokens in the training data, which pass these criteria, are tagged as "irregular", as they do not conform with the normal non-prepausal model. Data which do not pass are tagged as "regular", as they are indistinguishable from normal data. The data that are tagged as "irregular" can then be used to train up statistics associated with prepausal behaviour.

From the above experiments, prepausal effects appear to be related to slow speech and secondly, this is more profound when the following pause duration is longer than a minimum threshold. So it seems reasonable to impose the constraint that tokens must be slow and also that the corresponding pause has to exceed some duration. We also suggest that prepausal data should be "irregular" if they score poorly against the general model, trained from the entire training corpus as well as being slow. Scoring poorly implies that the word score falls below some predetermined value. This is a

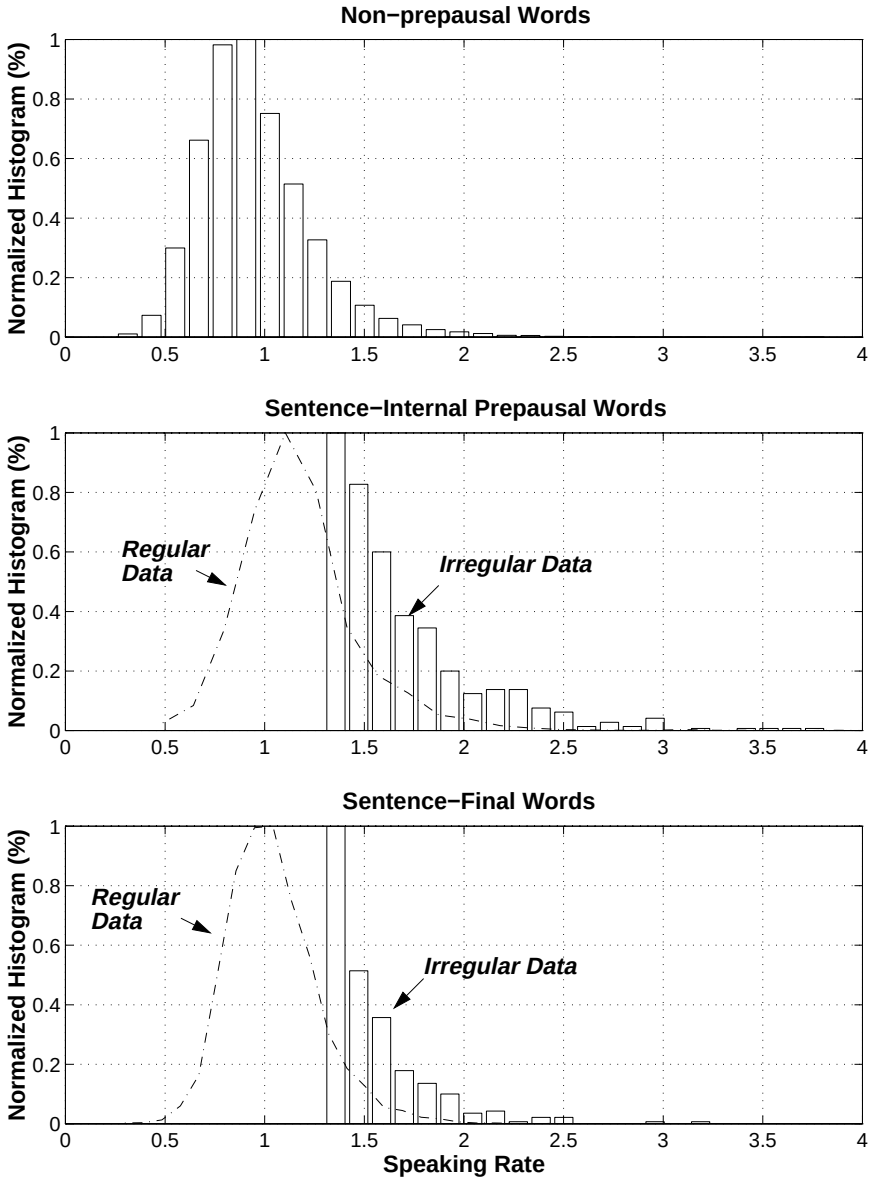

Figure 3-18: *Statistical Distributions of Speaking Rate for Non-Prepausal Words, Sentence-Internal Prepausal Words and Sentence-Final Words. The top histogram indicates the speaking rate distribution of non-prepausal data. For the middle and bottom plot, the histograms depict speaking rate distributions for data tagged as "irregular" while the dotted lines depict distributions for data tagged as "regular".*

Table 3.3: *Comparing the Characteristics of Regular and Irregular Prepausals and Non-prepausal Words.*

	Non-prepausal	Sentence-Internal		Sentence-Final	
		Regular	Irregular	Regular	Irregular
Count	33,760	588	3636	340	4323
Mean Score	1.62	1.72	0.87	1.64	0.71
Mean Speaking Rate	0.95	1.2	1.7	1.0	1.54
Mean Pause Duration	-	496ms	560ms	-	-

simple way of differentiating the disparity between normal and unusual phenonema within the pool of prepausal words. Tokens that score well are omitted because they are already represented by the general model and they do not require a separate model to describe their behaviour. By using word score as a criterion, any "irregular" effects are automatically detected without the need for knowledge of what these effects are in terms of sublexical duration. These can be discovered by examining the "irregular" data.

In summary, prepausal words are considered "irregular" if they:

1. are followed by a pause whose duration exceeds 110ms,
2. have a speaking rate greater than or equal to 1.3,
3. have word scores less than or equal to 1.5.

The above parameters are determined empirically and chosen to best highlight the disparity between the two groups. The word score, as in previous experiments, is determined as the total average of scores derived from the two-level subtrees within the ANGIE parse tree of the word. All scores are log probabilities with an additive offset of $\log 2\pi$. The average word score for the entire corpus is 1.61.

Figure 3-18 depicts the statistical distributions for non-prepausal data and the "regular" and "irregular" groups for prepausal data under sentence-internal and sentence-final conditions. For the 43,467 words, 5045 (12%) are sentence-internal prepausals and 4663 (11%) are sentence-final prepausals. For the 5045 sentence-internal prepausals, 821 tokens are discarded because of the following pause duration. The detailed results are tabulated in Table 3.3.

14% of the sentence-internal prepausals and 7% of the sentence-final prepausals are labelled as "irregular". Results indicate that sentence-final lengthening occurs to a lesser extent and "irregular" sentence-final words are also faster than the sentence-internal counterpart. It appears that "irregular" sentence-internal prepausals are associated with longer pauses, on average, than their "regular" counterparts. It is also found that when evaluating this criterion over the non-prepausal set, only 4.7% of the normal words satisfy the criteria of poor word score and slow speaking rate.

This further reinforces the validity of our choices in defining "irregularity". The "irregular" pool of prepausals can then be used to generate a set of prepausal models.

3.3.4 Secondary Effects of Prepausal Lengthening

The two-level subtree patterns of the ANGIE parse trees which occur in prepausal segments are trained according to the "irregular" criteria determined above and as a result, a set of relative duration statistics that correspond with sublexical elements occurring under prepausal conditions is now available. It is then of interest to examine various durational relationships for these sublexical patterns and compare them with their corresponding duration under normal non-prepausal conditions. The goal is to discover the patterns which are altered more by lengthening and those which are invariant under prepausal conditions.

Only the two-level subtrees which are directly adjacent to the word boundary and therefore, the pause, are considered as prepausal. This is because previous research has suggested that parts of a word which are further away from the pause, such as onsets, exhibit little or no effects [13].

Several examples of sublexical patterns from the word and morph layers are chosen to illustrate the extent of lengthening effects on relative duration. Below, the relative duration and total absolute duration are computed for two-level subtree patterns which occur under non-prepausal conditions and "irregular" conditions. The "irregular" patterns are those that are used to construct prepausal models. To facilitate comparisons, the statistics for the corresponding "regular" patterns are also included.

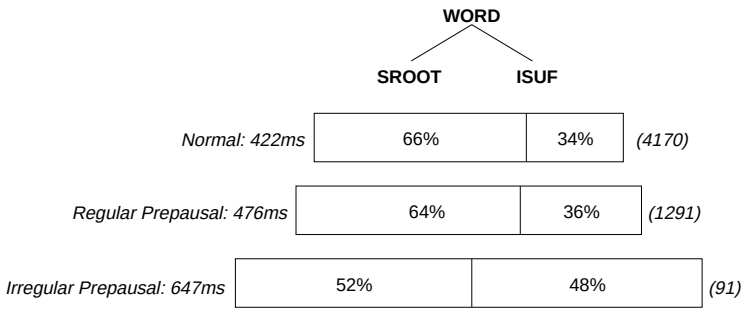

Figure 3-19: *Relative Duration for Normal, Regular Prepausal and Irregular Prepausal Speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example of this word pattern is in the word "flights".*

Word Level: Figures 3-19 and 3-20 illustrate two examples of word realizations. In the sequence, (SROOT ISUF), the ISUF lengthens substantially in proportion for the "irregular" prepausal case compared with both the normal and the "regular" prepausal. The absolute duration of the entire word is also lengthened to a larger extent when "irregular" while the effect is small for "regular" prepausals. Hence the absolute duration of the first syllable only lengthens slightly compared with the inflexional suffix. A similar phenomenon applies for the sequence, (SROOT UROOT). Lengthening effects are

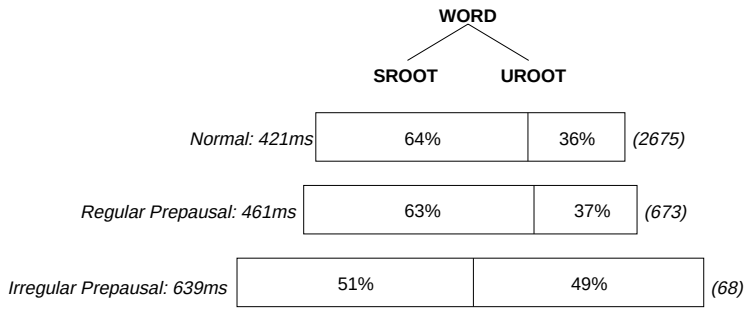

Figure 3-20: *Relative Duration for Normal, Regular Prepausal and Irregular Prepausal Speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example of this word pattern is in the word "travel".*

small for the "regular" prepausal case whereas for the "irregular" prepausal case, the absolute duration expands dramatically. The greater part of the lengthening is occupied by the unstressed root or the second syllable. These results are consistent with the speaking rate experiments in Section 3.2.2 in which it was found that the stressed root expands less compared with following syllables during slow speech.

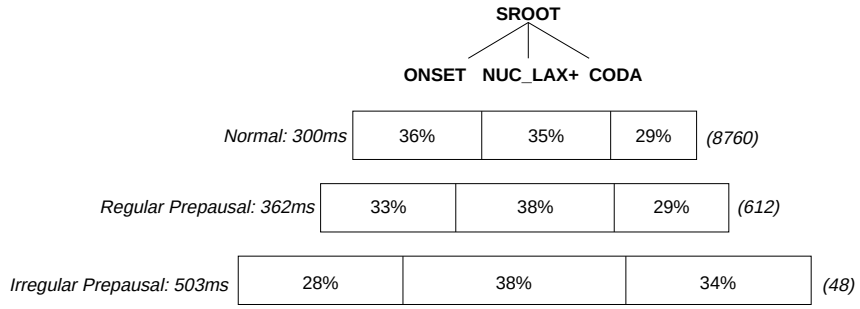

Figure 3-21: *Relative Duration for Normal, Regular Prepausal and Irregular Prepausal Speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example where this pattern occurs is in the word "slip"*

Morph Level: Figures 3-21 and 3-22 are examples of subtrees at the morphological layer. When a stressed root is realized by the sequence (ONSET NUC_LAX+ CODA), the "irregular" prepausal case shows a large lengthening in duration. And this expansion is distributed unevenly among the three syllable components, with the CODA absorbing more of the expansion and the ONSET absolute duration increasing only a small amount. When an unstressed root is realized by a sequence (UONSET NUC), an expansion of the nucleus occurs for "irregular" prepausal tokens as for slow speech. Here, the effect is more dramatic and the absolute duration of the unstressed onset actually falls in the "irregular" case.

Our results have produced some consistent disparities between prepausal patterns and normal patterns. At the very least, there is ample evidence that lengthening occurs as a result of the presence of a following pause, as consistent with the phenomena documented in the literature. The above

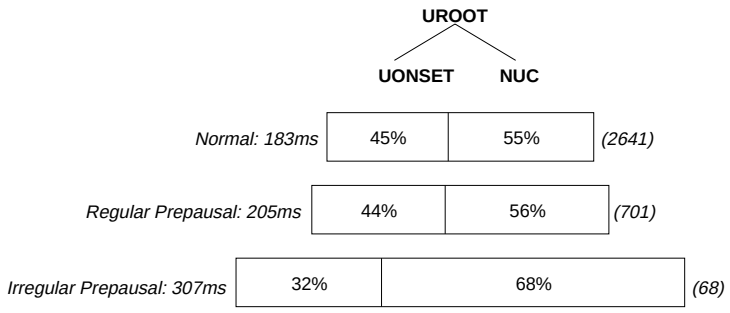

Figure 3-22: *Relative Duration for Normal, Regular Prepausal and Irregular Prepausal Speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example where this pattern occurs is in the final syllable of the word "tomorrow"*

examples illustrate that "regular" prepausal data show some deviation from the population of non-prepausal data, albeit less dramatic than the "irregular" tokens. The "irregular" data show dramatic increases in absolute lengthening, for one, and secondly display substantial nonlinear effects in which some sublexical units are significantly more susceptible to expansion. These manifestations are reminiscent of those that occur in slow speech.[2] It suggests that the criteria used to select "irregular" prepausals have been successful at delineating prepausal patterns that suffer large anomalous effects from those which behave similarly to non-prepausal data. The evidence conforms with the original conjecture that lengthening is a largely optional phenomenon whereby the speaker may choose to speak normally and still punctuate the sentence with pauses. In conclusion, it appears that a complete duration model should describe two types of behaviour—normal behaviour which comprises non-prepausal patterns as well as prepausal patterns that are not affected by the following pause, and an "irregular" behaviour which represents the lengthening effects which sometimes emerge from tokens preceding pauses.

3.4 Word Boundary Effects of Duration

This section will describe experiments pertaining to two specific phenomena which have similar effects on duration and which can be incorporated quite easily over the hierarchical duration modelling framework. We will first discuss gemination and its ramification on phoneme duration. Then we will discuss the effect of stop phonemes that span across word boundaries.

[2] We cannot, however, draw conclusions about the correlation between these nonlinear effects and those observed for slow speech. No attempt has been made to quantitatively compare or make distinctions between the two types of phenomena.

Table 3.4: *Geminate Phones: comparing the mean duration for phones under geminate and non-geminate conditions. μ_1: Mean of the non-geminate phone. μ_2: Mean of the geminate counterpart.*

Phone	μ_1 (ms)	(Count)	μ_2 (ms)	(Count)	$\frac{\mu_1}{\mu_2}$
l	62.3	(5781)	41.2	(47)	1.5
m	56.9	(3900)	41.4	(108)	1.4
n	50.9	(7112)	38.0	(78)	1.3
s	120.1	(6874)	84.3	(178)	1.4
sh	141.8	(1171)	67.8	(39)	2.1

3.4.1 Gemination

Gemination is the phenomenon where two identical phonemes occur adjacent to each other across word boundaries but their acoustic realizations are merged together so that little or no spectral change can be detected. In general, a recognizer would identify a geminate pair as one phonetic unit, even though this unit represents two phonemes. One way to detect gemination is by considering their duration. A geminate pair is usually longer than one single phoneme but shorter than the sum of two.

Our goal here is to study the extent to which geminate phone durations differ from their non-geminate counterparts. Ultimately, we would like to enhance our duration model by compensating for effects due to gemination which affect the first and final phone of a word.

In this experiment, we have collected mean duration for five phones, *l*, *m*, *n*, *s* and *sh*, which occur in both geminate and non-geminate situations in our training corpus. When the ANGIE parse tree encounters a geminate phone in the alignment, it automatically assigns half the duration to the word-final phone and the remaining half to the word-initial phone of the following word. Hence, we speculate that mean duration of geminate phones will be smaller than mean duration of the respective non-geminates but greater than half their value.

Results are tabulated in Table 3.4. As conjectured, non-geminate phones are generally 1.3–1.5 times the duration of their geminate counterpart with the exception of *sh* where the non-geminate is approximately twice the duration of the geminate. Alternatively, when the *sh* phone occurs in a geminate pair representing two phonemes, its total duration is not expected to be any different from when only one phoneme is present. This is perhaps not surprising as we have seen in Section 3.1.2 that *sh* has a smaller standard deviation to mean ratio, implying that it is not as elastic to changes due to phonetic contexts.

3.4.2 Word-Final Stop Closures

From observation, we suspect that stop closures are lengthened somewhat when they appear in word-final position and are followed immediately by stop releases in the word-initial position of the

Table 3.5: *Lengthening of Word-Final Closures: Mean duration (ms) and counts for word-final /tcl/ and /kcl/ when followed, in word-initial position, by the six stop releases compared with all other phones.*

Word-Final Stop Closures	Following Word-Initial Phone						
	p	*t*	*k*	*d*	*b*	*g*	Others
tcl	97.4 (189)	59.1 (10)	82.1 (72)	137.4 (2)	99.4 (10)	97.8 (62)	59.6 (9850)
kcl	71.9 (103)	- (0)	188.8 (3)	- (0)	- (0)	133.0 (3)	58.9 (3583)

next word. This effect is comparable to gemination, in that the word-final phone duration is altered due to a neighbouring phone across a word boundary. Similarly, the motivation is to find out the extent of this alteration and correct for its effect in the duration model. We would like to confirm the phenomenon by measuring the duration of all stop closures which appear in word-final position and by comparing those that precede stop releases with those that are not released.

One issue is that sparse data prevents us from examining the behaviour of all six stop closures. We are only left with *tcl* and *kcl* with a sufficient number of tokens. We select all tokens which are in word-final position and calculate their mean duration in several categories. These pertain to the particular stop release which follows and a collapsed category for all other phones. The results are illustrated in Table 3.5.

The results indicate that all word-final stop closures are lengthened by some amount when the next word begins with a stop release. This holds regardless of a discrepancy in the place of articulation between the closure and the release. In light of the sparse data problem from partitioning the closures into separate categories associated with stop releases, it is more informative to collapse these categories together. In general, all *tcl* followed by releases have a mean of 93.4 ms, 1.6 times greater than that which precedes other phones. Similarly, all *kcl* have a mean of 76.7 ms, 1.3 times longer than that which precedes other phones. In Chapter 4, we will describe how these scaling factors are used to compensate for this lengthening effect in order to further improve the duration model.

3.5 Summary

This chapter has explored many aspects of speech timing phenomena which, when utilized, can potentially increase the predictive power of a duration model. In our initial experiments, it has been demonstrated that the normalization schemes, both hierarchical based and speaking rate based, reduce model variance effectively, comparable to the variances derived from previous research in duration modelling. This leads us to conclude that a combined absolute and relative duration model

may be very useful for imposing additional constraints at recognition time. For an absolute duration model based on phonemes, our model accounts for variabilities derived from speaking rate as well as those originating from phonological realization.

Our series of experiments have uncovered issues not explicitly addressed by both our models. Speaking rate tends to have nonlinear effects on the expansion and compression of sublexical units and as expressed before, there are many contextual factors operating in concert that complicate the degree of influence. Our models compensate for some rate effects by assuming linearity, and it appears that this already provides us with significant gains. These assumptions hold in both the absolute and relative duration models. Nevertheless, much evidence has been presented revealing that speaking rate significantly alters relative duration among sublexical units. Different sublexical categories at various linguistic levels suffer different degrees of change. Similarly, the absolute duration of phonemes and phones also change differently with speaking rate according to their identity. For example, it is yet to be determined whether certain broad classes, manner or place of articulation are more susceptible to speaking rate distortion. Here lies the opportunity for further pursuit in finding an improved characterization of the effect induced by speaking rate.

Studies have also suggested that speaking rate itself can be useful additional knowledge for a complex duration model for a speech recognizer. It is possible to penalize recognize hypotheses if the speaking rate variance is exceedingly large. Given the finding that slow words are predominantly single syllable, multi-syllable words that are anomalously slow can also be penalized.

Experiments have increased our understanding of speech phenomena such as prepausal lengthening and gemination. We have devised a method to detect the occurrence of prepausal phenomena which are only manifested occasionally. Many instances of words preceding pauses reveal behaviour consistent with non-prepausal data. This method of identifying the presence of prepausal events is used to observe effects on relative duration of subword units and we have found nonlinear behaviour as expected. Our experiments in word boundary effects such as gemination and lengthening of word-final stop closures have also revealed consistent behaviour which we can correct for in the duration model.

We will see in the next chapter that knowledge gained from these experiments will be incorporated into the complete duration model in the speech recognizer where we can finally evaluate its utility in terms of performance gains.

Chapter 4

Phonetic Recognition Experiments

Thus far, we have explained the fundamental concepts behind our duration model which is capable of describing both relative and absolute duration, and we have provided evidence that it can potentially improve speech recognition. This chapter will present a set of experiments integrating the duration model with a speech recognition system at the level of phonetic recognition. The goal is to establish the contribution of durational constraints to the improvement of a phonetic recognizer. We will begin by discussing some implementation aspects of the computation and incorporation of the duration score with the recognizer's word score. The scoring mechanism is based on three separate duration scores from the normal, prepausal and geminate models. This will be elaborated upon in Section 4.1.2. We will also address how to combine different scores produced by the relative and absolute models. After presenting the details of our experiment, the chapter will conclude with an analysis of implications of the results.

In principle, durational information is likely to yield substantial gains in terms of recognition accuracy. Often, candidate hypotheses contain implausible durations proposed by a recognizer because the acoustic-phonetic characteristics are well-matched to the recognizer's models. As a consequence, incorrect words are proposed despite the presence of segments possessing improbable durations. Pitrelli [23] analyzed recognition errors and discovered duration as a significant contributor to over half the discrepancies. A duration score will encourage the rejection of faulty segmentations by penalizing unlikely hypotheses and boosting probable ones.

In phonetic recognition, however, the recognizer does not have knowledge of a word lexicon. In the absence of a word lexicon, it is legitimate for the recognizer to propose sequences of phones which would never combine to form an English word. But our duration model's strength stems from the embedded information regarding lexical stress and linguistic context which reside predominantly at syllable and morphological levels. Then, it is of interest to observe how much this duration model is able to assist phonetic recognition. It is our claim that even though the recognizer is not

required to hypothesize higher level linguistic units, lexical information, such as stress embedded in the ANGIE parse tree, is still likely to boost the preference for correct candidate phones. This is because linguistic constraint offered by the ANGIE parse tree alone is sufficient for the recognizer to take advantage of duration information based on the upper levels of the phonetic hierarchy. Moreoever, we claim that duration becomes progressively more useful, if the recognizer is allowed to access more implicit lexical information. In light of this fact, phonetic recognition experiments will be performed given various degrees of linguistic constraint with correspondingly different baseline performances. Our results will shed light on the benefit of duration to recognition as a function of increasing amounts of lexical knowledge.

4.1 Implementation Issues

4.1.1 Integration with the ANGIE Recognizer

In the ANGIE recognizer, words are built bottom up from rules, while tracking the lexicon along the phoneme layer. The lexicon is used implicitly to train the ANGIE subword models, that is, the ANGIE probabilities are trained on a set of phonetic sequences associated with orthographic transcriptions for ATIS sentences.

In past experiments, duration is usually employed as a post-processor in a recognition system where the N best candidate hypotheses are input to the duration component which subsequently rescores and reorders the sentence hypotheses. However, it is more desirable to use duration more actively in constraining hypotheses as they are being proposed. In the ANGIE phonetic recognizer, because of the absence of a word lexicon, pseudo-words are proposed periodically, at which point the duration processor is called upon to process these pseudo-words one at a time. A duration score associated with each pseudo-word is added with an empirically determined weight to the total word score, which is composed of the sum of acoustic and linguistic scores. Note that the core recognizer itself uses a simple phone duration model in its acoustic model as well, and any overall improvement from our duration model is over and above any gains realized from standard models already present in the baseline system.

4.1.2 Computation of Duration Scores

Essentially, the duration score output from the relative duration model accounts for contextual factors within the subword level only, but, as we have seen in Chapter 3, effects such as gemination and prepausal lengthening operate at the inter-word level. These are accounted for explicitly by the scoring mechanism.

The duration processor computes duration scores for three distinct cases—the normal case, the

prepausally lengthened case and the geminate case. Each score is deduced from a separate model. A higher score in any of the three cases will produce a preference for the following phone, which matches the particular condition. The prepausal score is a probability which will express a preference for choosing a following inter-word trash or *iwt* phone. The geminate score encompasses both the geminate condition and also a "closure geminate" condition where a word-final stop closure is followed by a word-initial stop release. Therefore the geminate score is a probability which will express a preference for choosing as the following phone, the corresponding geminate phone or any stop release phone if the word-final phone is a stop closure.

Model for Prepausal Data

The prepausal model is trained from utterances which are designated as being "irregular".[1] All "regular" tokens are treated the same way as all non-prepausal data in the training corpus for the normal model. As determined by experimentation in Chapter 3, "irregular" data are defined as sublexical patterns which both are slow (speaking rate greater than 1.3) and score poorly (score less than 1.5). In addition, the duration of the pause in the training data has to be greater than 110ms. As in Chapter 3, only patterns which are at the rightmost branch of a parse tree are modelled for any prepausal behaviour.

It must be noted that many sublexical patterns do not have sufficient training tokens which are prepausal or "irregular" prepausal. This limits our ability to characterize prepausal phenomena comprehensively. Most of the patterns available reside in the upper levels of the hierarchical tree, while patterns at the phoneme level suffer from sparse data problems. In general, if a pattern has fewer than 10 tokens, the model is replaced by that trained from the entire training corpus.

It is necessary to determine if a hypothesis is "irregular" or "regular" prepausal when scoring prepausal data. Thus, the prepausal score is taken as the maximum of the normal score and the score deduced from prepausal data. Effectively, if the prepausal score is chosen equal to the normal score, no preference is made to either condition. This is because the word is equally likely to be non-prepausal or prepausal, having manifested no observable lengthening, that is, it is a "regular" prepausal. By constrast, the presence of lengthening effects is confirmed if the prepausal score is the greater of the two. By experimentation, it is also determined that better results can be obtained if the prepausal score is only chosen if the speaking rate is greater than 1.3. This is the condition imposed for "irregularity" at training, and it is more reasonable that words faster than 1.3 should not be considered as "irregular" prepausal.

[1] Concepts relating to "regularity" and prepausal lengthening were introduced and discussed in Chapter 3.

Model for Geminate Data

As ANGIE proposes one word at a time, if the final phone is a geminate, under normal circumstances, it would be penalized incorrectly for having an unusually long final phone. A similar effect occurs for word-final stop closures which are followed by word-initial releases. To solve this, each hypothesized word is given a geminate score. The geminate score is the same as the normal score if the final phone is not a candidate for gemination or a stop closure. Otherwise, the final phone duration is multiplied by a predetermined scaling factor.[2] Various scaling factors are determined for each of the following phones: *l*, *m*, *n*, *s*, *sh*, *tcl*, *kcl*[3]. Having compensated for the respective lengthening effect in the last phone, the word can be scored against the normal model to obtain a geminate score. If this geminate score exceeds the normal score, there are several implications: (1) if the previous or word-final phone is *m*, *n*, *s* or *sh*, it is likely that the phonetic unit represented two adjacent phonemes of the same identity and that the next segment will be the *second* phone of the subsequent word (2) if the previous or word-final phone is a *tcl* or *kcl*, then it is likely that the following word begins with any stop release phone. And so the preferences are boosted accordingly.

As for the next word following the previous geminate, the initial phone is also tagged as being geminate and the duration model automatically compensates for this by multiplication with the scaling factor. And this is performed for all three duration scores.

4.1.3 Combining Duration Scores

All scores are targetted towards a mean value of zero, in line with the acoustic and linguistic scores in the recognizer. By substracting the mean score which is predetermined during training, duration scores realize a mean of zero. Segments with better than average duration will yield positive log probability scores.

The duration score is added to the total word score with an empirically determined weighting factor. In the hierarchical model, scores are computed for each two-level subtree pattern, with greater than one child, within the parse tree, and these log probabilities are combined to yield a word score. This is done by a weighted arithmetic average of all the individual log probability scores. The weights allow a bias towards scores at any particular layer of the tree to give them more influence and it has been determined by experimentation that greater weight at the syllable level yields improved results. This could imply that duration information plays a somewhat larger role at the syllabic level. For the case where the parse tree consists of a single branch only, an average score, (i.e., a score of zero), is given, to indicate that the duration component does not have a preference.

In the absolute duration model, the duration score is an arithmetic mean of the log probability

[2] Experiments for determining this are described in Chapter 3.
[3] Other closures are omitted because of sparse training data.

scores for all phones or phonemes. By averaging, word scores are essentially normalized by the number of phonemes and longer words are prevented from being penalized unfairly.

4.2 Details of Experiment

4.2.1 Training

The duration models are trained on the ATIS-3 set from 4644 utterances and the ANGIE parse tree is based on 217 unique subword categories. In the relative duration model, there are 654 distinct two-level subtree patterns. For the absolute duration model, the phoneme model is trained on 94 phonemes. /*ey_a*/ and /*ay_I*/ are omitted from the statistical models because they only appear in single-branch trees. In order to ensure that the duration score is neutral to their occurrence, they are given a mean score of zero whenever they are proposed by the recognizer. There are in total 64 phones in the phone model. Given the large number of patterns in the relative duration model, it is inevitable that some patterns are faced with the problem of sparse data. 64 patterns are based on statistics from less than 10 tokens. In order to minimize any adverse effects, the estimated mean values are calculated from these statistics and the standard deviation is computed as 0.15% of the estimated mean. When the recognizer proposes patterns which are previously unseen, the duration component will output a score of zero to maintain neutrality.

4.2.2 Linguistic Constraint

Experiments are conducted under three separate scenarios of varying levels of linguistic constraint. In the first case, recognition is performed with the sole constraint of the ANGIE parse tree. For the remaining two cases, we have performed experiments by providing the recognizer with the added linguistic constraint given by implicit lexical knowledge. This is accomplished by adopting a two-tiered approach to the lexicon. The word lexicon is represented by sequences of intermediate morph units with which phoneme sequences are associated. The word and morph lexicon are both derived from the ATIS corpus. The recognizer filters partial theories in accordance with morphs available in the lexicon. The morphs provide the extra linguistic constraint by permitting only those syllable patterns which appear in the lexicon. We experiment with this constraint in which random sequences of lexically licensed morphs are allowed regardless of whether they actually appear in this sequence in the word lexicon. This should boost the recognizer's performance because it is only allowed to propose phone sequences that exist within syllables in a vocabulary. In the final scenario, we add another constraint that only morph sequences which appear in the word lexicon are allowed. That is, theories producing morph sequences that do not exist in the word lexicon are pruned. This not only amounts to a reduction of search space for the recognizer, it also approaches word

recognition through adding more lexical knowledge. All the pseudo-words which are proposed by the recognizer are actual words that appear in the word lexicon and, in principle, could be retrieved for word recognition results which corresponds with a word recognizer devoid of a word language model. Consequently, the phonetic error rate should provide a better baseline performance given these linguistic constraints.

Our aim is to observe the amount of improvement that hierarchical duration modelling is capable of offering in all three cases. In principle, greater lexical knowledge represents greater knowledge in the upper echelons of the linguistic hierarchy and consequently duration models based on these linguistic units become more applicable and therefore more helpful, promising greater gains. The experiments will also compare the performance between augmenting with the relative versus absolute duration models. Phonetic recognition results are obtained for the three scenarios without applying duration, and with the relative and absolute duration models applied both individually and combined.

4.3 Results

The phonetic recognizer is evaluated against its own phonetic labels, as obtained during forced alignment of the orthographic transcription. This is the phonetic transcription that the system would need to produce to perform correct lexical access. ANGIE's phonetic inventory consists of 64 distinct units, some of which are context dependent. The forced alignments for this experiment were generated using a system which also employed a duration model, using a combined relative duration and phoneme absolute duration models with a weight of 100.

The test data consist of 905 sentences drawn from the December 1993 test set. In the following, results for the three scenarios are presented for various cases with and without the addition of duration. We compare results using the absolute duration model based on both phones and phonemes and results produced using the relative duration model. Duration weights are empirically chosen and optimal values are different for each scenario. For the case where relative and absolute duration were combined, relative and absolute phoneme duration are not given equal weights, and their relative ratios are also empirically chosen during experimentation.

4.4 Analysis

In general, duration modelling provides some positive gain to all three cases of linguistic constraint. However performances tend to vary for each case. The following conclusions can be drawn:

- Generally, duration modelling is more effective at greater levels of linguistic constraint, as predicted. At best, the duration model offers an optimal 2.3% gain or 7.7% relative gain

Table 4.1: *Results of Phonetic Recognition Experiment Using the* ANGIE *Parse Tree with No Additional Constraint. The percentage error rate with their component substitutions, deletions and insertions are given.* Δ *represents the percentage error reduction from error rate using no duration.*

Scheme	Weight	Error Rate (%)	Subs (%)	Del (%)	Ins (%)	Δ (%)
No duration	n/a	33.2	16.5	11.2	5.6	-
Phones Only	50	33.0	16.5	10.6	5.9	0.6
Phonemes Only	50	33.0	16.1	11.2	5.7	0.6
Relative Duration without Prepausal Models	55	32.7	16.3	10.7	5.7	1.5
Relative Duration with Prepausal Models	60	32.6	16.2	10.6	5.8	1.8
Phoneme + Relative	30 + 30	32.7	16.3	10.6	5.8	1.5

Table 4.2: *Results of Phonetic Recognition Experiment Using Morph Constraints. The percentage error rate with their component substitutions, deletions and insertions are given.* Δ *represents the percentage error reduction from error rate using no duration.*

Scheme	Weight	Error Rate (%)	Subs (%)	Del (%)	Ins (%)	Δ (%)
No duration	n/a	31.8	15.1	12.7	4.0	-
Phones Only	100	31.5	15.11	12.1	4.4	1.0
Phonemes Only	100	31.2	14.6	12.6	4.0	1.9
Relative Duration without Prepausal Models	75	31.0	14.9	12.2	4.0	2.5
Relative Duration with Prepausal Models	75	31.1	14.9	12.1	4.1	2.2
Phoneme + Relative	33 + 66	30.9	14.7	12.1	4.1	2.8

when word constraints are imposed and only 0.6% or 1.8% relative gain when only the ANGIE parse tree is used. This can be attributed to the availability of higher level lexical knowledge, encouraging the recognizer to propose more sensible words whose linguistic properties are more suitable for the duration scoring. Duration weights also tend to be larger, which indicates that the duration component becomes more reliable for evaluating a hypothesis. On the other hand, when these added constraints are removed, the recognizer tends to hypothesize nonsensical pseudo-words which are not real words and are not well-matched to the training data for the duration model. Therefore, even when the phone sequence is correct, the ANGIE parse at the higher linguisitic levels is meaningless and yields poor duration scores. To yield the best performance gains, the duration weight is set comparatively low for the case with the least constraint, inidicative of the limitations of the duration model to boost performance for this scenario.

- For the first case with the most basic level of constraint, the relative duration model appeared to be the most beneficial, whereas the absolute duration model only offers marginal gain. In fact, both phone and phoneme models add very little improvement. Similarly, when morph units are used without word constraints, relative duration yields better performance. When

Table 4.3: *Results of Phonetic Recognition Experiment Using Morphs with Word Constraints. The percentage error rate with their component substitutions, deletions and insertions are given. Δ represents the percentage error reduction from error rate using no duration.*

Scheme	Weight	Error Rate (%)	Subs (%)	Del (%)	Ins (%)	Δ (%)
No duration	n/a	29.7	14.3	10.8	4.6	-
Phones Only	100	28.5	13.9	9.9	4.8	4.0
Phonemes Only	200	27.7	13.3	9.3	5.1	6.7
Relative Duration without Prepausal Models	75	28.6	13.8	10.1	4.7	3.7
Relative Duration with Prepausal Models	75	28.8	14	10	4.7	3.0
Phoneme + Relative	170 + 85	27.4	13.3	8.9	5.2	7.7

full constraints are imposed, the absolute phoneme model offered significantly better performance. The reason for this is possibly due to the greater lexical knowledge which causes the recognizer to choose, in many cases, actual words in which the derived speaking rate parameter becomes more applicable and the rate-normalized phoneme model becomes the most effective. In cases where the pseudo-words are meaningless, the speaking rate parameter refers, then, to a normalized duration which is derived not from a real word at all. It is likely to be an inaccurate reflection of actual word speaking rate, and so the normalized absolute duration may be mismatched to the rate-normalized absolute duration model and hence degrade its performance.

- In all three cases, the phoneme model performs as well if not better than the phone model. This is expected because the words are lexicalized at the phoneme level. And the phoneme models themselves have been normalized by both speaking rate and their phonological realization, accounting for possibly two of their greatest sources of variability. Phone durations by nature suffer more variabilities, and as terminal categories, the phone models embed less linguistic knowledge. In addition, absolute phone models are probably somewhat more redundant with the standard duration models already present in the system.

- Attempts were made to combine the two duration models together and an incremental improvement resulted from using both relative and phoneme model. The relative weight given to each model is different in all three cases depending on how well they perform individually for that case. It is conceivable that combining the models does not offer substantial further improvement because of the level of redundant information between the two models. Future work can concentrate on optimizing parameters for combining these models into one duration score, taking into account the large correlation between them.

- The relative duration model performed better by 0.1% using a separate model for prepausal words when no morphs were employed. For the other two cases, using a separate model for

prepausal words degraded performance by 0.1–0.2%. In all cases, the impact of using a separate model seems minimal. We can speculate that the phenomenon of prepausal lengthening is not dramatic enough to warrant separate modelling. Few prepausal tokens in the test data exhibit large lengthening effects which score poorly against the normal data and therefore, the overall error rate does not improve significantly. Also our prepausal models suffer from sparse data. Many patterns in the training data do not have examples of being "irregular" prepausal and so these were replaced by the model for the normal condition. There are potentially other ways to deal with prepausal lengthening such as compensating for absolute durations of phones by a correction factor prior to scoring. However, this is difficult because it is necessary to determine the number of phones away from the word boundary to correct for. This can be overcome by, say, only correcting for the phones which are pathologically long, up to some predetermined distance from the pause or for the entire word.

In conclusion, this duration model has been successful at reducing error rate and we have demonstrated its utility in phonetic recognition. Our results show that a complex duration model, which implicitly incorporates a large amount of contextual knowledge, has improved performance in phonetic recognition which does not have access to this knowledge. Our duration model has managed to produce gains under all three levels of linguistic constraint although it is clear that as we approach word recognition, duration plays a much greater role.

Chapter 5

Wordspotting Experiments

The first part of this thesis has introduced our hierarchical duration model and in the previous chapter, we have already seen its utility in improving phonetic recognition performance. This is encouraging because durational knowledge can potentially offer even greater gains to word recognition.[1]

As a first step towards demonstrating the benefit of durational information to word recognition, we choose to evaluate our duration model on the task of keyword spotting in the ATIS domain, using an ANGIE-based wordspotting system[2]. Experiments are conducted in three stages. During the initial stage, some preliminary studies are undertaken in order to compare duration scores assigned to alignment outputs of a wordspotting system. In principle, for a duration model with good discriminating power, poorly aligned or incorrect outputs, which constitute false alarms, should, in many cases, score poorly compared with alignments which constitute hits. As an intermediate stage, the duration model is applied as a post-processor. We have selected the specific task of disambiguating between an acoustically confusable word pair in the ATIS domain. The results produced by the wordspotting system are input to the duration model for subsequent rescoring. The effectiveness of the duration model will be reflected by its power to further reduce the total error rate by decreasing the number of false alarms without decreasing the number of hits. This is a pilot study which attempts to illustrate the important role played by duration and its success encourages us to allow durational information to be applied more directly into a wordspotting task. Finally, the duration model is integrated with the wordspotting system with the goal of enhancing the overall performance of the system. We shall see that duration modelling brings substantial benefits to the overall performance of the wordspotting system.

[1] Word recognition experiments are beyond the scope of this thesis although this duration model is entirely applicable to continuous speech recognition.

[2] This system is currently under development by Ray Lau and a detailed description of it will be found in [27].

5.1 Preliminary Investigations

Preliminary investigations compare the distribution of duration scores among correct and erroneous outputs of the ANGIE wordspotting system. A collection of alignments corresponding with hits and false alarms are generated by the wordspotting system from training data and these tokens are analyzed and scored by the duration model. The result is a series of plots comparing the probability distributions of the duration scores of alignments of hits versus false alarms.

This experiment is conducted using a random subset of the ATIS-3 training set consisting of 2182 spotted keywords, of which 1467 are hits and 715 are false alarms. The keywords consist of a list of city names which occur in the ATIS domain. In Figures 5-1, 5-2 and 5-3, the probability distributions of duration scores are plotted in normalized histograms. All the score values plotted here have been subtracted by a mean score value computed during training, that is, scores have been targetted to a mean value of zero. In addition, the relative speaking rates for all spotted words have been calculated and their histograms are also plotted in Figure 5-4.

For all three duration models, scores corresponding with false alarms are, on average, substantially worse than those of hits. In all cases, the distances between mean scores of hits and false alarms are more than one standard deviation of the distribution of hit scores. This indicates that there is some separation between the two distributions. Characteristically, score distributions for false alarms have large standard deviations and many words have anomalously poor duration scores which lie outside the range for the distribution of hit scores. For the speaking rate distributions, some tokens have anomalously slow speaking rates in the set of false alarms. These are noticeably absent for hits.

It is difficult, from these results, to assess the relative effectiveness of the different duration models or to gauge the amount of improvement that duration modelling will contribute to a wordspotter, in actuality. For one, many false alarms yield scores above zero. This implies many of the errors incurred cannot be eliminated through the help of duration because their alignments are not considered poor by the duration model. All distributions tend to have a skewed shape in which a long tail extends toward the negative scores. Many hits also score poorly according to duration and so a large weight given to the duration score could be detrimental to them. However, the plots do indicate that there is disparity between duration scores for correct and incorrect alignments. In particular, tokens yielding extremely poor duration scores are very likely to be false alarms. Similarly, it is very rare for correct alignments to yield a relative speaking rate greater than 2.

5.2 Duration as a Post-processor

In the second experiment, we have chosen to demonstrate the importance of duration by using the duration component as a post-processor in a specific wordspotting task. Two acoustically confusable

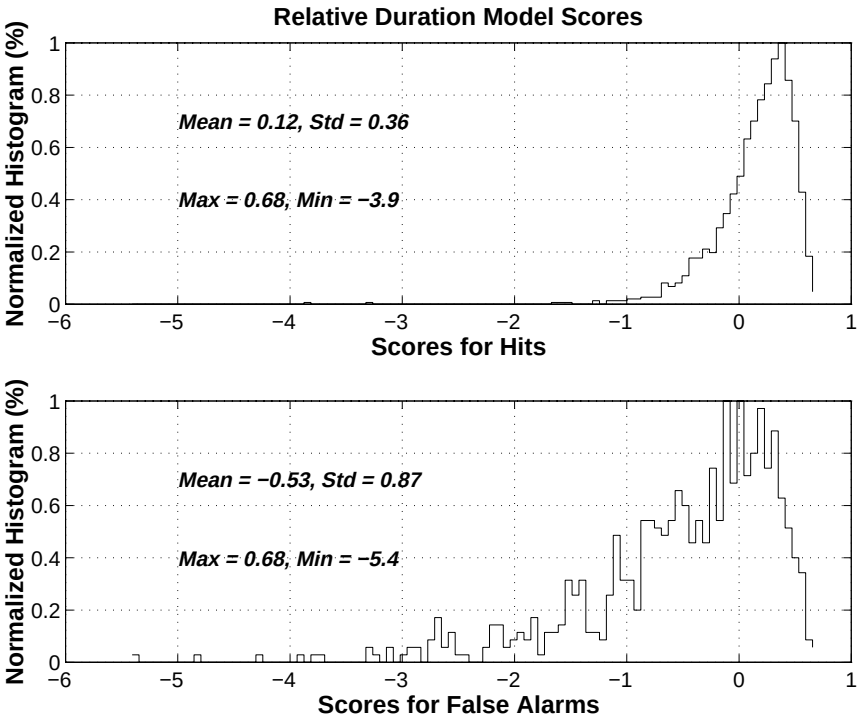

Figure 5-1: *Probability Distributions of Relative Duration Scores for Hits and False Alarms.*

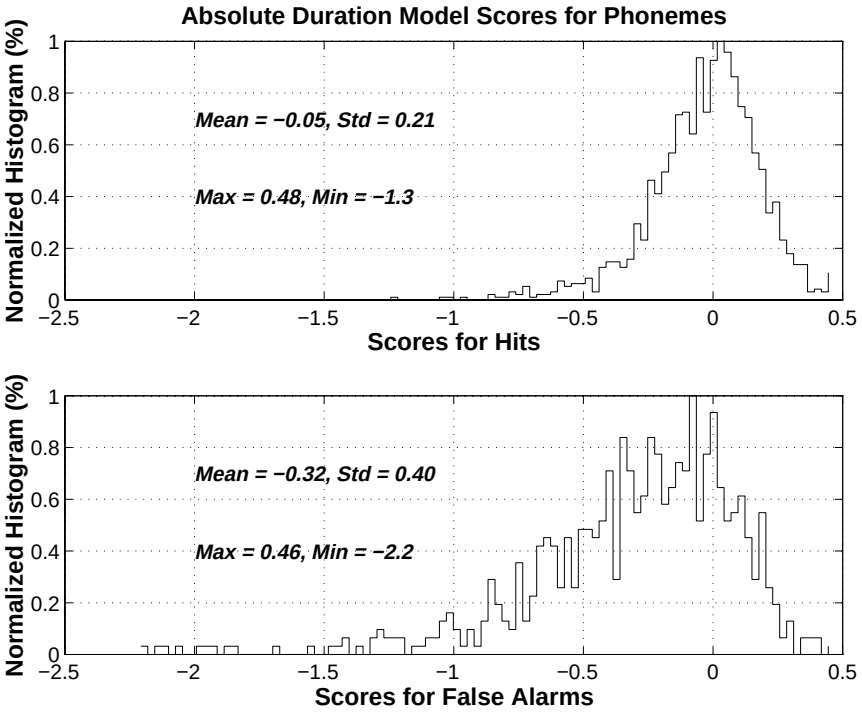

Figure 5-2: *Probability Distributions of Absolute Duration Scores for Phonemes for Hits and False Alarms.*

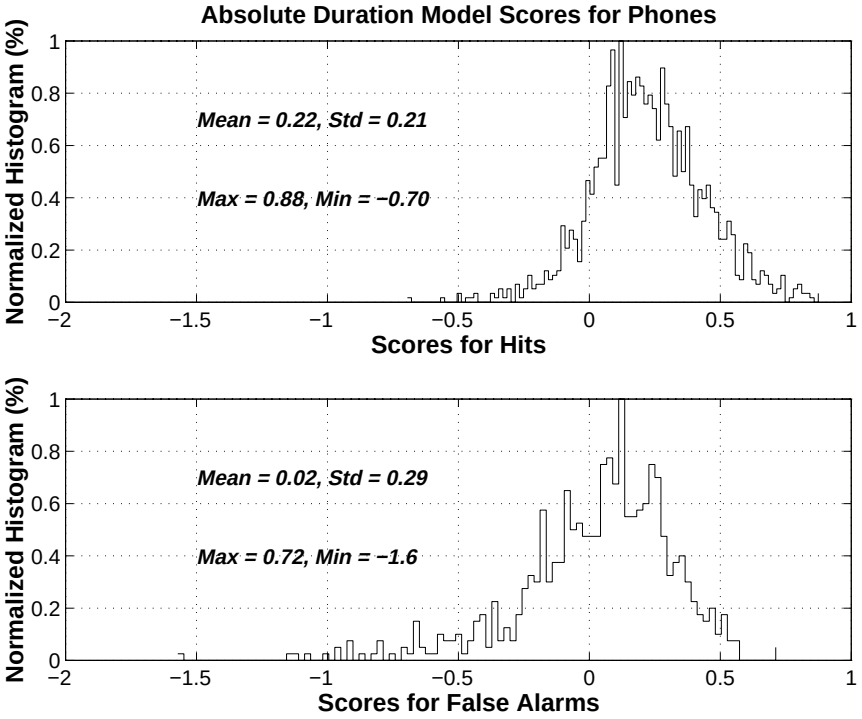

Figure 5-3: *Probability Distributions of Absolute Duration Scores for Phones for Hits and False Alarms.*

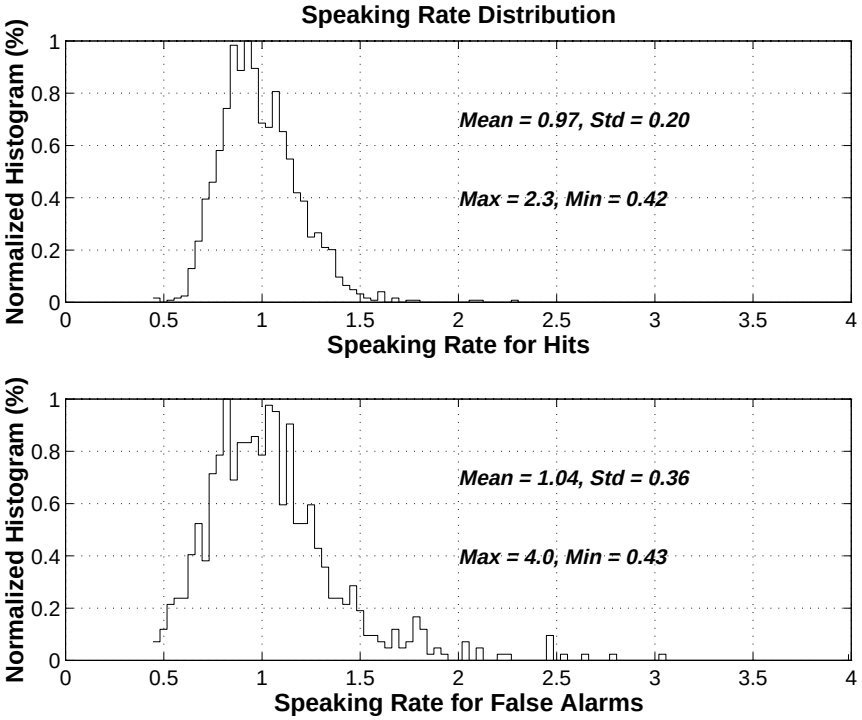

Figure 5-4: *Probability Distributions of Speaking Rates for Hits and False Alarms.*

keywords, "New York" and "Newark", are targetted. Because these words are very well-matched acoustically, their confusion is by far the largest source of error for the wordspotting system involved and a processor aimed exclusively at solving this problem can lead to significant overall gains. In fact, duration is possibly a very suitable feature candidate for disambiguating the two. Here, the duration model is designed to act as a post-processor and, at this stage, no attempt is made to integrate duration more closely into the wordspotting system.

5.2.1 Details of Experimental Procedure

The wordspotter has been assigned the task of spotting keywords from a set of utterances which contain "New York" in the ATIS domain. The most outstanding alignment error for the wordspotter is the tendency to align "New York" to waveforms corresponding with "Newark". The role of the duration post-processor is specifically to reduce the total number of errors by reducing the large number of "Newark" false alarms while minimizing the number of misses corresponding with "New York" waveforms.[3] So for all output alignments, only waveforms which contain either "New York" or "Newark" are used as input to the processor and those which are known to contain neither keyword are eliminated by hand because they are irrelevant to this task[4].

The duration processor itself consists of its own ANGIE forced alignment system along with a duration model[5]. Once an utterance with a detected "New York" is available, two new forced alignments, corresponding with the "New York" and "Newark", are performed at the given temporal boundaries where "New York" is detected. This is necessary because the wordspotter does not output a "Newark" alignment whenever it spots a "New York" and for consistency, we align both words. Each alignment is also scored for duration. Henceforth, each utterance now has duration, acoustic and linguistic scores matching both "New York" and "Newark". These scores can be combined to yield a total score. The acoustic and linguistic scores are added together without optimization, whereas the duration score is scaled by a weight before combining. The post-processor makes a decision between "New York" and "Newark" by choosing the one with the higher total score.

5.2.2 Results and Analysis

Due to disparities between the ANGIE wordspotter and the alignment system of the duration post-processor, some utterances fail to align given the supplied temporal boundaries. It is likely that this

[3] Only waveforms detected as "New York" are considered because firstly, that is the primary source of error, that is, a predominance of "New York" keywords being detected and secondly, we would like to simplify the task as much as possible, our goal being only to illustrate the power of durational information. Conceivably, in a wordspotting system, all detected "Newark" waveforms could also be passed onto the post-processor and an even greater number of errors can be reduced.

[4] In any case, these errors cannot be eliminated because the duration post-processor has only been assigned to disambiguate between "Newark" and "New York" and has no knowledge of other words.

[5] This configuration of realigning data is not ideal in reality but is used here for the purpose of this experiment whose goal is solely to demonstrate the utility of the duration model.

can be attributed to the different search strategies between the alignment and wordspotting system[6] and differing scoring thresholds upon which the system decides to fail on an utterance. Thus, failure upon alignment is a strong indicator that a mismatch between the waveform and given transcription exists. The utterances that fail to align as "New York" are designated by our system as "Newark" and those that fail to align as "Newark" are designated as "New York". Any that fail in both cases are counted as unknowns and are therefore errors which are not rectifiable.

For 323 output alignments spotted as "New York", 60 were "Newark" false alarms, which translates to a baseline error rate of 19%. During realigment, 30 utterances failed to align as "Newark", 4 failed to align as "New York" and 1 utterance failed to align in both cases. All utterances which failed to align as "Newark" are "New York" waveforms and 3 utterances which failed as "New York" are "Newark" waveforms. According to the scheme outlined above, failures in the alignment system have contributed to two errors, prior to duration scoring. Hence, 288 utterances were scored for duration of which 57 are "Newark" waveforms and 231 were "New York" waveforms.

All tokens were scored with respect to (1) relative duration model, (2) absolute duration model for phonemes and (3) absolute duration model for phones. Figures 5-5, 5-6 and 5-7 present the results for different empirical weights from the 288 waveforms which were scored for duration. For each plot, the total number of errors made is the sum of the number of misses and false alarms. And these are calculated with a duration weight that is varied from 0 to 2000.

It can be seen that the best results can be obtained from the phoneme model, for which the total number of errors is lowest and the results are relatively stable with the weight given. Also the trends for the number of misses and false alarms are quite similar. On the other hand, when the relative duration model is given a boost, the number of misses escalates while the number of false alarms remains stable. This implies that many more "New York" waveforms are given higher duration scores with their respective "Newark" alignment. The relative duration score is also more sensitive to the duration weight. Overall, the performance of the phone model does not match that of the phoneme model. This confirms as was seen in Chapter 4 that the phoneme model is more effective than the phone model.

Table 5.1 tabulates a summary of the best results obtained by using optimal duration weights. The final error rates include additional errors made when the realignment failed prior to duration scoring. Also included is the performance obtained from total scores using duration models by themselves without the augmentation of acoustic and linguistic scores. These results show that the phoneme model is superior to both the phone and relative duration model and that large duration weights can be afforded. The optimal performance yielded a 68% error reduction. The phoneme model itself, without acoustic or linguistic scores, provides a large error reduction already. It must

[6] We will briefly discuss the search strategy of the word-spotting system later in this chapter but for a detailed description, consult [27].

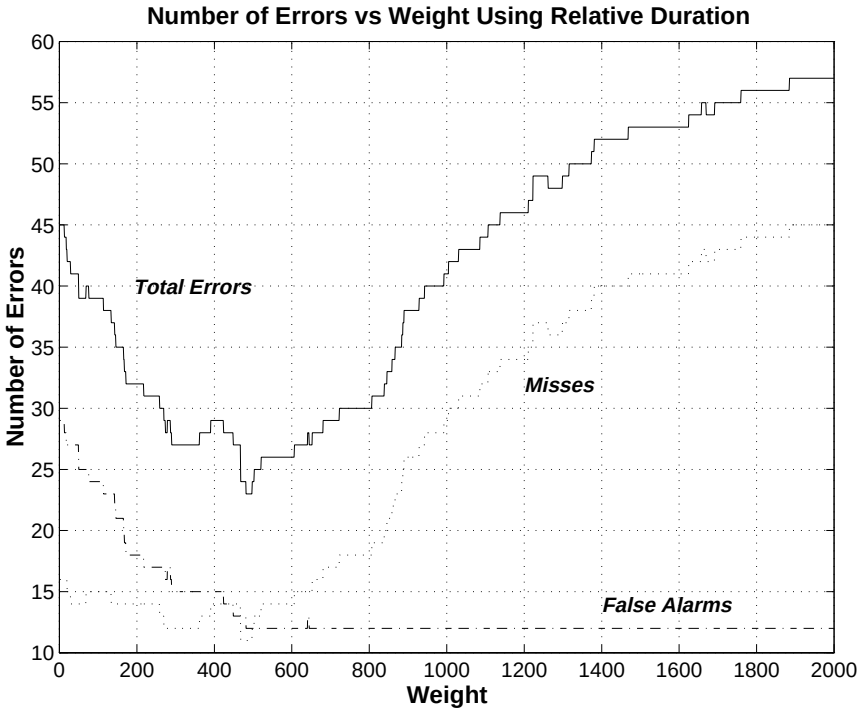

Figure 5-5: *Number of Errors versus Weight Using Relative Duration Model. The total number of errors is the sum of the number of misses and false alarms. 288 tokens are scored in total.*

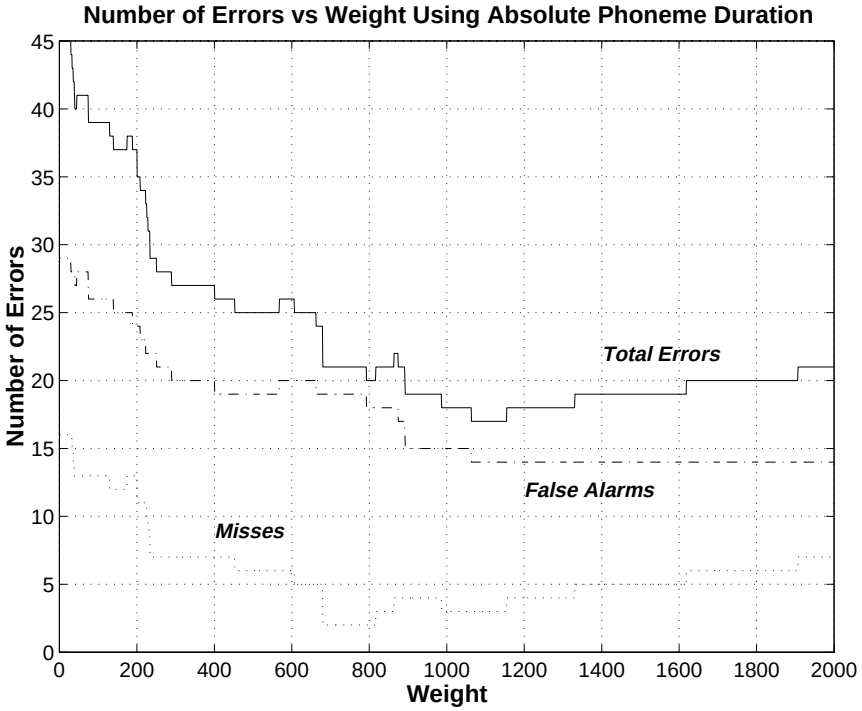

Figure 5-6: *Number of Errors versus Weight Using Absolute Duration Model for Phonemes. The total number of errors is the sum of the number of misses and false alarms. 288 tokens are scored in total.*

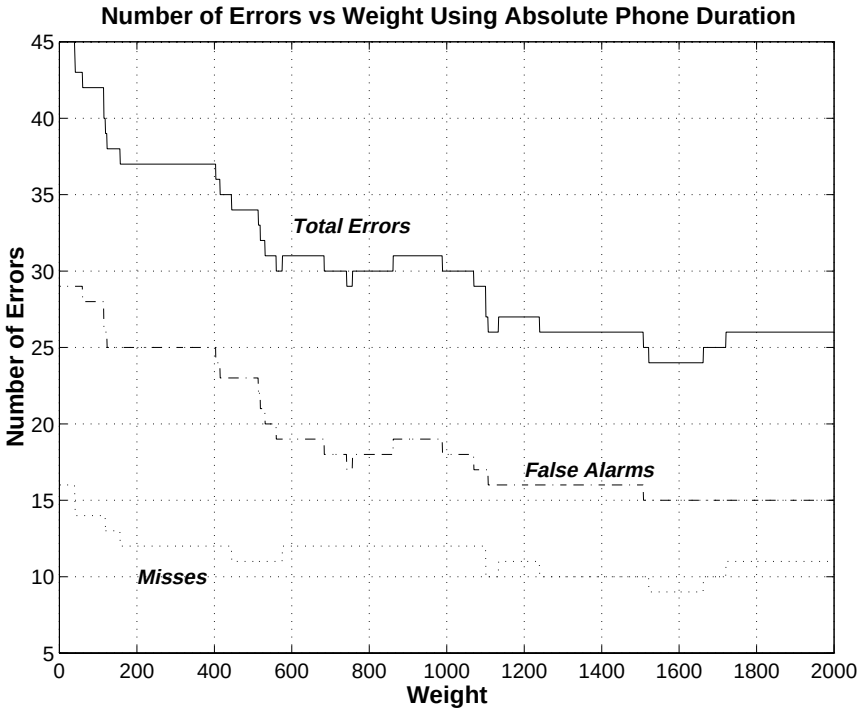

Figure 5-7: *Number of Errors versus Weight Using Absolute Duration Model for Phones. The total number of errors is the sum of the number of misses and false alarms. 288 tokens are scored in total.*

be pointed out that the realignment itself has contributed to a 32% reduction in error. Again, this disparity is most likely attributed to the difference in search stategies between the two alignment systems. The original wordspotter makes recognition errors when correct theories are inadvertantly pruned. These errors can be recovered by realigning a second time with a different strategy. Adding the duration scores from the different duration models together did not yield further performance gains. This may be explained by large redundant information in the duration models in which combining only offers an incremental gain.

In conclusion, duration modelling has generated improvement on a dramatic scale for this particular task. It has served to demonstrate that duration is an important candidate for consideration, especially for specific instances in which word pairs have confusable acoustic-phonetic features. It must be highlighted that duration weights are very large compared to those used in the phonetic recognition experiments. This indicates that, firstly, duration plays a more important role when the task involves recognizing whole words. Secondly, duration is a feature which can be applied specifically for circumstances where it is known apriori to be particularly effective over other acoustic features. Duration may not be as reliable compared with the acoustics, when other words are being considered, and so large duration weights may be counterproductive. We will evaluate the performance of our duration model in a more general task in the next section.

Table 5.1: *Results of using Duration Processor to Rescore Hypothesized "New York"s with Optimized Duration Weights. 323 waveforms were processed in total.*

Method	False Alarms	Misses	Failures	Total Errors	%	Error Reduction
Original	60	0	0	60	19%	-
Realignment	29	16	2	41	13%	32%
Relative Duration Only	11	69	2	82	25%	-37%
Relative Duration+Word Score (Weight = 490)	12	11	2	25	8%	58%
Phoneme Duration Only	15	20	2	37	11%	38%
Phoneme Duration+Word Score (Weight = 1100)	14	3	2	19	6%	68%
Phone Duration Only	12	41	2	55	17%	8%
Phone Duration+Word Score (Weight = 1600)	15	9	2	29	9%	52%

5.3 A Wordspotting System with Fully Integrated Duration Component

The duration model is fully integrated into a wordspotting system based on ANGIE. These experiments were completed in conjunction with Ray Lau and details are provided in his forthcoming PhD thesis. We will briefly outline the experiments here and present the results. For further details of the ANGIE-based wordspotting system, consult [27].

The ANGIE wordspotting system takes advantage of the hierarchical subword structure of ANGIE to model filler words in wordspotting. The ANGIE-based wordspotting system, instead of proposing only pseudo-words (as in the ANGIE phonetic recognizer), will propose both pseudo-words and keywords periodically. Each keyword has its own additive boost which is determined experimentally to optimize performance. Unlike the phonetic recognizer which employs a *best-first* search strategy, the wordspotting system uses a phone level stack decoder largely based on that proposed by Doug Paul [21]. This is essentially a *breadth-first* search with beam pruning which maintains computational tractability. At each time point determined by the acoustic-phonetic network, the number of active theories is set to a predetermined N while the rest are pruned. These theories are advanced time-synchronously. A word graph or N-best list is also incorporated. A word graph has arcs which correspond with various possible word hypotheses and nodes which correspond with possible word boundaries.

A duration score is computed whenever a keyword from the ATIS lexicon is proposed and is combined with the acoustic and linguistic scores to produce a total word score. Here, the prepausal score and geminate score are not employed because of the limitations of the word graph which does not support forward looking scores.

As in our phonetic recognition experiments, varying levels of implicit lexical knoweledge are imposed to provide progressively greater linguistic constraint and consequently leading to better baseline performances. As described in Chapter 4, the word lexicon is represented by sequences of intermediate morph units which, in turn, are represented by phonemic sequences in a predefined morph lexicon. These morph units are based on syllabic structures. Three experiments of varying linguistic constraint are performed:

1. Morph units are employed to constrain the search space, that is, only phoneme sequences that combine to form existing morphs in the lexicon are permitted. The sequences of morphs themselves are not constrained in any way so that they do not necessarily combine to form words.

2. Phoneme sequences are required to form legal morph units and additionally morph sequences forming words that are not represented in the word lexicon, are given a penalty. Under this scenario, the degree of higher order linguistic constraint is increased, and the recognizer is encouraged to propose words which are present in the lexicon. The level of constraint can be tweaked by the penalty imposed. When the penalty is very large, we approach the third experiment.

3. Phoneme sequences are required to form legal morph units and additionally, morph sequences are required to form words that appear in the ATIS word lexicon. This scenario of using full word constraints bears most resemblance to continuous word recognition. The difference is that the recognizer is not required to propose words from the ATIS lexicon but instead, only a select set of keywords.

As the degree of linguistic constraint is increased, we expect the baseline performance to improve. These experiments are evaluated using the ATIS December 93 test set with 965 utterances and 39 keywords representing the city names.

5.3.1 Results and Analysis

The standard evaluation metric in wordspotting is computed from the *Receiver Operator Characteristic (ROC)* curve. This ROC curve plots the detection rate as a percentage versus the number of false alarms normalized by the number of keywords and the hours of speech (fa/kw/hr). Generally, a *Figure of Merit* summarizes the ROC curve and is used when comparing performance among wordspotting systems. This FOM yields a mean performance over zero to ten false alarms per keyword per hour of data. A detailed description of how this FOM is computed is given in [39].

In all three experiments, we compare results of the baseline performance with no duration model with performance of the wordspotter augmented by a combined absolute phoneme and relative duration model is used. Results are tabulated in Table 5.2.

Table 5.2: *Results of Wordspotting Experiments.*

Experiment	No Duration (FOM)	Duration (FOM)	% Improvement
Using Morph Constraints	88.4	89.8	12.1
Using Morph Constraints with Word Penalty	88.6	90.0	12.3
Using Full Word Constraints	89.3	91.6	21.5

Our results indicate that duration benefits the wordspotting performance over all three scenarios. We have yielded performance improvements of 1.5%, 1.6% and 2.3% (FOM) for the three scenarios respectively. First of all, this seems to indicate that, as we speculated, duration is more useful in word recognition than in phonetic recognition because of the presence of higher level linguistic knowledge. And this is consistent with results obtained so far in this chapter. Moreover, better results are obtained as we supply greater implicit lexical knowledge in the form of morph and word constraints. This again is consistent with phonetic recognition results from Chapter 4.

In addition, a number of points should be noted from this experiment:

- The prepausal and geminate scores have been disabled due to the use of the word graph in the wordspotting system. So at present, a general model is employed for the three cases described in Chapter 4. In principle, if these separate models become enabled again, they may potentially provide greater gains. In a continuous speech recognition task, this could be reinstated where links in the word graph are introduced.

- An addition experiment is conducted to investigate the contribution of the confusable pair "New York" and "Newark" in which duration is used on these two keywords exclusively. We discovered that this alone yielded a 1% improvement, hence supporting our original claim that this pair alone is a large source of error. And as shown in the previous experiment of Section 5.2, for this specific case, duration is a valuable feature candidate for reducing error.

- Duration modelling does not offer improvement when it is applied to the word fillers. This may be explained by the fact that the word fillers have not been optimized for recognition accuracy. Consequently, many function words and one syllable words are proposed. Most of these cannot benefit from the hierarchical duration model.

5.4 Summary

This chapter has demonstrated the benefit of a duration model to a wordspotting system through a number of experiments. After some preliminary studies, we have evaluated the duration model in a task, disambiguating a pair of acoustically confusable words. The success of this particular experi-

ment shows that duration can play a major role in recognition, especially where the acoustic-phonetic attributes alone are inadequate features for discrimination. Thus, one way to utilize duration is as a post-processor for confusions which we know a priori are related to duration. The question, then, is whether duration is equally beneficial in a more general experiment where (1) duration is not targetted only for certain words and (2) the duration component is more integrated with the recognizer. When the duration model is applied to all keywords for the wordspotter, duration is unlikely to be as important a candidate for consideration compared to the acoustics. Nonetheless, the results have shown that duration offers significant gains performance; the best result was a 21.5% improvement. This indicates that our model can take advantage of complex durational information to improve overall recognition performance.

Chapter 6

Conclusion and Future Work

This thesis has implemented a complex statistical duration model which captures multiple contextual effects simultaneously throughout various levels of the phonetic hierarchy below the word level. While evidence suggests that durational information provides important perceptual cues to listeners, as yet, extensive use of durational knowledge is by and large absent in speech recognition systems. Our understanding of durational phenomena is still incomplete, and their complex nature has hindered successful quantitative modelling in the past. Most previous studies of duration effects have been driven by synthesis applications while most current recognition systems use rudimentary context-independent duration models. This study is unique because it attempts to extract complex durational knowledge for the purpose of continuous speech recognition. Our goal has been to quantitatively describe factors that are operating in concert at the morphological, syllable, phonemic and phonological levels into one comprehensive statistical model. In particular, few experiments in the past have considered multi-speaker corpora of continuous *spontaneous* speech because this is not required for speech synthesis. Our study is one of the first to be developed using spontaneous continuous speech which is greatly more suitable for recognition conditions.

Our hierarchical model is based upon the ANGIE framework which captures sublexical phenomena in the form of parse trees. Using this novel framewok, we obtain one set of models based on relative duration among sublexical components within a word and another set of speaking-rate-normalized absolute duration models based on phones and phonemes. Our models are derived from a normalization procedure where sublexical contextual effects are compensated for successively in the parse tree from the bottom upwards.

Using this paradigm, we have fulfilled our goals of (1) investigating various speech timing effects and (2) demonstrating the benefit of duration by improving recognition performance. The hierarchical structure has shown to be a useful tool for exploring various temporal effects, particularly relating to speaking rate. We have mainly been concerned with quantifying characteristics of speaking rate,

prepausal lengthening and gemination. The underlying goal has been to capture these effects for a final comprehensive model and incorporate this model into a recognition system.

The ultimate goal of this research is to eventually incorporate duration into a continuous word recognizer, and this thesis has taken first steps into demonstrating the potential gains through a series of experiments. Initially, we showed that duration can aid phonetic recognition and that the value of duration is increased as we move closer towards word recognition due to the presence of higher level linguistic knowledge. Next, duration is employed in several wordspotting experiments and again duration has proved to offer substantial improvement, with the greatest gain of 21.5% in a general wordspotting task. Thus the success of our experiments indicate that in spite of the complexity of durational effects and their interactions, they are an important factor and should be utilized to provide added constraint to a speech recognition system.

6.1 Future Work

There are multiple directions in which we can pursue further study. Our current knowledge of durational effects is incomplete, particularly in the way effects combine to obscure each other. Our model provides a good framework in which to perform more experiments investigating speech timing phenomena. We have only touched on some effects of speaking rate using the relative speaking rate parameter. More detailed experimentation can be directed towards studying the relationship between speaking rate and duration of sublexical units. This topic deserves more detailed scrutiny of the types of systematic behaviour occurring within sublexical pattern with respect to rate and can be better explored using larger amounts of data. Alternatively, speaking rate not only affects the duration relationships but is also correlated with the actual phonological realization of phonemes. For example, vowel reduction may be more likely to be associated with fast speech. The rate parameter can be utilized to discover how such phonological rules relate with speech rate. This knowledge may be useful for the recognizer as well.

Due to the constraints of time, we have not examined further speaking rate patterns at a sentence level. Our word speaking rate parameter is particularly useful for studying rate variations within a sentence or paragraph and conducting experiments on rate patterns at a discourse or semantic level. Again, this can become additional knowledge for a recognizer. For example, we have seen that words which are realized extremely slowly on a relative scale tend to be function words. We have no conclusive information about changes in speaking rate throughout a sentence but the ability to predict speaking rate is of interest both from an academic point of view and for potential application to recognition.

The work on prepausal lengthening also merits further study. Lengthening effects vary according to the length of pauses and the position within the sentence. Such effects are interrelated to the

syntactic structure of a sentence and the degree of lengthening varies at clause and phrase boundaries. It may be useful to quantify lengthening effects as a function of the distance away from the corresponding syntactic boundary. An improved model of prepausal lengthening can be developed simply by increasing the amount of data available. In our study a separate prepausal model offered small gains, mainly because of insufficient data.

Our studies have also raised more subtle issues regarding our hierarchical model. This model best captures linear or first order effects of speaking rate and assumes that duration changes uniformly as speaking varies. Evidence has suggested that this assumption largely accounts for durational changes but higher order effects are also present. This is at least one source of variability that has not been addressed by our duration model. Both the relative and absolute models assume a linearity between duration of sublexical units and rate. Further studies are required to investigate which sublexical categories are more elastic to durational changes and how this inherent nonlinearity can be modelled in a statistical framework.

This thesis has dealt with contextual effects from the phone ranging to the word level. In principle, it is possible to extend the ANGIE structure beyond the word level and incorporate syntactic and semantic categories. This would allow the modelling of paragraph and discourse level effects in the same statistical manner and also promises to be a useful tool for investigating temporal phenomena at higher levels. At this point, our knowledge of higher level duration effects remains sparse and studies in this realm will require large amounts of training data.

Finally, our novel hierarchical framework need not be confined to modelling only durational effects. One could conceivably apply a similar strategy for other prosodic parameters such as fundamental frequency and energy. These parameters, like duration, embed linguistic information and factors such as stress and position within the sentence affect them in a relative manner. As in our work on duration, the knowledge contained in such prosodic parameters can be extracted and used to provide added constraint for a speech recognition system.

Appendix A

ANGIE Categories

Layer 2	
DSUF	Derivational suffix
FCN	Function word e.g. the, from, and, a, in, you, I ,does
ISUF	Inflexional suffix
PRE	Prefix: unstressed syllable preceding the stressed root
SROOT	Stressed root: stressed syllable
UROOT	Unstressed root: unstressed syllable following the stressed root
Layer 3	
CODA	Coda after a stressed nucleus e.g. */m, n, s, st/*
DNUC	Nucleus in derivational suffix e.g. */eh, aa, ah, ow/*
FCODA	Coda in function word e.g. */t, d, m, v/*
FNUC	Nucleus in function word e.g. */iy_the, ra_from, ae, aar/*
FONSET	Onset in function word e.g. */sh!, b!, w!, l!/*
FSUF	Suffix in function word e.g. */v, d*ed, s*pl, m/*
LCODA	Coda following a long nucleus e.g. */m, n, p, r/*
LNUC+	Stressed long nucleus e.g */ey+, ow+, ay+, iy+/*
NUC	Unstressed nucleus e.g. */ae, eh, aa, ow/*
NUC+	Stressed nucleus e.g. */el+, aol+, ehr+, aor+/*
NUC_LAX+	Stressed lax nucleus e.g. */ae+, eh+, uh+, ih+/*
ONSET	Onset before a stressed nucleus e.g. */s!, f!, m!, dh!/*
UCODA	Coda after an unstressed nucleus e.g. */m, dh, n, k/*
UMEDIAL	Medial consonant between two syllable suffixes e.g. */m!, s!, t!, w!/*
UONSET	Onset before an unstressed nucleus e.g. */s!, s!p, sh!, t!/*
Inflexional suffixes	-ABLE, -AL, -ER, -EST, -ING, -LY, -PAST, -PL, -TH, -TON

Layer 4: Phoneme Set				
/aa+/	/aa/	/aw+/	/g/	/m/
/aar+/	/aar/	/ay_I/	/g!/	/m!/
/ae+/	/ae/	/uw_to/	/th/	/n/
/ah+/	/ah/	/ux_you/	/th!/	/n!/
/ao+/	/ao/	/en_and/	/f/	/ng/
/aor+/	/aor/	/ey_a/	/f!/	/w/
/ay+/	/ay/	/ix_in/	/s/	/w!/
/yu+/	/yu/	/iy_the/	/s*pl/	/r/
/uh+/	/uh/	/ra_from/	/s!/	/r/!
/uw+/	/uw/	/ah_does/	/sh/	/l/
/eh+/	/eh/	/t/	/sh!/	/l!/
/ehr+/	/ehr/	/t!/	/dh/	/h!/
/el+/	/el/	/k/	/dh!/	
/er+/	/er/	/k/!	/v/	
/ey+/	/ey/	/p/	/v!/	
/ow+/	/ow/	/p/!	/z/	
/ih+/	/ih/	/d/	/z!/	
/oy+/	oy	/d*ed/	/ch/	
/iy+/	/iy/	/d!/	/ch!/	
/aol+/	/en/	/b/	/jh/	
/ihr+/	/ing/	/b/!	/jh/	

Layer 5: Phone Set			
aa	*eh*	*g*	*n*
aar	*ehr*	*gcl*	*ng*
ae	*er*	*d*	*r*
aen	*ey*	*dcl*	*w*
ah	*iy*	*th*	*l*
ao	*ow*	*f*	*y*
aor	*uh*	*scl*	*dx*
aw	*t*	*sh*	*fr*
ax	*tcl*	*s*	*ti*
axr	*p*	*v*	*tr*
ay	*pcl*	*dh*	*ts*
ih	*k*	*z*	*hh*
ix	*kcl*	*ch*	*hl*
uw	*b*	*jh*	*hv*
ux	*bcl*	*m*	*epi*

Appendix B

Tables

Table B.1: *Hierarchical Normalization of Morphological Units: reduction in standard deviation. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.*

Sublexical Unit	Count	μ (ms)	σ_1 (ms)	σ_2 (ms)	Δ%
DSUF	5,475	200	96	67	31%
FCN	13,409	183	114	78	32%
ISUF	5,305	150	75	63	14%
PRE	3,817	157	87	63	30%
SROOT	30,058	291	113	85	25%
UROOT	4,787	156	87	64	24%

Table B.2: *Speaking Rate Normalization of Morphological Units: reduction in standard deviation. μ Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discarded. Δ% : Percentage reduction of standard deviation to mean ratio.*

Sublexical Layer	Count	μ (ms)	σ (ms)	Δ%
DSUF	5,475	202	44	55%
ISUF	5,305	146	41	43%
PRE	3,817	163	47	47%
SROOT	15,786	284	46	56%
UROOT	4,787	149	45	46%

Table B.3: *Hierarchical Normalization of Syllablic Units: reduction in standard deviation. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.*

Sublexical Unit	Count	μ (ms)	σ_1 (ms)	σ_2 (ms)	Δ%
ABLE	54	238	77	72	19%
AL	208	114	51	56	-11%
ER	25	100	57	57	0%
EST	413	217	79	67	11%
ING	901	152	64	53	16%
LY	17	172	57	57	0%
PAST	135	95	50	32	21%
PL	3,436	143	69	64	7%
TH	143	163	93	78	11%
CODA	13,865	84	57	42	29%
DNUC	5,475	111	64	44	29%
FCODA	4,505	79	58	50	13%
FNUC	13,409	84	63	46	26%
FONSET	9,011	89	64	47	27%
FSUF	2,073	81	47	36	16%
LCODA	7,897	74	55	44	19%
LNUC+	12,505	148	61	56	8%
NUC	8,982	100	61	50	20%
NUC+	6,403	169	68	56	17%
NUC_LAX+	11,150	105	49	42	14%
ONSET	25,178	115	64	50	21%
UCODA	2,344	127	79	59	25 %
UMEDIAL	123	90	48	32	45%
UONSET	7,588	82	44	35	21%

Table B.4: *Hierarchical Normalization of Phonemic Units: reduction in standard deviation for vowels. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. $\Delta\%$: Percentage reduction of variance.*

Sublexical Unit	Count	μ (ms)	σ_1 (ms)	σ_2 (ms)	$\Delta\%$
Stressed Vowels					
/*aa+*/	1341	130	45	44	2%
/*aar+*/	351	165	57	55	4%
/*ae+*/	2884	127	48	44	8%
/*ah+*/	1976	104	51	50	1%
/*ao+*/	720	140	42	42	1%
/*aol+*/	733	193	78	73	6%
/*aor+*/	1310	174	67	58	4%
/*aw+*/	793	144	47	47	0%
/*ay+*/	4844	161	53	50	5%
/*yu+*/	166	140	75	75	0%
/*uh+*/	55	71	29	29	0%
/*uw+*/	1023	146	70	68	3%
/*eh+*/	2500	100	42	41	2%
/*ehr+*/	1737	190	75	57	24%
/*el+*/	175	157	58	58	0%
/*er+*/	945	135	49	48	0%
/*ey+*/	1869	155	69	60	14%
/*ow+*/	1754	132	60	57	5%
/*ih+*/	2420	69	30	29	2%
/*ihr+*/	48	170	74	74	0%
/*oy+*/	79	168	47	45	6%
/*iy+*/	2335	131	59	54	9%
Unstressed Vowels					
/*aa*/	30	103	29	29	1%
/*aar*/	343	132	64	63	0%
/*ae*/	1059	78	55	37	33%
/*ah*/	1902	71	51	37	27%
/*ao*/	868	90	66	49	25%
/*aor*/	618	130	77	69	10%
/*ay*/	254	159	57	39	31%
/*yu*/	344	156	82	82	1%
/*uh*/	119	62	44	44	0%
/*uw*/	167	64	40	29	28%
/*eh*/	986	64	29	26	12%
/*ehr*/	171	157	72	66	9%
/*el*/	1392	113	53	53	0%
/*en*/	1500	102	55	49	11%
/*er*/	1892	107	60	57	3%
/*ey*/	827	155	79	70	11%
/*ow*/	1187	139	70	67	5%
/*ih*/	2320	67	36	34	6%
/*iy*/	4537	107	55	44	20%
/*ing*/	986	146	64	53	18%

Table B.5: *Speaking Rate Normalization of Phonemic Units: reduction in standard deviation for vowels. μ: Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discared.. $\Delta\%$: Percentage reduction of standard deviation to mean ratio.*

Sublexical Layer	Count	μ (ms)	σ (ms)	$\Delta\%$
Stressed Vowels				
/*aa+*/	1341	132	38	17%
/*aar+*/	339	156	33	32%
/*ae+*/	2884	128	37	23%
/*ah+*/	1976	108	39	26%
/*ao+*/	720	140	37	11%
/*aol+*/	311	179	45	41%
/*aor+*/	1310	177	41	40%
/*aw+*/	793	152	39	22%
/*ay+*/	4842	163	39	28%
/*yu+*/	156	119	38	37%
/*uh+*/	55	72	21	27%
/*uw+*/	1023	140	44	34%
/*eh+*/	2500	96	30	25%
/*ehr+*/	807	188	40	47%
/*el+*/	172	150	32	43%
/*er+*/	945	136	32	35%
/*ey+*/	1641	154	39	44%
/*ow+*/	1683	134	33	44%
/*ih+*/	2420	70	22	27%
/*ihr+*/	48	172	57	23%
/*oy+*/	79	171	38	22%
/*iy+*/	2324	130	36	39%
Unstressed Vowels				
/*aa*/	30	98	20	30%
/*ae*/	1059	79	28	48%
/*ah*/	1902	70	27	47%
/*ao*/	868	89	32	51%
/*aor*/	524	123	42	43%
/*ay*/	254	155	43	21%
/*yu*/	344	154	64	21%
/*uh*/	119	66	29	38%
/*uw*/	167	62	22	44%
/*eh*/	986	66	24	21%
/*ehr*/	171	164	41	45%
/*el*/	1392	112	42	21%
/*en*/	1500	100	37	32%
/*er*/	1892	103	42	27%
/*ey*/	827	148	48	36%
/*ow*/	1187	141	53	27%
/*ih*/	2320	68	25	30%
/*iy*/	4537	107	32	43%
/*ing*/	986	146	40	37%

Table B.6: *Hierarchical Normalization of Phonemic Units: reduction in standard deviation for function word specific phonemes. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.*

Sublexical Unit	Count	μ (ms)	σ_1 (ms)	σ_2 (ms)	Δ%
/*ah*/ (does)	179	71	51	41	21%
/*ay*/ (I)	451	83	59	56	5%
/*uw*/ (to)	2219	75	68	47	30%
/*ux*/ (you)	61	125	58	58	0%
/*en*/ (and)	505	104	68	56	18%
/*ey*/ (s)	488	75	67	37	44%
/*ix*/ (in)	443	60	35	30	14%
/*iy*/ (the)	1730	63	50	33	34%
/*ra*/ (from)	1971	87	62	51	17%

Table B.7: *Speaking Rate Normalization of Phonemic Units: reduction in standard deviation for function word specific phonemes. μ Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discared.. Δ% : Percentage reduction of standard deviation to mean ratio.*

Sublexical Layer	Count	μ (ms)	σ (ms)	Δ%
/*ah*/ (does)	179	72	26	49%
/*ay*/ (I)	94	83	23	51%
/*uw*/ (to)	2219	73	22	67%
/*ux*/ (you)	22	120	39	17%
/*ix*/ (in)	443	68	17	57%
/*iy*/ (the)	1730	64	17	66%
/*ra*/ (from)	1971	86	30	51%

Table B.8: *Hierarchical Normalization of Phonemic Units: reduction in standard deviation for affricates, stops and fricatives. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.*

Sublexical Unit	Count	μ (ms)	σ_1 (ms)	σ_2 (ms)	Δ%
Affricates					
/*ch*/	635	150	72	72	0%
/*jh*/	11	104	38	28	27%
/*ch!*/	224	143	54	47	14%
/*jh!*/	295	106	43	36	16%
Unvoiced Stops					
/*t*/	9218	58	41	32	21%
/*k*/	1663	81	38	33	13%
/*p*/	1117	91	34	34	2%
/*t!*/	5253	117	65	53	19%
/*k!*/	2173	115	43	41	5%
/*p!*/	1506	111	51	40	23%
Voiced Stops					
/*b*/	110	55	16	15	4%
/*d*/	1965	56	41	34	17%
/*d*ed*/	212	74	50	30	39%
/*g*/	152	67	45	38	16%
/*b!*/	1382	74	36	30	29%
/*d!*/	2427	77	47	40	40%
/*g!*/	786	73	37	31	30%
Unvoiced Fricatives					
/*th*/	370	123	79	69	12%
/*sh*/	254	122	33	33	0%
/*s*/	3935	119	64	63	1%
/*s*pl*/	3438	142	69	64	7%
/*f*/	575	95	42	42	0%
/*th!*/	304	130	73	73	0%
/*sh!*/	1600	139	47	47	0%
/*s!*/	3016	139	49	49	1%
/*f!*/	7126	120	59	59	0%
Voiced Fricatives					
/*z*/	2051	115	45	44	3%
/*v*/	2123	58	41	41	0%
/*dh*/	56	38	13	13	0%
/*z!*/	151	114	32	32	1%
/*v!*/	601	69	32	32	0%
/*dh!*/	2531	52	30	28	8%

Table B.9: *Speaking Rate Normalization of Phonemic Units: reduction in standard deviation for affricates, stops and fricatives. μ: Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discared.. $\Delta\%$: Percentage reduction of standard deviation to mean ratio.*

Sublexical Layer	Count	μ (ms)	σ (ms)	$\Delta\%$
Affricates				
/*ch*/	635	157	52	31%
/*jh*/	11	101	24	36%
/*ch!*/	224	155	40	32%
/*jh!*/	295	112	35	22%
Unvoiced Stops				
/*t*/	9218	59	25	40%
/*k*/	1663	86	30	26%
/*p*/	1117	91	25	25%
/*t!*/	5253	120	37	44%
/*k!*/	2173	118	36	18%
/*p!*/	1506	116	38	29%
Voiced Stops				
/*b*/	110	59	16	8%
/*d*/	1965	57	25	40%
/*d*ed*/	212	75	23	55%
/*g*/	152	66	29	38%
/*b!*/	1382	76	26	29%
/*d!*/	2427	74	27	40%
/*g!*/	786	76	27	30%
Unvoiced Fricatives				
/*th*/	370	117	50	33%
/*sh*/	254	121	35	-6%
/*s*/	3935	119	55	14%
/*s*pl*/	3438	139	40	40%
/*f*/	575	96	31	27%
/*th!*/	304	132	60	18%
/*sh!*/	1600	154	47	9%
/*s!*/	3016	143	37	27%
/*f!*/	7126	121	46	23%
Voiced Fricatives				
/*z*/	2051	119	37	21%
/*v*/	2123	53	24	35%
/*dh*/	56	37	11	11%
/*z!*/	151	112	27	15%
/*v!*/	601	61	26	11%
/*dh!*/	2531	51	14	52%

Table B.10: *Hierarchical Normalization of Phonemic Units: reduction in standard deviation for nasals, semivowels and aspirants. μ: Mean duration. σ_1: Unnormalized standard deviation. σ_2: Normalized standard deviation. Δ%: Percentage reduction of variance.*

Sublexical Unit	Count	μ (ms)	σ_1 (ms)	σ_2 (ms)	Δ%
Nasals					
/*m*/	3287	76	44	36	18%
/*m*!/	2824	67	28	27	1%
/*n*/	7987	55	43	39	9%
/*n*!/	2063	65	36	36	0%
/*ng*/	145	74	43	43	1%
Semivowels and Aspirants					
/*w*/	304	48	22	22	0%
/*r*/	1281	37	19	20	-5%
/*l*/	3962	69	38	39	-2%
/*w*!/	2751	71	46	46	0%
/*y*!/	300	39	16	14	13%
/*r*!/	960	51	31	30	4%
/*l*!/	3099	68	37	37	0%
/*h*!/	600	68	37	36	2%

Table B.11: *Speaking Rate Normalization of Phonemic Units: reduction in standard deviation for nasals, semivowels and aspirants. μ Normalized mean duration. σ: Normalized standard deviation with deterministic tokens discared.. Δ% : Percentage reduction of standard deviation to mean ratio.*

Sublexical Layer	Count	μ (ms)	σ (ms)	Δ%
Nasals				
/*m*/	3287	76	27	39%
/*m*!/	2824	67	21	23%
/*n*/	7987	54	27	34%
/*n*!/	2063	61	22	34%
/*ng*/	145	79	41	11%
Semivowels and Aspirants				
/*w*/	304	50	22	4%
/*r*/	1281	37	19	1%
/*l*/	3962	71	38	1%
/*w*!/	2751	67	31	30%
/*y*!/	300	40	10	37%
/*r*!/	960	50	22	26%
/*l*!/	3099	67	26	28%
/*h*!/	600	71	30	23%

Bibliography

[1] Anastasakos, A., Schwartz, R., Shu, H., "Duration Modeling in Large Vocabulary Speech Recognition," *Proceedings ICASSP '95*, May 1995, pp. 628-631.

[2] Campbell, W. N., Isard, S. D., "Segment Durations in a Syllable Frame," *Journal of Phonetics*, Vol. 19, 1991, pp. 37-47.

[3] Campbell, W. N., "Syllable-based Segmental Duration," in G. Bailly, C. Benoit and T. R. Sawallis, eds., *Talking Machines: Theories, Models, and Designs*, (Elsevier Science Publishers B. V., 1992), pp. 211-224.

[4] Campbell, W. N., "Predicting Segmental Durations for Accomodation within a Syllable-Level Timing Framework," *Proceedings EUROSPEECH '93*, Berlin, Germany, September 1993, pp. 1081-1084.

[5] Crystal, T. H., House, A. S., "Segmental Durations in Connected Speech Signals: Preliminary Results," *J. Acoust. Soc. Am.*, Vol. 72, No. 3, September 1982, pp. 705-716.

[6] Crystal, T. H., House, A. S., "Segmental Durations in Connected Speech Signals: Current Results," *J. Acoust. Soc. Am.*, Vol. 83, No. 4, April 1988, pp. 1553-1573.

[7] Crystal, T. H., House, A. S., "Segmental Durations in Connected Speech Signals: Syllabic Stress," *J. Acoust. Soc. Am.*, Vol. 83, No. 4, April 1988, pp. 1574-1585.

[8] Grover, C., Terken, J., "Rhythmic Constraints in Durational Control," *Proceedings ICSLP '94*, Yokohama, Japan, September 1994, pp. 363-366.

[9] Harris, M. S., Umeda, N., "Effect of Speaking Mode on Temporal Factors in Speech," *J. Acoust. Soc. Am.*, Vol. 56, No. 3, September 1974, pp. 1016-1018.

[10] House, A. S., "On Vowel Duration in English," *J. Acoust. Soc. Am.*, Vol. 33, No. 9, September 1961, pp 1174-1178.

[11] Jones, M., Woodland, P.C., "Using Relative Duration in Large Vocabulary Speech Recognition," *Proceedings EUROSPEECH '93*, Berlin, Germany, September 1993, pp. 311-314.

[12] Klatt, D. H., "Interaction between Two Factors that Influence Vowel Duration," *J. Acoust. Soc. Am.*, Vol. 54, No. 4, 1973, pp. 1102-1104.

[13] Klatt, D. H., "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence," *J. Acoust. Soc. Am.*, Vol. 59, No. 5, May 1976, pp. 1208-1221.

[14] Lea, W. A., "Prosodic Aids to Speech Recognition," in Lea, W. A., ed., *Trends in Speech Recongition*, (Englewood Cliffs: Prentice-Hall, 1980), pp. 166-205.

[15] Levinson, S. E., "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition," *Computer Speech and Language*, Vol. 1, 1986, pp. 71-78.

[16] Luce, P. A., Charles-Luce, J., "Contextural Effects on Vowel Duration, Closure Duration, and the Consonant/Vowel Ration in Speech Production," *J. Acoust. Soc. Am.*, Vol. 78, No. 6, December 1985, pp. 1949-1957.

[17] Marshall, C. W., Nye, P. W., "Stress and Vowel Duration Effects on Syllable Recognition," *J. Acoust. Soc. Am.*, Vol. 74, No. 2, 1983, pp. 433-443.

[18] Moore, K., Zue, V. W., "The Effect of Speech Rate on the Application of Low-Level Phonological Rules in American English," paper presented at the 109th Meeting of the Acoustical Society of America, April 10, 1985.

[19] Osaka, Y., Makino, S., Sone, T., "Spoken Word Recognition Using Phoneme Duration Information Estimated from Speaking Rate of Input Speech," *Proceedings ICSLP '94*, Yokohama, Japan, September 1994, pp. 191-194.

[20] O'Shaughnessy, D., "A Multispeaker Analysis of Durations in Read French Paragraphs," *J. Acoust. Soc. Am.*, Vol. 76, No. 6, December 1984, pp. 1664-1672.

[21] Paul, D. B., "Efficient A^* Stack Decoder Algorithm for Continuous Speech Recognition with a Stochastic Language Model," *MIT Lincoln Laboratory*, TR 930, July 1991.

[22] Peterson, G. E., Lehiste, I., "Duration of Syllable Nuclei in English," *J. Acoust. Soc. Am.*, Vol. 32, No. 6, June 1960, pp. 693-703.

[23] Pitrelli, J. F., "Hierarchical Modeling of Phoneme Duration: Application to Speech Recognition," Ph. D. Thesis, M. I. T., May, 1990.

[24] Pols, L. C. W., Wang, X., ten Bosch, L. F. M., "Modelling of Phoneme Duration (using the TIMIT database) and its Potential Benefit for ASR," *Speech Communication*, Vol. 19, 1996, pp. 161-176.

[25] Port, R. F., "Linguistic Timing Factors in Combination," *J. Acoust. Soc. Am.*, Vol. 69, No. 1, January 1981, pp. 262-274.

[26] Port, R. F., Reilly, W. T., Maki, D. P., "Use of Syllable-Scale Timing to Discriminate Words," *J. Acoust. Soc. Am.*, Vol. 83, No. 1, January 1988, pp. 265-273.

[27] Lau, Ray, "Subword Lexical Modelling for Speech Recognition," *Ph.D. Thesis*, To be completed 1998.

[28] Riley, M. D., "Tree-based Modeling of Segmental Durations," in G. Bailly, C. Benoit and T. R. Sawallis, eds., *Talking Machines: Theories, Models, and Designs*, (Elsevier Science Publishers B. V., 1992), pp. 265-273.

[29] Riley, M. D., "Statistical Tree-based Modeling of Phonetic Segmental Durations," *J. Acoust. Soc. Am.*, Vol. 85, S44, 1989.

[30] Seneff, S., Lau, R., Meng, H., "ANGIE: A New Framework for Speech Analysis Based on Morpho-phonological Modeling," *Proc. ICSLP '96*, Philadelphia, PA, 1996, pp. 225-228.

[31] Umeda, N., "Vowel Duration in American English," *J. Acoust. Soc. Am.*, Vol. 58, No. 2, August 1975, pp. 434-445.

[32] Umeda, N., "Consonant Duration in American English," *J. Acoust. Soc. Am.*, Vol. 61, No. 3, March 1977, pp. 846-858.

[33] Van Santen, J. P. H., "Contextual Effects on Vowel Duration," *Speech Communication*, Vol. 11, February 1992, pp. 513-546.

[34] Van Santen, J. P. H., "Deriving Text-to-Speech Durations from Natural Speech," *Talking Machines: Theories, Models, and Designs*, (Elsevier Science Publishers B. V., 1992), pp. 275-285.

[35] Wang, X, Pols, L. C. W., ten Bosch, L. F. M., "Analysis of Context-Dependent Segmental Duration for Automatic Speech Recognition," *Proceedings ICSLP '94*, Philadelphia, PA USA, October 3-6, 1996.

[36] Wang, X, ten Bosch, L. F. M., Pols, L. C. W., "Integration of Context-Dependent Durational Knowledge into HMM-Based Speech Recognition," *Proceedings ICSLP '94*, Philadelphia, PA USA, October 3-6, 1996.

[37] Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P., "Segmental Duration in the Vicinity of Prosodic Phrase Boundaries," *J. Acoust. Soc. Am.*, Vol. 91, 1992, pp. 1707-1717.

[38] Zue, V., Seneff, S., Polifroni, J., Phillips, M., Pao, C., Goddeau, D., Glass, J., Brill, E., "The MIT ATIS System: December 1993 Progress Report," *Proc. ARPA Spoken Language Technology Workshop*, Plainsboro, NJ, 1994, pp. 67-71.

[39] The Road Rally Word-Spotting Corpora (RDRALLY1), "The Road Rally Word-Spotting Corpora (RDRALLY1)" *NIST Speech Disc 6-1.1*, September, 1991.