ELSEVIER

Speech Communication 27 (1999) 113–134

SPEECH COMMUNICATION

A hierarchical duration model for speech recognition based on the ANGIE framework [1]

Grace Y. Chung *, Stephanie Seneff

Spoken Language Systems Group, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 8 May 1998; received in revised form 3 November 1998; accepted 6 November 1998

Abstract

This paper presents a hierarchical duration model applied to enhance speech recognition. The model is based on the novel ANGIE framework which is a flexible unified sublexical representation designed for speech applications. This duration model captures duration phenomena operating at the phonological, phonemic, syllabic and morphological levels. At the core of the modelling scheme is a hierarchical normalization procedure performed on the ANGIE parse structure. From this, we derive a robust measure for the rate of speech. The model uses two sets of statistical models – a first set based on relative duration between sublexical units and a second set based on absolute duration that has been normalized with respect to the speaking rate. We have used this paradigm to explore some speech timing phenomena such as the secondary effects on relative duration due to variations in speaking rate, the characteristics of anomalously slow words, and prepausal lengthening effects. Finally, we successfully demonstrate the utility of durational information for recognition applications. In phonetic recognition, we achieve a relative improvement of up to 7.7% by incorporating our model over and above a standard phone duration model, and similarly, in a word spotting task, an improvement from 89.3 to 91.6 (FOM) has resulted. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Duration modelling; Prosodic modelling; Speech recognition

1. Introduction

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.

Given that such durational information is of perceptual importance to the human listener, it follows that durational information may be extracted for improving speech recognition performance. It has also been observed that recognition error rates are higher for particularly fast speakers (Pallet et al., 1995) and consequently the ability to handle such variations could boost recognition performance. However, our current understanding of durational patterns and the many sources of variability which affect them, is still sparse. To

* Corresponding author. Tel.: +1 617 253 3043; fax: +1 617 258 8642; e-mail: graceyc@mit.edu

[1] This research was supported by a contract from Bell Atlantic and by the National Science Foundation under Grant No. IRI-9618731.

0167-6393/99/$ – see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(98)00071-5

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 phenomena reside at multiple levels of the phonological hierarchy. Duration values are associated intrinsically with individual phones but are also altered by factors that include phonetic contexts within the syllable, position within the word, lexical stress, syntactic structure and those that operate at higher levels such as rhythm, speaking rate and semantic novelty. Such factors have been studied previously for many years (House, 1961; Klatt, 1976; Umeda, 1975; Umeda, 1977). Because of this abundance of factors that coexist at multiple levels, their interactions obscure one another's manifestations, rendering their presence difficult to detect or analyze.

For these reasons, most speech recognition systems tend to under-utilize the knowledge provided by durational cues. Much extensive research in speech timing has been directed towards synthesis applications (Campbell, 1992; Van Santen, 1992), but duration models developed solely for synthesis do not address the issue of variabilities among speakers. This is because in synthesis, only one rendition of a given text is necessary, and so researchers have been concerned with predicting one duration value given a set of contextual conditions. As a consequence, these models are not suitable for use for recognition purposes.

In this paper, we propose a computational statistical duration model which captures speech timing phenomena at various linguistic levels. The framework is developed on a hierarchical sublexical representation known as ANGIE, which provides a basis for a variety of tasks, namely, letter-to-sound/sound-to-letter generation (Seneff et al., 1996), phonetic recognition, and word spotting (Lau and Seneff, 1997). In this work, we use ANGIE to model contextual effects in duration from the phone level up to the word level. We demonstrate its capability as a mechanism for investigating speech timing phenomena by examining secondary effects of speaking rate changes and prepausal lengthening effects. The duration model is incorporated into a speech recognizer, and our experiments show that the utilization of durational effects offers substantial performance gains in both phonetic recognition and word spotting, thereby demonstrating the value of duration modelling as an aid to speech recognition.

This paper begins with an introduction of the hierarchical representation, ANGIE, followed by an explanation of the principles of our duration model, which accounts for effects pertaining to both relative and absolute duration, partly through the use of a rate of speech measure. We then demonstrate the value of this new duration model by showing the reduction in variance of data as a result of the normalization procedure associated with this model. We then present some experiments that apply this model to investigate various speech timing phenomena such as secondary effects of speaking rate variations, characteristics of anomalously slow words, and the characteristics of prepausal lengthening. We will also present results of the performance gains achieved when this duration model is applied to aid (1) phonetic recognition and (2) word spotting. More background and details of our duration model and experimental results are documented in (Chung, 1997; Lau, 1998).

2. The ANGIE framework

ANGIE is a trainable probabilistic paradigm in which word substructures are jointly characterized by a hand-written context-free grammar and are represented in a multi-layered hierarchical structure. This framework has been adopted in multiple tasks such as speech recognition and letter-to-sound/sound-to-letter generation. Details of the probability model and grammar are described in (Lau and Seneff, 1997; Seneff et al., 1996).

A typical ANGIE parse structure is shown in Fig. 1. There are four layers below that of the WORD node. The upper layers, from top to bottom, capture morphology and syllabification, while the preterminal layer represents phonemics, and the bottom terminal categories are phones. The mappings from the preterminal phoneme to

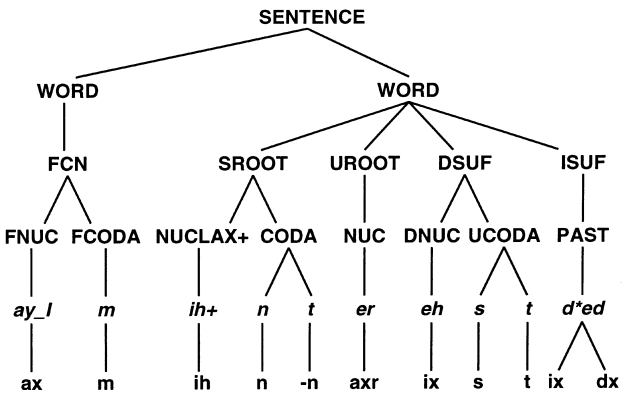

Fig. 1. Sample ANGIE parse tree for the phrase "I'm interested ..." Below the WORD node, the layers, from the top, represent morphology, syllabification, phonemics and phonology. Mappings from the preterminal phoneme to terminal phone layers capture phonological rules.

the terminal phone layers capture rules that govern phonological events. Lexical stress information is explicitly conveyed throughout the morphological, syllabic and phonemic layers by the use of distinct nodes to indicate the presence of stress. For example, SROOT denotes a stressed root, and /*ih* + / [2] denotes a stressed version of /*ih*/. Consonants are explicitly marked for occurring in the onset position. In the morphological and syllabic layers, nodes which appear in the context of a set of predefined function words are labelled accordingly. In addition, we have also expanded our set in the phoneme layer for modelling purposes to include (1) phonemes that are unique to certain morphemes (e.g., /*d*ed*/ for the past tense morpheme "ed" and /*s*pl*/ for the plural morpheme), (2) phonemes which occur in specific function words (e.g., /*uw_to*/ for the /*uw*/ in "to" and /*ay_I*/ for the /*ay*/ in "I") and also some word-specific diphones (e.g., /*ra*/ in "from"). There is a total of approximately 100 such preterminal pseudo-phonemes. In the phone set, geminations and deletions are explicitly tagged for a left context by a "−". For example, as shown in Fig. 1, the [−n] indicates that the parent phoneme /*t*/ was deleted during a realization following an /*n*/.

We will see that this representation is a particularly powerful paradigm for constructing computational models for duration because it has the ability to capture durational effects associated with the aforementioned layers of the phonological hierarchy as well as some important linguistic elements with direct relevance to durational patterns such as lexical stress.

3. The duration model

Durational relationships of sublexical units encode linguistic information, residing at various levels of the phonological hierarchy. 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, consider the two-syllable words "strongest", "longest" and "oldest". All of these 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 syllable. There are several reasons for this. Firstly, the initial syllables of "strongest" and "longest" begin with one or more consonants at the onset whereas a consonant onset is entirely absent in "oldest." Moreover, we expect the first syllable of "strongest" to have an even greater duration than that of "longest" because of the presence of the consonant cluster. If we correct for the lengthening effect of the consonant cluster in "strongest", by shortening it with some scaling factor, we can compare "strongest" and "longest" with the same probability model, effectively compensating for variations in sublexical durations below that of the syllable level. And in order to compare the duration of these two words against "oldest", it is necessary to lengthen the latter with another scaling factor to correct for the shorter

[2] Throughout this paper, we use a modified ARPAbet nomenclature where a single pair of enclosing []'s indicates phonetic units and a pair of enclosing //'s indicates pseudo-phonemic units.

duration induced by the absence of a consonant onset altogether. In addition, we also need to account for the different intrinsic durations of the vowel phonemes through further multiplicative factors, that is, more normalization at the phoneme level. Henceforth, we are able to model all three words by the 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 phonological variations. For example, a /t/ might be realized either as a flap [dx] or a closure followed by a release [tcl,t] which tends to be considerably longer in duration than the former. Both of these effects can be corrected for by different scaling factors, so that all instances of the phoneme /t/ can be compared against one another in the same distribution. This normalization process can be used successively throughout the phonological hierarchy, and forms the basic principle behind our duration model. Incidentally, when all morphological, syllabic and 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 realizations below the level of the word. And it can be argued that the remaining major sources of variabilities operate at and above the word level, of which the rate of speech is a major contributing factor. Therefore, we can define this normalized word duration as the inverse of a relative speaking rate for which a word is spoken. We will elucidate upon this in Section 3.4.

Our hierarchical representation forms the basis for two sets of statistical models – one based on relative duration and another based on speaking-rate-normalized absolute duration. The model involves a two-pass procedure where, initially, statistics for all nodes are gathered to perform a successive hierarchical normalization. After normalization, we are now ready to collect statistics based on relative duration to formulate the relative duration model. The hierarchical normalization procedure also yields a measure for rate of speech which, in turn, is used to build speaking-rate-normalized absolute duration models. Our overall scheme is illustrated in Fig. 2. We will detail the procedures in the following sections.

3.1. Hierarchical normalization

At the heart of our duration model is a hierarchical normalization procedure which is performed on each node or sublexical unit of the completed ANGIE parse tree. Under this scheme, model variance is reduced significantly at each node and sparse data problems are ameliorated. Our stategy involves a simple scaling of node durations, based on their respective realizations as represented by their child nodes, and is propagated from the bottom nodes upwards to the top node of the parse tree.

Given a nonterminal node, its normalized duration is equivalent to the sum duration (also normalized) of its child nodes in the immediate

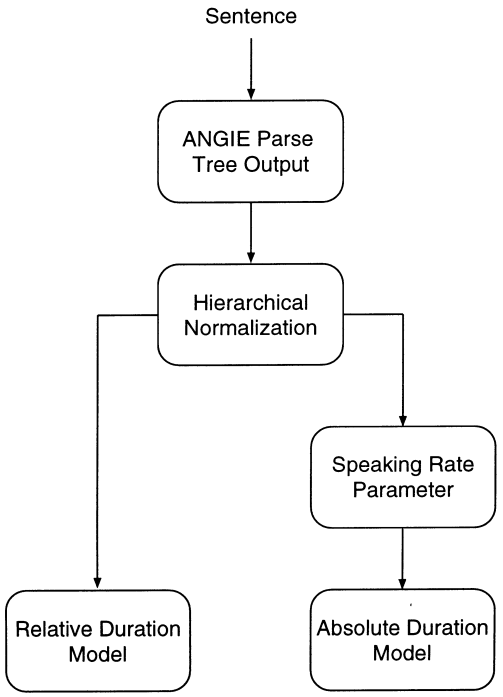

Fig. 2. The Hierarchical Duration Modelling Scheme: an ANGIE parse tree undergoes a successive hierarchical normalization described in Section 3.1. After this is completed (1) the relative duration is computed and (2) a speaking rate parameter is computed which in turn is used to normalize absolute duration for a second set of models.

layer below, scaled by some ratio predetermined during training. This ratio is the mean duration of the parent node, over the entire training data, divided by the mean duration when the parent node has been realized exclusively by the corresponding child nodes.

For example, in Fig. 3, the pseudo-phoneme $/d^*ed/$ is phonetically realized by a front schwa [ix] followed by a flap [dx], as in "... interes*ted* in ..." This $/d^*ed/$ can be realized by any number of sequences of phones as governed by our context-free rules. Then, the normalized duration for the ith instance, $\mathrm{DUR}_i(d^*ed)$, as illustrated by Eq. (1) below, is equivalent to the sum of the [ix] and [dx] durations, $\mathrm{DUR}_i(ix)$ and $\mathrm{DUR}_i(dx)$, measured from their time alignment, and scaled thereafter by a ratio. This ratio is given by the overall mean duration of $/d^*ed/$, $\mu_{\mathrm{DUR}}(d^*ed)$, taken over all instances of $/d^*ed/$ in the training set, divided by the mean duration of $/d^*ed/$ that is conditioned exclusively upon instances where it is realized as an [ix] followed by a [dx], represented by $\mu_{\mathrm{DUR}}(d^*ed \mid ix, dx)$.

$$\mathrm{DUR}_i(d^*ed) \triangleq (\mathrm{DUR}_i(ix) + \mathrm{DUR}_i(dx)) \times \frac{\mu_{\mathrm{DUR}}(d^*ed)}{\mu_{\mathrm{DUR}}(d^*ed \mid ix, dx)}. \quad (1)$$

So now the normalized duration of $/d^*ed/$ represents a duration which has compensated for the durational characteristics uniquely associated with this particular phonological realization. Similarly, all node durations in the phoneme layer are adjusted accordingly with respect to each of their child node sequences. After this is completed, we continue upwards to adjust node durations in the syllabic layer with respect to the phonemic realizations of each node. In this manner, the normalization scheme is propagated upwards successively throughout each layer of the parse tree. Node durations in each layer are scaled by a ratio determined by the identity of nodes in the layer immediately below.

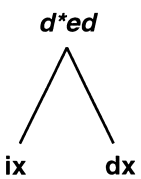

Fig. 3. A subtree shows the pseudo-phoneme $/d^*ed/$, where the "$^*ed$" marks past tense context. It is realized phonetically by the [ix dx] sequence.

According to this strategy, individual probability distributions [3], corresponding to different realizations of the one parent node, are all effectively scaled to have the same global mean, that is, each of these distributions are collapsed together. By merging these distributions, we can construct models which account for various contextual information without the need to split training data, thereby enabling us to mitigate sparse data problems.

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, as mentioned earlier in Section 3, the phoneme $/t/$ has several allophonic realizations. A specific instance may be realized as a flap [dx] which would have an inherently shorter duration. A particular instance may have an unusually long duration compared with the average duration of a [dx], and yet it may be considered relatively short among the total pool of $/t/$ tokens which have many other phonological realizations. In fact, when the $/t/$ phoneme has a comparatively long duration, it tends not to be flapped. In multiplying by a scaling factor and shifting the distributions as specified by our normalization scheme, we effectively correct for these variations and 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 by this normalization procedure.

3.2. The relative duration model

Upon completion of normalization, we build a set of models that is based on relative duration. At

[3] In our case, we have chosen to use Gaussian models.

each node in the parse tree, its duration is computed as a percentage of the total duration of its corresponding parent node. During training, we collect and store Gaussian statistics for all two-level subtrees where there is more than one child node. When scoring, probabilities are computed for the relative duration of all children nodes, and these are all subsequently combined and averaged to yield a score for the particular subtree. A duration score for the corresponding word is computed by averaging all scores computed from each available two-level subtree.

3.3. 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 phonological levels simultaneously, and, in so doing, account for various contextual factors that occur and interact at these hierarchical levels. Researchers in the past (Pitrelli, 1990) have been faced with the problem of choosing a single appropriate linguistic representation to best capture durational phenomena because no one linguistic unit can completely depict durational information which operates at multiple levels. Here, by using the ANGIE structure, we have overcome that problem. Also previous attempts to statistically 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 or rule out any particular effects or interactions which are significant, or determine which factors actually manifest in the model. The tendency to do this has constituted shortcomings in previous approaches (Klatt, 1976). Our data-driven model has the flexibility to learn statistically any durational effects from training data within the context of our framework.

The philosophy behind our relative duration models is founded upon the hypothesis that proportionate relationships among sublexical units are more important than raw absolute duration values. In light of the inherent variabilities, especially with respect to the rate of speech among realizations corresponding to different speakers, as well as each different realization corresponding to the same speaker, we expect relative durations to be better preserved in comparison with unnormalized absolute durations, which can range significantly. On the other hand, models based on raw durations will repeatedly penalize unusually slow and fast speakers.

3.4. Speaking rate parameter

In recognition, variations in speaking rate are particularly difficult to deal with, 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. As yet, there is no consensus on a reliable measure for speaking rate nor is there a generally accepted ideal linguistic unit at which rate of speech is best interpreted. Moreover, our knowledge of how rate affects segmental duration is sparse. Particularly in recognition, it is necessary for a speaking rate measure to be text-independent and computable in real time.

In the following, we define a parameter that hypothesizes a speaking rate value at the end of each word. After normalization is propagated up to the top WORD node, its duration is expected to be independent of inherent durations of its sublexical realization represented by its descendents in the parse tree. This means that we have accounted for the various sublexical contextual factors, including its phonemic realizations, that influence the duration of the various segments for this WORD node, and we now attribute the one remaining source of variability to a rate of speech at which the speaker realized the word. Note that we have not modelled explicitly syntactic or semantic factors that influence duration, and so these effects, resident at or above the word level, are responsible for some of the factors that act in concert to influence this word-level rate of speech.

Thus, the value of the normalized WORD duration provides an indication for this rate of speech. A word-level inverse speaking rate parameter is defined mathematically as the ratio of the normalized WORD duration over the global average normalized WORD duration which is collected from the training corpus:

$$\text{Inverse Speaking Rate} \triangleq \frac{\text{DUR}(\text{WORD})}{\mu_{\text{DUR}}(\text{WORD})}. \quad (2)$$

Effectively, this is an *inverse* measure of *relative* speaking rate whereby a large value corresponds with a slower than average rate of speech and a small value corresponds with a faster than average rate of speech. [4]

This parameter has the advantage that it is computed at a word level, which enables it to capture speaking rate variations within a sentence and also renders it compatible for use at the end of each word hypothesis during recognition time. In addition, it is a particularly reliable reflection of the actual rate of speech, due to its independence from all inherent durations of component sublexical units, attributable to the nature of the normalization scheme. As we shall demonstrate, this parameter also provides an amenable paradigm for investigating speech timing effects.

3.5. Absolute duration model

While the relative duration model exploits the proportionate distribution of total duration among sublexical units of the parse tree, there is also information encoded within the absolute durations of nodes. However, absolute durations cannot be modelled successfully without some normalization to account for the vast variabilities that we have already discussed. Our strategy involves constructing speaking-rate-normalized absolute duration models at the phonemic level by dividing phoneme durations by the speaking rate parameter associated with the given word:

$$\text{NDUR}_i(\text{NODE}) \triangleq \frac{\text{DUR}_i(\text{NODE})}{\text{Speaking Rate}_i}. \quad (3)$$

In the above equation, DUR_i denotes unnormalized absolute duration of an instance i of a certain node. This is simply the sum duration of its child nodes. NDUR_i denotes the normalized absolute duration of this node which is computed by dividing with the speaking rate parameter. In effect, a speaking-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 the word. Our goal is to compensate for the effects of speaking rate on node duration, and subsequently construct absolute duration models that are speaking-rate-normalized. This formula acts on the assumption that speaking rate affects segmental duration linearly and uniformly on each sublexical unit. It does not account for any evidence that some nodes are more inelastic to speaking rate changes than others. This is in the same vein as our relative duration models, which underlyingly assume that durations throughout the hierarchical structure, on a relative scale, are, by and large, preserved for different realizations. The absolute duration models embody essentially the same philosophy and make an even stronger assumption that all phoneme durations, within the same word, are perturbed linearly by the one speaking rate parameter. We will observe later that these assumptions hold well, based on the encouraging results from recognition experiments, but also we have evidence to suggest that speaking rate, to some extent, does not act linearly or uniformly on all segments, as will be seen in Section 6.

The subsequent statistical models use the two-parameter Gamma probability distribution functions. These functions have been used previously to model absolute duration (Crystal and House, 1988) and are considered to have characteristics well-matched to those of absolute duration distributions.

[4] This is inverse to the conventional notion of speaking rate which is measured by the average number of occurrences of some linguistic unit in unit time. And so a large numerical value usually corresponds with fast speech. However, our measure produces a large numerical value for words with longer than average duration and correspondingly a slow rate of speech. Effectively, it is an inverse parameter. Henceforth, we will simply refer to this as the speaking rate parameter.

4. Corpus and training procedure

Our experiments are conducted on spontaneous speech from the *Air Travel Information Service* (ATIS) corpus (Dahl et al., 1994; Zue et al., 1994). 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 speaking rate experiments. This subset consists of 88 speakers and 43,467 words in total. There are over 220,000 phones. We have chosen the ATIS December'93 test set of 965 utterances and 27 speakers as test data in our recognition experiments. Our relative duration models contain 654 distinct two-level subtree patterns while our absolute duration models comprise 94 phonemes. The phonetic inventory consists of 64 unique categories. Phonetic alignments and their temporal boundaries are obtained by an automatic procedure given the orthographic transcription and using an ANGIE-based forced alignment system. This is done for the training data for use in subsequent speech timing experiements, and for the test data as reference for recognition experiments.

In the following, we will give a summary of the implementation procedures required for training and scoring.

1. Computing parse trees: ANGIE parse trees for all words are constructed from forced alignments that correspond with utterances in the training set. The temporal boundaries of the phone terminals in each parse tree are derived from the forced alignment output. These are used to compute temporal boundaries for the remaining nodes throughout the parse tree.
2. Normalization: This is an iterative process where computation is performed for the entire training data, one layer at a time, from the bottom layer upwards. Within each iteration, beginning at the preterminal phoneme layer, the computation is a two-pass strategy. During the first pass, we compile the mean and conditional mean duration for each node in the current layer which are used to derive the ratio or scaling factor, equivalent to the one we first introduced in Eq. (1). And during the second pass, the scaling factor is applied to every node in the current layer over the entire training set. This yields hierarchically normalized duration for all the nodes in this layer, and we then proceed to the next layer above to perform the same two-pass operation in the next iteration.
3. Relative duration model: To derive model parameters we first represent node duration in relative terms and then collect single-dimensional Gaussian statistics for each and every node with one or more siblings in the parse trees. Thus during scoring, we have log likelihoods available for each node which has one or more siblings in the tree. The number of scores is variable among individual trees, depending on the word and the number of nodes with one or more siblings, and therefore we must derive a total relative duration score by averaging over the number of nodes available.
4. Absolute duration model: An average of the normalized word durations is computed and subsequently we derive the inverse speaking rate for each word by scaling with this average as described in Section 3.4. All the hierarchically normalized phoneme durations are scaled by their corresponding speaking rate as in Eq. (3). Subsequently, we estimate parameters for each Gamma model associated with each phoneme. During scoring, we have a log likelihood available for each phoneme within a word. An absolute duration score for a word is computed as the average log probability score for the set of phonemes within the word.

5. Variance reduction

Model variance is an indicator of how well a model fits the data. A smaller variance corresponds with greater constraint and reliability for a given statistical model. In the following, we demonstrate variance reduction achieved for both hierarchical normalization and speaking rate normalization. We use our automatically generated forced alignments to construct ANGIE parse trees wherein the prescribed normalization procedures are performed. In this section, we examine the changes in variance that result.

5.1. Hierarchical normalization

We consider the amount of variance reduction attributed to hierarchical normalization at the WORD level in the training data. It is found that the standard deviation for the WORD node is reduced from 180 ms to 109 ms, a 39% reduction. The normalized histogram is plotted in Fig. 4.

The resultant normalized distribution is much smoother in appearance and considerably better fitted by the Gaussian function.

5.2. Speaking-rate-normalized absolute duration

At the phoneme layer, average standard deviation is reduced from 50 ms to 33 ms, a 33% reduction. This is attributed to two simultaneous effects: (1) correcting for phonological realization via the hierarchical normalization procedure and (2) compensating for speaking rate effects by normalization with the speaking rate parameter. Large variance reduction is indicative of the potential success of these models.

Fig. 5 plots the histogram of data for the special phoneme /*uw*/ that appears exclusively in the function word "to", before and after the combined normalization procedures. Standard deviation is reduced from 68 ms to 22 ms while mean duration is 75 ms prior to normalization and 73 ms after normalization. Normalization achieves greater smoothness in the statistical distribution. The range of values for duration is widely reduced and the skewed shape exhibited by the original unnormalized distribution has been corrected.

5.3. Non-uniform effects of speaking rate

We have implemented a linear speaking rate normalization, which implicitly assumes a linearly proportionate influence of speaking rate on durational segments. However, when observing gains in variance reduction in individual sublexical categories such as those in the phoneme set, the resultant reduction differs from one category to another. While our normalization imposes the assumption that speaking rate influences each

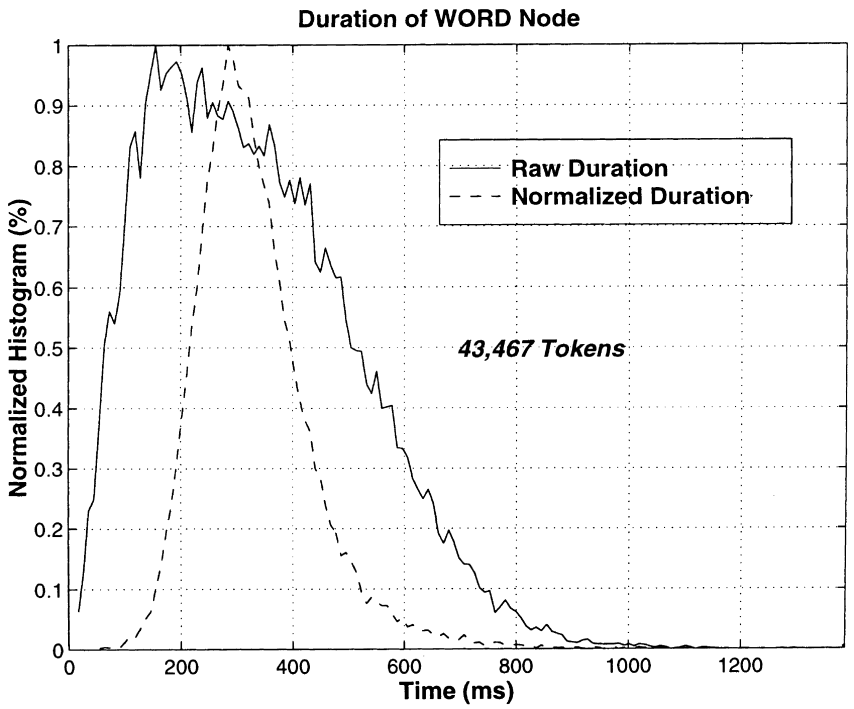

Fig. 4. Hierarchical normalization for the WORD node: standard deviation is reduced from 180 ms to 109 ms, a 39% relative reduction.

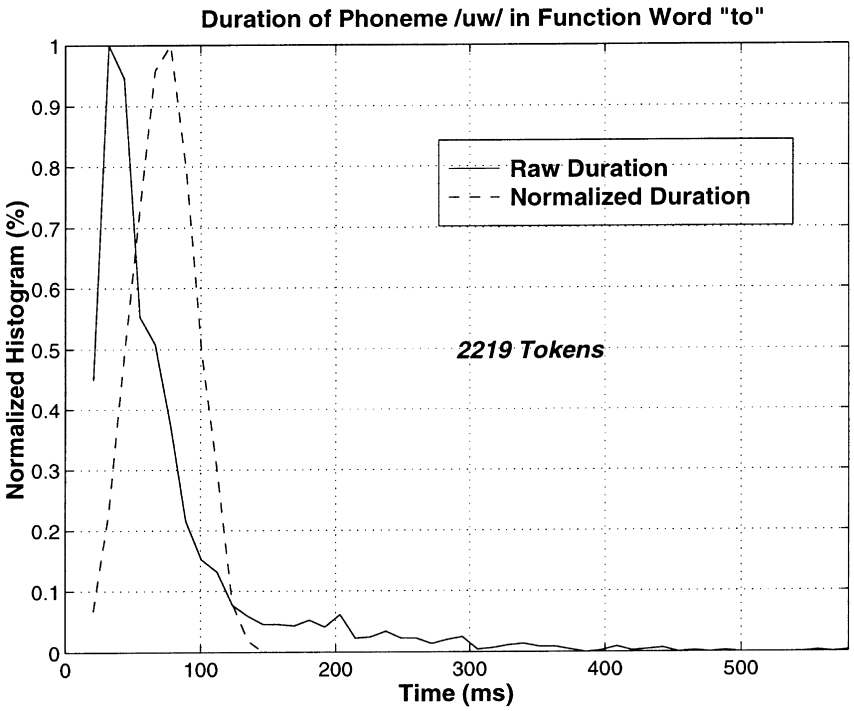

Fig. 5. Hierarchical and speaking rate normalization for the special phoneme /*uw*/ in the function word "to". Standard deviation is reduced from 68 ms to 22 ms, a 68% relative reduction.

sublexical unit in a uniform manner, it is apparent that this is not true. To the contrary, there is evidence to suggest that speaking rate affects different nodes to a varying extent and some sublexical units are more susceptible to speaking rate changes than others. We will observe this in Section 6. However, in spite of this, for all phonemes, we found that speaking-rate-normalization yields a reduction of variance, indicating that segments do compress and expand as we expect for fast and slow speech, and a first order approximation has been effective, although the extent of compression/expansion varies according to the segment identity, and remains to be explained or accounted for more precisely in a model.

The challenge, then, is to somehow improve 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/compress as a function of speaking rate. The problem is to (1) identify these nodes and (2) determine the relative weights with which we can perform speaking rate normalization. It may also be beneficial to characterize these nonlinear effects in terms of broad classes or place or manner of articulation. For example, vowels tend to vary in duration much more significantly than stop closures. Nonetheless, the knowledge of how speech rate affects duration nonlinearly can potentially provide significant additional gains to our model. We will see in the next section more evidence of this type of nonlinear behavior. We would ultimately like to incorporate such second order nonlinear effects into a more sophisticated duration model.

6. Speaking rate experiments

Our speaking rate parameter and hierarchical paradigm allow us the flexibility to conduct various experiments exploring speech timing phenomena, particularly those associated with the rate of speech. In the following, we present two studies, the first of which investigates the influence of the rate of speech on relative durations of sublexical units and the second will study closely the type of words that tend to be spoken anomalously slowly.

We are motivated by the potential for incorporating more durational knowledge into a computational model to enhance speech recognition.

6.1. Speaking rate effects

We have chosen to examine the effects of varying speaking rate on relative durations of sublexical units. The training set, of 44,467 words, is partitioned into three equal sized subsets of slow, medium and fast speech. The average inverse speaking rates for these subsets are 1.35, 0.95 and 0.70 respectively, where 1.0 represents an average speaking rate for the entire training set. The mean relative durations for sublexical units are computed in each subset and compared across varying rates. We will present here a few examples taken from various layers of the ANGIE parse tree to illustrate the nature of these secondary effects induced by speaking rate changes.

We consider a WORD node that is realized by the sequence (PREFIX, STRESSED ROOT, SUFFIX), such as in the word, "December." It is found that in our training data, the SUFFIX node progressively occupies proportionately more of the WORD duration from 29% to 34% of the total WORD duration, as speaking rate changes from fast to slow. This is shown in Fig. 6. On the contrary, the PREFIX occupies proportionately less of the total WORD duration, from 28% to 23%, and the percentage duration occupied by the STRESSED ROOT remains constant at 43%. This indicates the presence of a non-uniform effect of speaking rate on sublexical components and a resistance to change on the stressed syllable of the word. Consequently, the SUFFIX part of the WORD absorbs more expansion and contraction at different rates of speech.

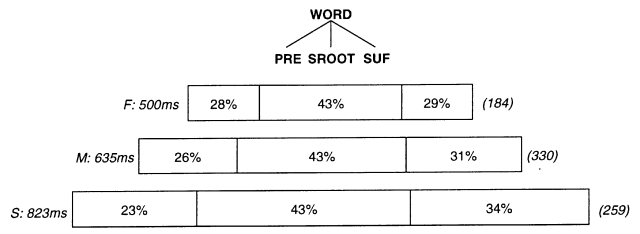

Fig. 6. Relative duration for parts of WORD node corresponding with 3 speaking rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. An example of this word pattern is in the word "December".

Similarly, within subsyllabic units, speaking rate also displays nonlinear effects. As shown in Fig. 7, an unstressed sequence, (ONSET, NUCLEUS), behaves differently depending on its position within a word. Within a PREFIX, (e.g. in "*to*morrow"), proportionate relationships remain largely constant with respect to speaking rate changes, while within an UNSTRESSED ROOT, (e.g. in "tomor*row*"), subsyllabic components expand non-uniformly. The NUCLEUS occupies from 48% to 59% of the total syllable duration as speaking rate varies from fast to slow. This implies that the NUCLEUS expands proportionately more than the ONSET when speaking rate is comparatively slower. Despite the absence of stress in both types of syllables, their differing position within a word affects significantly their behavior with respect to the rate of speech.

When we examined our phoneme layer, which includes items such as various diphones and morph-specific pseudo-phonemes, and their respective phonological realizations, we found the same nonlinear phenomena to persist as well. In Fig. 8, the special function-word specific diphone */ra/* found in "from" is realized as [r,ah] sequence. When comparing realizations of slow and fast speech, the [ah] phone virtually accounts for all the expansion for slow speech as we expand from 73% to 86% of the total diphone duration. On the other hand, the [r] phone occupies a much smaller portion of the total diphone when the speech is spoken slowly. As a second example in Fig. 8, another diphone */en/*, which exclusively occurs in "and", is examined and we find that, by contrast, the nasal lengthens proportionately more than the preceding vowel when speech is realized comparatively slowly, from 39% to 49% of total duration of the */en/*. Therefore we cannot conclude whether, in general, the vowel or consonant tends to bear more of the change in duration for a particular diphone pair.

It should be noted that, within all the patterns where we have observed nonlinear lengthening, the final sublexical unit consistently lengthens the

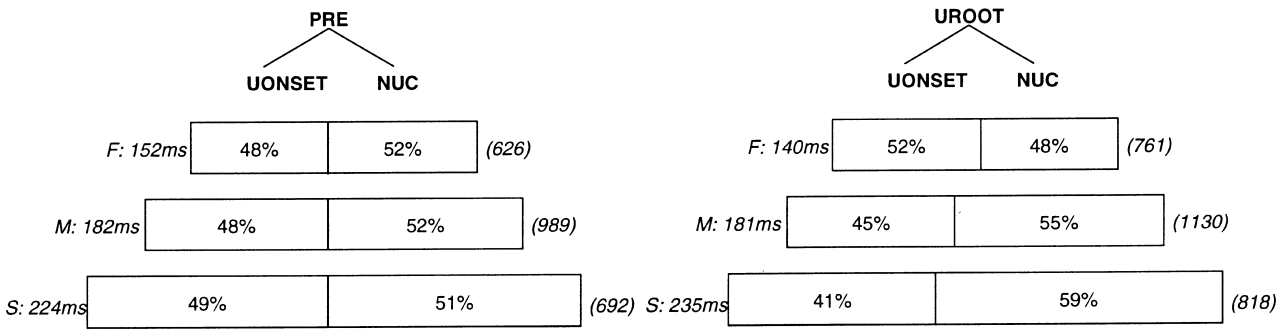

Fig. 7. Relative duration for subsyllabic segments of the PREFIX and UNSTRESSED ROOT nodes corresponding with 3 speaking rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. Both are realized by the (UONSET NUC) sequence. Examples are the prefix and unstressed root in the word "tomorrow".

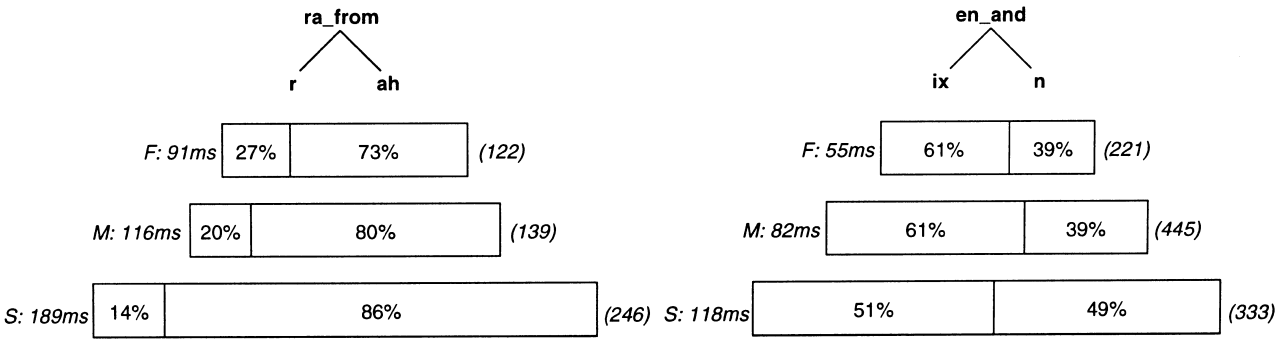

Fig. 8. Relative duration for phone segments within phoneme units corresponding with 3 speaking rates: F: Fast speech, M: Medium speech, S: Slow speech. Average absolute duration at each speaking rate is given in ms. The number of tokens in each set is indicated in brackets. Examples are function-word specific diphones in the words "from" and "and".

most in proportion. Strikingly, this trend ensues from the subsyllabic level to the morphological level. This can be seen as a consistent increase in capacity for proportionate lengthening towards the segments positioned later in time. This applies for any unit in the hierarchy, whether it be a word, syllable or subsyllabic unit. Within a WORD node, we have observed the burden of lengthening to be carried proportionately more by the SUFFIX. While below that level, lengthening appears to be linearly proportionate within the word-initial PREFIX node. But within the UNSTRESSED ROOT, which occurs in a later position in the word, its segments or child nodes are characterized by greater lengthening towards the final. This repeated manifestation of greater elasticity in later segments at each layer of the ANGIE tree, in a somewhat recursive manner, is particularly intriguing, and may suggest the possibility for a more formal model for describing duration phenomena via a new hierarchical elasticity principle.

Finally, we 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, as would be expected. Their absolute durations change only slightly, while the closure expands and compresses more responsively according to speech rate. This is depicted in Fig. 9. Here, all stop phonemes in the onset position are partitioned into 5 equal subsets of data, ordered 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 according to the identity of each stop.

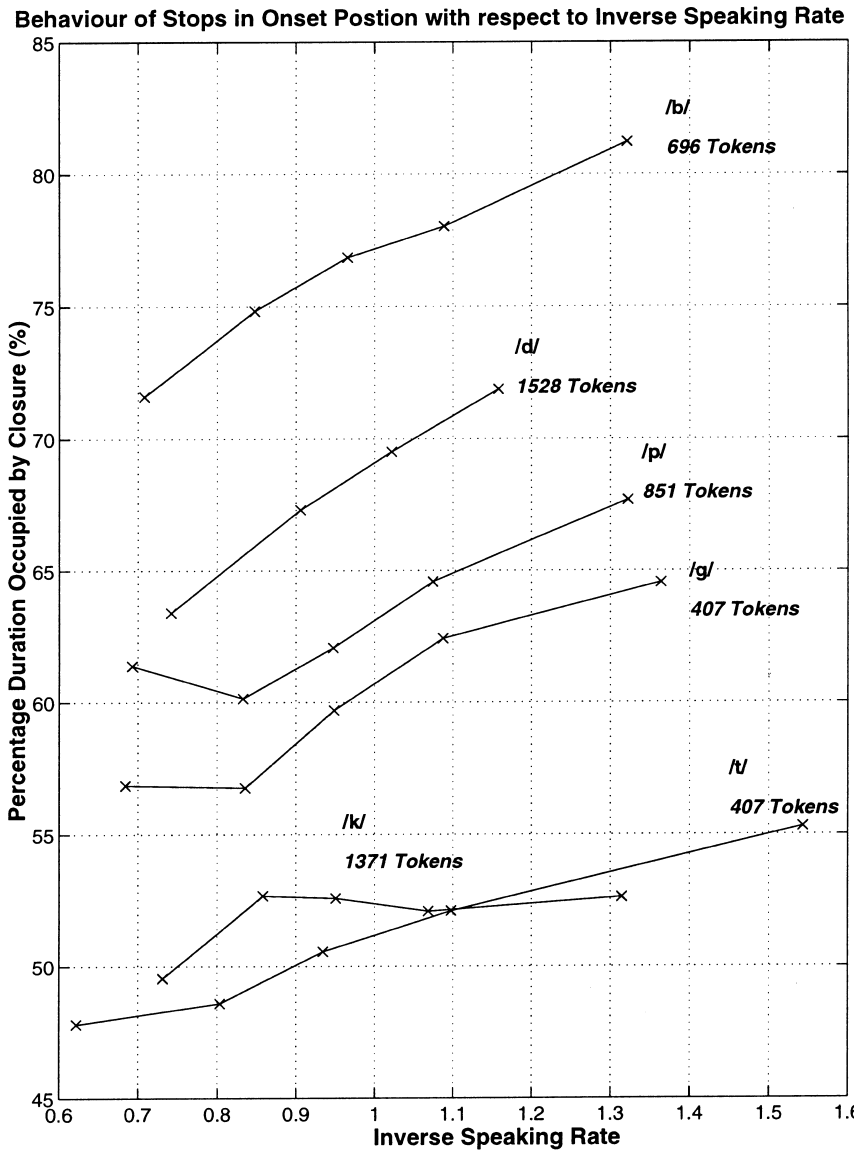

Fig. 9. Behaviour of stops in onset position with respect to speaking rate: each of the six stops are divided into five equal subsets of data in accordance with speaking rate of the word they occur in. Average percentage duration occupied by stop closure within a stop phoneme is plotted as a function of average speaking rate for each subset, represented by a "×" on the plot.

The above results have provided evidence of the non-uniform effect of speaking rate on sublexical units and given some insights on which linguistic units are more susceptible or resistent to changes. Not only is this non-uniformity complex but it is also confounded by many contextual factors which are not well understood. There is not yet a systematic way of explaining such behavior, and sparse data prevents us from investigating vast numbers of examples of sublexical patterns among the multitude of combinations possible.

6.2. 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 pathological nature. All words which have a speaking 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

Table 1
A list of anomalously slow words, specified as function words or otherwise, and their frequency of occurrence in the training set

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		

author to be inaccurate or incorrect are omitted from analysis. The result is 102 correctly aligned words with inverse rate greater than 2.5. The maximum speaking rate is 5. Table 1 sets out the list of these words and their frequency of occurrence.

Fifty-five (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. At a semantic level, it is useful to note that slow speech occurs at function words, just prior to content words. In most cases, pauses are inserted before these content words. This is not surprising as the speakers tend to pause before words of high lexical content in spontaneous speech and the function words preceding them are prepausally lengthened. We will explore prepausal lengthening further in the next section.

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.

7. Studies characterizing prepausal lengthening

Our current paradigm addresses contextual effects that take place below the word level. But it is known that many factors above this determine the behavior of segmental duration. One particular phenomenon is *prepausal lengthening*. This refers to the effect where segments followed by pauses are lengthened in duration. Studies report a 60–200 ms increase in syllable duration (Klatt, 1976). Many studies have reported on the nature and effects of lengthening before pauses (Campbell, 1991; Klatt, 1976; Umeda, 1975, 1977; Wightman et al., 1992). Although we do not explicitly model prepausal lengthening, we would like to gain greater knowledge about its characteristics using our representation. We investigate here (1) the relationship between prepausal lengthening and the rate of speech as defined by our parameter as well as (2) the influence of the duration of the pause.

7.1. Speaking rate and prepausal lengthening

The first experiment explores the relationship between speaking rate and prepausal speech. 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, sentence-internal pauses are marked by "inter-word trash" or [iwt] and words which are followed by these are tagged. Words which appear at the end of an utterance are also marked as prepausal words. For this experiment, the two types of prepausal words are considered indistinguishable. In Fig. 10, 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 by inserting pauses afterwards. In any case, slow speech and prepausal effects are interrelated and speaking rate can be used to anticipate the presence of a pause.

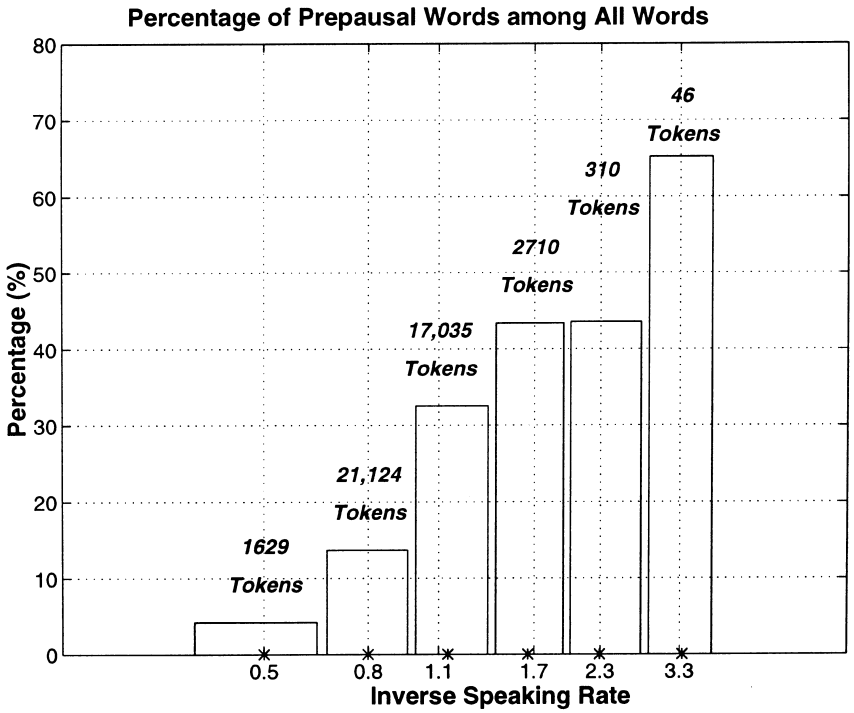

Fig. 10. Percentage of prepausal words among all words: all words are partitioned into subsets according to their measured speaking rate and arranged on a histogram plot. Each subset has approximately equal bin width on the logarithmic x-axis. The percentage of words that are prepausal in each bin is plotted.

7.2. 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 Fig. 11, 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 150 ms. 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 200 ms, the average speaking rate is above 1.2 and is significantly higher when the pause is extremely long (>700 ms). This suggests that prepausal data which are followed by pauses of less than 100–150 ms may not be considered prepausally lengthened but rather treated as the same as non-prepausal data.

8. Phonetic recognition

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, in spite of unlikely segmental durations. A duration score will encourage the rejection of faulty segmentations by penalizing the unlikely hypotheses and

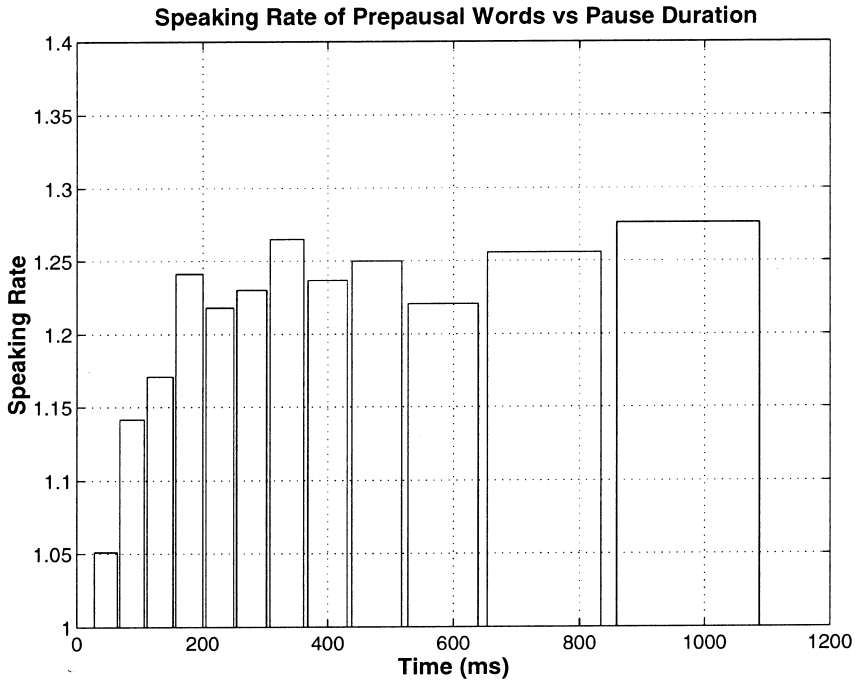

Fig. 11. Average inverse speaking rate of prepausal words versus duration of pause: all sentence-internal prepausal words are divided into equal bins of 400 tokens according to duration of corresponding pause. In each bin, the average speaking rate of the prepausal words is plotted.

boosting probable ones. We would like to demonstrate the utility of durational information for speech recognition by first incorporating our model at the level of phonetic recognition. We use an ANGIE-based phonetic recognizer, previously described in (Seneff et al., 1996). As we shall see, this ANGIE-based system offers higher level linguistic constraint by way of sublexical parse trees wherein our duration model can derive contextual information that is highly beneficial to the performance of the recognizer. We conducted recognition experiments under three scenarios of varying levels of linguistic constraint. We did so by altering the amount of lexical knowledge available to the recognizer. The experiments will show that when more linguistic information is provided, the duration model is more effective and yields higher gains.

8.1. The ANGIE recognizer

In the ANGIE phonetic recognizer, the recognizer does not have explicit knowledge of the word lexicon although the ANGIE subword model probabilities have been trained on phonetic sequences associated with orthographic transcriptions for ATIS sentences; that is to say, lexical knowledge is only implicitly used during training. During recognition time, word-like structures or pseudo-words are proposed whenever it is allowable by the ANGIE probabilities for the current sequence of phones. At this point the duration component is called to produce a duration score. This score is a linear combination of scores derived individually from the relative and absolute duration models. These are log probabilities averaged over all nodes within a pseudo-word in the parse tree for the case of relative duration and over all phonemes in a pseudo-word for the case of absolute duration. This score becomes a weighted component of the total word score. All weights are empirically determined and optimized over an independent development set. 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.

8.2. Linguistic constraint

Experiments are conducted under three separate scenarios of varying degrees of linguistic constraint. In the first case, recognition is performed with the sole constraint of the ANGIE parse tree. That is, phonetic sequences are allowable whenever an ANGIE parse can be completed. For the remaining two cases, we have performed experiments by providing the recognizer with the added linguistic constraint given by further implicit lexical knowledge. This is accomplished by adopting a two-tier approach to the lexicon. The word lexicon is represented by sequences of intermediate morph or syllable units which themselves are represented by phoneme sequences. The word and morph lexicon are both derived from the ATIS corpus.

So under the second case, the recognizer filters partial theories of phonetic sequences in accordance with these syllables or morphs available in the lexicon. Phoneme sequences are only permissible if they form legal morphs or syllables as set out in the morph lexicon. In this way, the morphs provide the extra linguistic constraint by permitting only those syllable patterns which appear in the lexicon and disallowing those not specified in the lexicon. However, random sequences of lexically licensed morphs are allowed regardless of whether they actually appear in this sequence in the word lexicon, that is, whether they form legitimate words. These constraints are expected to boost the recognizer's performance because by only allowing phoneme sequences licensed by syllables in the vocabulary, the search space is somewhat reduced.

In the final scenario, we add the additional constraint that only morph sequences which appear in the 1200-sized 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 due to the extra lexical knowledge. All the pseudo-words which are proposed by the recognizer will then be able to form actual words that appear in the word lexicon and, in principle, could be retrieved for word recognition results which corresponds with a word recognizer. However, this recognizer is devoid of a word language model since no unigram or bigram probabilities are employed. All linguistic constraints operate at the sublexical level. In any case, the phonetic error rate should provide a much improved 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. Phonetic recognition results are obtained for the three scenarios without applying duration, and with a combination of relative and absolute duration models applied (as described in Section 4).

8.3. Results and discussion

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. Our results, tabulated in Table 2, show that the duration model yields improvement in all three cases. A maximum gain is achieved when word constraints are applied, and using duration yields a 7.7% relative error reduction. The baseline system already contains standard context-independent phone duration models and all gains reported are

Table 2
Phonetic error rates and relative error reduction (Δ) due to adding hierarchical duration model. Results correspond with three scenarios of linguistic constraint as described in Section 8.2. The baseline system contains a standard phone duration model whereas the hierarchical model combines scores from both relative and absolute duration models

Linguistic constraint	Baseline error (%)	Error with hierarchical duration (%)	Δ (%)
ANGIE tree only	33.2	32.7	1.5
Morph constraints	31.8	30.9	2.8
Word constraints	29.7	27.4	7.7

over and above those realized from these baseline models.

It can be seen that our duration model is more effective with increasing linguistic constraint or lexical knowledge. The reason is that for case two and three, at better baseline performance, the recognizer not only proposes greater numbers of correct phones but also more sensible phonetic sequences which constitute syllabic units, and whose corresponding parse trees are more useful to the duration model. This enables the model to apply its contextual knowledge embedded among the multiple layers of the sublexical tree. On the other hand, when these added constraints are removed, the recognizer's performance degrades, for one, and, secondly, the presence of several incorrectly hypothesized phones in a sequence will cause nonsensical pseudo-words to be proposed. Although nonsensical syllables are permitted to occur, their ANGIE scores will be low because they are not well supported by the training data. In any case, the resultant incorrect ANGIE parse structure is likely to deviate significantly from that of the underlying word we are attempting to recognize, and so the information that we pass on to our duration model is largely meaningless. So in this case, much of the higher level linguistic information that duration can provide cannot be effectively utilized.

We can conclude that this duration model has been successful at reducing error rate and we have demonstrated its utility in phonetic recognition. Our results show a complex duration model, which implicitly incorporates a large amount of contextual 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. We will see this trend again in the next section.

9. Word spotting

As a first step towards demonstrating the benefit of duration to word recognition, we choose to evaluate our model in word spotting using an ANGIE-based system described in (Lau and Seneff, 1997; Lau, 1998). We begin with a pilot study in which duration is used as a post-processor. The task is to disambiguate between two acoustically confusable keywords, "New York" and "Newark". This pair constitutes the largest source of error in our recognizer for this ATIS task, being very acoustically similar. The results produced by the word spotting system are input to the duration component for subsequent rescoring. Duration is potentially a powerful feature for distinguishing between the two words. The success of this experiment will demonstrate the power of duration as an aid to word recognition. We will proceed to describe our experience in fully integrating the duration model into the word spotting system. We shall see that duration modelling brings substantial benefits to the overall performance of the word spotting system.

9.1. Pilot experiment

The word spotter has been assigned the task of spotting keywords from a set of utterances which propose "New York" in the ATIS domain. The most outstanding alignment error for the word spotter 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. [5] 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

[5] 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 word spotting system, all detected "Newark" waveforms could also be passed on to the post-processor and an even greater number of errors can be reduced.

neither of the two keywords are eliminated by hand because they are irrelevant to this task. [6]

The duration processor itself consists of its own ANGIE forced alignment system along with a duration model. This configuration of realigning data in the post-processor is not ideal in the sense of computational efficiency, but is used here for the purpose of this experiment whose goal is solely to demonstrate the utility of the duration model. 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" has been detected. This is necessary because the word spotter does not output a "Newark" alignment whenever it spots a "New York". For consistency, we realign both words, using the ANGIE system within the post-processor. Each alignment is then scored for duration. Subsequently, each utterance now has a set of scores derived from duration, acoustics and linguistics, associated with both "New York" and "Newark". These scores are combined to yield a total score. The acoustic and linguistic scores are added together without optimization, whereas the duration score is scaled by an empirically determined weight before combining. The post-processor makes a decision between "New York" and "Newark" by choosing the one with the higher total score.

Table 3 tabulates the baseline performance and the improvement gained by adding the relative and absolute duration model separately. 323 utterances were passed from the word spotter to the duration component. The absolute duration model appears to offer the best performance. The use of duration modelling has generated improvement on a dramatic scale. However, firstly, it should be noted that the duration weight has been optimized for performance over the same 323 utterances. And secondly, in this specific task, we have precluded all other false alarms except for "Newark" waveforms. Therefore, the duration post-processor cannot rectify these other errors nor can it reduce the number of misses. Nonetheless, this task has demonstrated that duration is an important candidate for consideration, especially for specific instances in which word pairs have confusable acoustic-phonetic features, allowing duration weights to be much larger compared with those used in the phonetic recognition experiments. We believe that the performance gains can be attributed to two reasons:

1. Durational information takes on greater importance when recognizing whole words. Here, we have overcome the problem encountered in phonetic recognition where entirely incorrect ANGIE parse trees emerge due to the presence of a few incorrect phones in the hypothesis, thereby preventing optimal use of the duration model. Instead, we are able to take full advantage of various sublexical structures in the hierarchical representation.
2. This task is of a particularly specific nature where it is known a priori that duration should be advantageous over other acoustic features. A very large duration weight is used and this may not have been applicable for other word pairs or more general tasks. One can expect that for other choices of words durational information is less valuable.

We will evaluate the performance of our duration model in a more general task in the next section.

Table 3
Error rate and relative improvement (Δ) of using duration post-processor to disambiguate "New York" from "Newark" in word spotting in a pilot experiment described in Section 9.1

Duration model	Error rate (%)	Δ (%)
None	19	–
with relative duration	8	58
with absolute duration	6	68

9.2. Fully integrated system

For this experiment, the same duration model is fully integrated into the ANGIE-based word spotter in which the model is employed for a full set of keywords. The details of the word spotting

[6] In any case, during recognition time, 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.

system are documented in (Lau and Seneff, 1997; Lau, 1998). A duration score is computed whenever a keyword is proposed and combined with the acoustic and linguistic scores to produce a total word score. As in our phonetic recognition experiments, varying levels of implicit lexical knowledge are imposed to provide progressively greater linguistic constraint, and consequently leading to better baseline performances. As previously described, the word lexicon is represented by sequences of intermediate morph or syllable units which, in turn, are represented by phoneme sequences in a predefined morph lexicon. These morph units, as in Section 8, are based on syllable structures. As in our phonetic recognition experiments, 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 finetuned by the penalty imposed. When the penalty is very large, morph sequences which do not form words are highly discouraged and we effectively 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 in our phonetic recognition experiments, when 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.

Table 4
Results of word spotting experiments using three levels of linguistic constraints as described in Section 9.2. The baseline system contains a standard phone duration model whereas the hierarchical model combines scores from both relative and absolute duration models

Experiment	Baseline (FOM)	Hierarchical duration (FOM)
Using morph constraints	88.4	89.8
Using morph constraints with word penalty	88.6	90.0
Using full word constraints	89.3	91.6

Results are tabulated in Table 4. We use a standard measure for evaluating word spotting performance, the *Figure of Merit* (FOM), as set out by the procedures in (The Road Rally Word-Spotting Corpora, 1991). The best performance yielded a significant improvement from 89.3 to 91.6 (FOM), by employing a combination of the absolute and relative duration models. This demonstrates that duration can be applied in a more general task and still yield substantial gains. Here, as we would expect, the relative improvement is not as dramatic as that produced in our pilot study given our greater generality. But we speculate that duration is particularly useful for a general keyword spotting task such as this where the keywords consist of a set of proper nouns. Most keywords contain complex syllabic structures, enabling their durational relationships to be utilized by our models efficiently. Words such as function words and single syllable words contain less structure from which the duration model could derive information.

10. Conclusion and future work

We have developed a statistical hierarchical duration model which captures contextual phenomena at multiple sublexical levels. This novel paradigm which encompasses a speaking rate

measure constitutes a useful mechanism for conducting experiments in speech timing. One can discover and quantify various phenomena in speech timing under this regime. Our motivation is to use durational information to aid speech recognition. And in our experiments, the combination of using relative and speaking-rate-normalized absolute duration has brought performance improvement in both phonetic recognition and word spotting. Our results support the claim that this model, which captures information embedded at morphological, syllabic and phonemic levels, becomes more useful when more lexical knowledge is available. In addition, our experiments suggest that duration can play a particularly important role for certain specific words.

Our experiments have shown that, at a secondary level, duration of sublexical units in the word change non-uniformly with respect to speaking rate. As we have seen in our experiments, these nonlinear changes can be analyzed at different levels of the phonological hierarchy and are affected by all the contextual variables at these levels. In contrast, we impose the assumption in our models that the effect of the rate of speech on each sublexical component is linear. This amounts to a first order approximation. Our normalization procedures and relative duration models rely upon the claim that sublexical units preserve their relationships with one another and compress/expand by the same relative amount. In actuality, this assumption has yielded very good results in terms of recognition improvement. We believe that a more sophisticated model can progress beyond the first order linear approximation and incorporate knowledge about the second order interactions due to speaking rate. This holds the potential for even better performance for our duration model.

We would like to extend our representation to include phenomena that reside above the word level, as we observe that much variability in the duration patterns may be linked to syntactic and semantic level. These effects can be incorporated under a similar paradigm. Our ultimate goal then is to explore the benefit of employing duration in a continuous word recognition task, taking into account the higher level linguistic effects as well.

Acknowledgements

The authors would like to thank Victor Zue for many helpful suggestions on earlier versions of this paper. Also, the authors thank Joshua Koppelman and Raymond Lau for their assistance on numerous aspects of the implementation of this work.

References

Campbell, W.N., 1991. Segment durations in a syllable frame. Journal of Phonetics 19, 37–47.

Campbell, W.N., 1992. Syllable-based segmental duration. In: Bailly, G., Benoit, C., Sawallis, T.R. (Eds.). Talking Machines: Theories, Models, and Designs. Elsevier, Amsterdam, pp. 211–224.

Chung, G., 1997. Hierarchical duration modelling for a speech recognition system. S.M. Thesis, MIT Department of Electrical Engineering and Computer Science.

Crystal, T.H., House, A.S., 1988. Segmental durations in connected speech signals: Current results. J. Acoust. Soc. Amer. 83 (4), 1553–1573.

Dahl, D., et al., 1994. Expanding the scope of the ATIS task: the ATIS-3 corpus. In: Proceedings ARPA Human Language Technology Workshop, Plainsboro, NJ, pp. 43–48.

House, A.S., 1961. On vowel duration in English. J. Acoust. Soc. Amer. 33 (9), 1174–1178.

Klatt, D.H., 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. J. Acoust. Soc. Amer. 59 (5), 1208–1221.

Lau, R., 1998. Subword lexical modelling for speech recognition. Ph.D. Thesis, MIT Department of Electrical Engineering and Computer Science.

Lau, R., Seneff, S., 1997. Providing sublexical constraints for word spotting within the Angie framework. In: Proceedings Eurospeech '97, Rhodes, Greece, September 1997, pp. 263–266.

Pallet, D., Fiscus, J., Fisher, W., Garofolo, J., Lund, B., Martin, A., Przybocki, M., 1995. 1994 benchmark tests for the ARPA spoken language program. In: Proceedings ARPA Spoken Language Systems Technology Workshop, Austin, TX, January 1995, pp. 5–36.

Pitrelli, J.F., 1990. Hierarchical modeling of phoneme duration: Application to speech recognition. Ph.D. Thesis, MIT Department of Electrical Engineering and Computer Science.

The Road Rally Word-Spotting Corpora (1991). The Road Rally Word-Spotting Corpora (RDRALLY1). NIST Speech Disc 6-1.1, September 1991.

Seneff, S., Lau, R., Meng, H., 1996. ANGIE: A new framework for speech analysis based on morpho-phonological modelling. In: Proceedings of the IC-SLP Philadelphia, PA, October 1996, Vol. 1, pp. 110–113.

Umeda, N., 1975. Vowel duration in American English. J. Acoust. Soc. Amer. 58 (2), 434–445.

Umeda, N., 1977. Consonant duration in American English. J. Acoust. Soc. Amer. 61 (3), 846–858.

Van Santen, J.P.H., 1992. Contextual effects on vowel duration. Speech Communication 11, 513–546.

Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P., 1992. Segmental duration in the vicinity of prosodic phrase boundaries. J. Acoust. Soc. Amer. 91, 1707–1717.

Zue, V., Seneff, S., Polifroni, S., Phillips, M., Pao, C., Goddeau, D., Glass, J., Brill, E., 1994. The MIT ATIS system: December 1993 Progress Report. In: Proceedings of the ARPA Spoken Language Technology Workshop, Plainsboro, NJ, pp. 67–71.