HIERARCHICAL DURATION MODELLING FOR SPEECH RECOGNITION USING THE ANGIE FRAMEWORK[1]

Grace Chung and Stephanie Seneff

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
http://www.sls.lcs.mit.edu, mailto:{graceyc, seneff}@mit.edu

ABSTRACT

We describe a novel hierarchical duration model for speech recognition. The modelling scheme is based on the ANGIE framework, a flexible unified sublexical representation for speech applications. Our duration model captures contextual factors that influence duration of sublexical units at multiple linguistic levels simultaneously, using both relative and absolute duration information. The modelling procedure involves a normalization scheme which produces a new measure for relative speaking rate at a word level. This may be used to explore phenomena in speech timing and we present studies on secondary effects of speaking rate here. This duration model demonstrates its ability to aid speech recognition in phonetic recognition experiments where it has yielded a relative improvement of up to 7.7%. In word spotting, a study employing duration as a post-processor in disambiguating between 2 acoustically similar keywords reduces relative error by 68%. Furthermore, a fully integrated duration model in an ANGIE based word spotter improves performance by 21.5%. All gains are over and above any gains realized from standard phone duration models present in the baseline system. All experiments were conducted in the ATIS domain, using continuous spontaneous speech.

1. INTRODUCTION

Durational patterns of phonetic segments and pauses convey information about the linguistic content of an utterance. Given that such duration 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 ([4]) and consequently the ability to handle such variations could boost performance. However, there is a vast array of factors that influence speech timing. These reside at multiple levels of the phonetic hierarchy and their complex interactions are not well understood. For this reason, most speech recognition systems grossly underutilize the knowledge provided by durational cues, while most extensive studies in speech timing have been directed towards synthesis applications ([1, 6]).

Here, 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 has been applied to a variety of tasks, namely, letter-to-sound/sound-to-letter generation, phonetic recognition and wordspotting ([5, 3]). In this work, we use ANGIE to model contextual effects in duration from the phone level up to the word level. Here, we demonstrate its capability as a mechanism for investigating speech timing phenomena by examining secondary effects of speaking rate changes. The duration model is incorporated into a speech recognizer and our experiments show that duration offers substantial performance gains in both phonetic recognition and word spotting, thereby demonstrating the value of duration as an aid to speech recognition. For further details, consult [2].

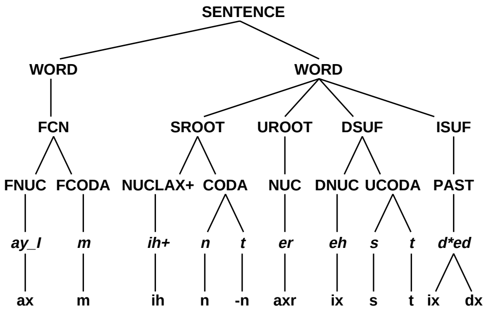


Figure 1: *Sample parse tree for the phrase "I'm interested...".*

2. THE ANGIE FRAMEWORK

ANGIE is a trainable probabilistic paradigm in which word substructures are jointly characterized by a context-free grammar and are represented in a multi-layered hierarchical structure. An example is shown in Figure 1. The upper layers capture syllabification, morphology, and stress, while the preterminal layer represents phonemics, and the bottom terminal categories are phones. Further details of the probability model and grammar are described in [3, 5].

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, 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. We present a novel normalization procedure which merges statistical distributions together so that sublexical units such as syllables can be independent of their phonemic realizations, and, similarly, phonemic durations are normalized by their phonetic variations. Upon normalization,

[1] This research was supported by a research contract from NYNEX Science and Technology. G. Chung is also supported by the Australian Fulbright Commission.

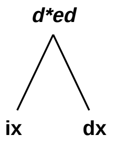


Figure 2: *Phoneme /d*ed/ realized as ix followed by dx.*

the framework is used to derive two sets of statistical models — one based on relative duration and another based on speaking-rate-normalized absolute duration.

3.1. Hierarchical Normalization and the Relative Duraton Model

We have formulated a normalization scheme which reduces the model variance at each node. Our stategy 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 of the parse tree.

Given a nonterminal node, its normalized duration is equivalent to the sum duration of its child nodes in the immediate layer below, scaled by some ratio predetermined from training data. This ratio is the mean duration of the parent node over the mean duration when the parent node has been realized by the corresponding child nodes. For example, as in Figure 2, when an instance of the pseudo-phoneme */d*ed/*[2] is phonetically realized by a front schwa *ix* followed by a flap *dx*, as in "..interes*ted* in..," its normalized duration, as illustrated by Eqn 1 below, is equivalent to the sum of the *ix* and *dx* durations, derived from their time alignment, and scaled thereafter by a ratio. 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 an *ix* followed by a *dx*.

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

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. According to this strategy, individual probability distributions, corresponding to different realizations of the one parent node, are collapsed together, and, as a consequence, model variance at each node is reduced.

Upon normalization, Gaussian statistics are collected from training data at each node throughout the parse tree to build the relative duration model. The duration of each node is calculated as a percentage of the total duration of its corresponding parent node. 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.

3.2. Speaking Rate Parameter and Absolute Duration Model

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. We define a parameter that hypothesizes a speaking rate at the end of each word. After normalization is propagated up to the WORD node, its duration is expected to be independent of inherent durations of its descendents, and thus is an indicator of speaking rate. Henceforth, a word-level speaking rate parameter can be formulated as the ratio of the normalized WORD duration over the global average normalized WORD duration:

$$\text{Speaking Rate}_i \triangleq \frac{\mathrm{DUR}_i(\textsc{word})}{\mu_{\mathrm{DUR}}(\textsc{word})} \tag{2}$$

Effectively, this is a measure of *relative* speaking rate whereby a large value corresponds with slow speech and a small value corresponds with fast speech. Computed at a word level, it has the ability to capture speaking rate variations within a sentence. As we shall demonstrate, it is a useful paradigm for investigating speech timing effects.

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. We construct speaking-rate-normalized absolute duration models at a phonemic level by dividing phoneme durations by the speaking rate associated with the given word:

$$\mathrm{NDUR}_i(\textsc{node}) \triangleq \frac{\mathrm{DUR}_i(\textsc{node})}{\text{Speaking Rate}_i} \tag{3}$$

NDUR represents node duration which has been normalized, and hence is independent of speaking rate effects. The subsequent statistical models use Gamma probability functions.

4. CORPUS

All experiments are conducted in the ATIS domain, using spontaneous speech. Our models are trained from approximately 5000 utterances in the ATIS-3 subset. This set is also used in our speaking rate experiments. For testing, we use the Dec '93 test set. Our relative duration models contain 654 distinct subtree patterns while our absolute duration models comprise 94 phonemes. The phonetic inventory consists of 64 unique categories.

5. VARIANCE REDUCTION

We consider variance reduction due to normalization at word and phoneme levels in the training data. It is found that the standard deviation for the WORD node is reduced from 180ms to 109ms, a 39% reduction. The normalized histogram is plotted in Figure 3. At the phoneme layer, average standard deviation is reduced from 50ms to 33ms, a 33% reduction. This is attributed to 1) correcting for phonological realization and 2) compensating for speaking rate effects. Large variance reduction is indicative of the potential success of these models.

6. SPEAKING RATE EFFECTS

One application of our duration model is to investigate various phenomena in speech timing. We have chosen to examine the effects of varying speaking rate on relative durations of sublexical units. The training set, of approximately 43,000 WORD tokens, is partitioned into three equal sized subsets of slow, medium and fast speech. The mean relative durations for sublexical units are computed in each subset and compared across varying rates.

[2] **ed* is a past tense marker

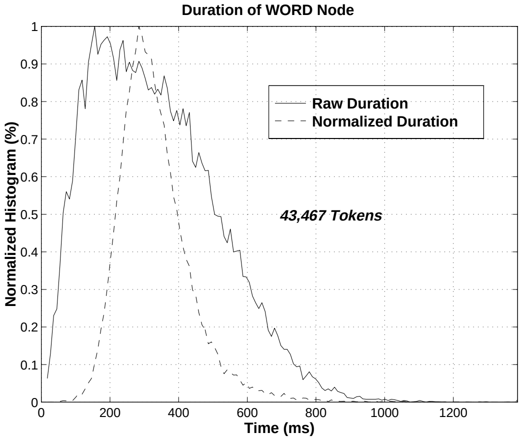
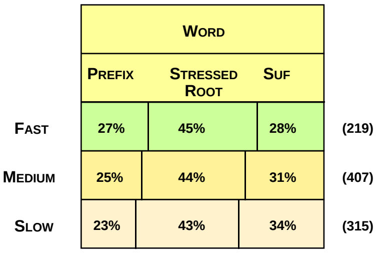


Figure 3: *Reduction of standard deviation due to hierarchical normalization for the* WORD *node*



	WORD			
	PREFIX	STRESSED ROOT	SUF	
FAST	27%	45%	28%	(219)
MEDIUM	25%	44%	31%	(407)
SLOW	23%	43%	34%	(315)

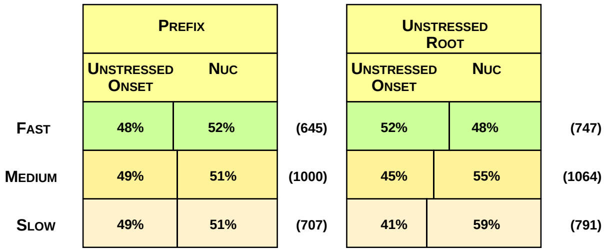
Figure 4: *Mean relative duration for subword patterns at 3 speaking rates.*

It is found that whenever a WORD is realized by the sequence (PREFIX, STRESSED ROOT, SUFFIX), such as in the word, "December," the SUFFIX node progressively occupies proportionately more of the WORD duration as speaking rate slows. This is shown on Figure 4. On the contrary, the PREFIX occupies proportionately less of the total WORD duration and the percentage duration occupied by the STRESSED ROOT remains constant. This indicates the presence of a non-uniform effect of speaking rate on sublexical components.

Similarly, within subsyllabic units, speaking rate also displays nonlinear effects. As shown in Figure 5, 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 rate changes, while within an UNSTRESSED ROOT, (e.g. in "tomor*row*"), subsyllabic components expand non-uniformly, that is, the NUCLEUS expands proportionately more than the ONSET. These observations suggest that the effects of speaking rate are complex and contingent upon other variables such as position within word and within syllable.

7. PHONETIC RECOGNITION

We would like to demonstrate the utility of durational information by first incorporating our model into an ANGIE based phonetic recognizer, previously described in [5]. In the ANGIE recognizer, pseudo-words are proposed periodically, at which 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 word in the parse tree for the case of relative duration and over all phonemes in a word for the case of absolute duration. This score becomes a weighted component of the total word score.



	PREFIX			UNSTRESSED ROOT		
	UNSTRESSED ONSET	NUC		UNSTRESSED ONSET	NUC	
FAST	48%	52%	(645)	52%	48%	(747)
MEDIUM	49%	51%	(1000)	45%	55%	(1064)
SLOW	49%	51%	(707)	41%	59%	(791)

Figure 5: *Mean relative duration for subsyllabic patterns at 3 speaking rates*

7.1. Linguistic Constraint

We experiment with varying the levels of lexical constraint and observing the gains offered by adding duration at each level. In the first case, recognition is performed with the sole constraint of the ANGIE parse tree. For the remaining two cases, we have adopted a two-tiered approach to the lexicon. The word lexicon is represented by sequences of intermediate morph/syllable units, which are themselves associated with phoneme sequences. By allowing only phoneme sequences that exist in predefined morph patterns, we have added increased linguistic constraint which should boost baseline performance. In the final scenario, only morph sequences licensed by the 1200-sized word lexicon are allowed. That is, theories producing morph sequences that do not exist in the word lexicon are pruned. This amounts to a reduction of search space for the recognizer and therefore leads to even better baseline performance.

7.2. 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. Our results, tabulated in Table 1, show that the duration model yields improvement in all three cases. The baseline system already contains standard context-independent phone duration models and all gains reported are over and above those realized from these baseline models.

Duration is more effective with increasing linguistic constraint or lexical knowledge. The reason is that at better baseline performance, the recognizer proposes more sensible words whose corresponding parse trees are more useful to the duration model, enabling it to apply its higher level linguistic knowledge, embedded above the phoneme level. On the other hand, when these added constraints are removed, the recognizer tends to hypothesize nonsensical pseudo-words where, despite correct phoneme sequences, much higher order linguistic information is absent and cannot be utilized as effectively by the duration model.

8. WORD SPOTTING

As a first step towards deomonstrating the benefit of duration to word recognition, we choose to evaluate our model in word spotting using an ANGIE based system described in [3]. In this pilot study, duration is used as a post-processor to disambiguate between two acoustically con-

Linguistic Constraint	Baseline Error (%)	Error w/ 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

Table 1: *Phonetic error rates and relative error reduction (Δ) due to adding duration.*

Duration Model	Error Rate (%)	Δ (%)
None	19	-
w/ Relative Duration	8	58
w/ Absolute Duration	6	68

Table 2: *Results of using duration post-processor to disambiguate "New York" from "Newark" in word spotting.*

fusable keywords, "New York" and "Newark." This pair constitutes the largest source of error in the ATIS task, being very acoustically similar. Duration is potentially a more suitable feature for distinguishing between the two words.

8.1. Pilot Experiment

The wordspotter is assigned to spotting the keyword "New York" from a set of utterances that contain both "New York" and "Newark." These waveforms are passed on to a duration post-processor whose role is to reduce the total number of errors primarily caused by "Newark" false alarms. The duration processor consists of its own ANGIE forced alignment system along with a 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" is detected. Each alignment is then scored for duration which, combined with the acoustics and linguistics, forms the total word score. For each waveform, the post-processor makes a decision between "New York" and "Newark" by choosing the one with the higher total score.

Table 2 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 and 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 1) the greater importance of duration in recognizing whole words and 2) the specific nature of this task where it is known apriori that duration should be advantageous over other features.

8.2. Fully Integrated System

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. These experiments are detailed in [3]. The best performance yielded a 21.5% improvement from 89.3 to 91.6 (FOM), by employing a combination of the absolute and relative duration models. This suggests that duration can be applied in a more general task and still yield substantial gains. Although the relative improvement is not as dramatic as that produced in our pilot study, we speculate that duration is particularly useful for a general keyword spotting task because most keywords contain complex syllable structures and their durational relationships can be utilized by our models efficiently.

9. CONCLUSION

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 utilizing 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 higher linguistic 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 ultimate goal then is to explore the benefit of employing duration in a continuous word recognition task.

10. REFERENCES

[1] W. N. Campbell, "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.

[2] G. Chung, *Hierarchical Duration Modelling for a Speech Recognition System*, S.M. thesis, MIT Department of Electrical Engineering and Computer Science, 1997.

[3] R. Lau and S. Seneff, "Providing Sublexical Constraints For Word Spotting Within The Angie Framework," These proceedings.

[4] D. Pallet, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, A. Martin and M. Przybocki, "1994 Benchmark tests for the ARPA Spoken Language Program," in *Proc. ARPA Spoken Language Systems Technology Workshop*, Austin, TX, Jan. 1995, pp. 5–36.

[5] S. Seneff, R. Lau, and H. Meng, "Angie: A new framework for speech analysis based on morpho-phonological modelling," in *Proc. ICSLP '96*, Philadelphia, PA, vol. 1, pp. 110–113, Oct. 1996. URL http://www.sls.lcs.mit.edu/raylau/icslp96_angie.pdf

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