# TEMPORAL ORGANIZATION OF SPEECH FOR NORMAL AND FAST RATES

*Geetha Krishnan and Wayne Ward*

School of Computer Science, Carnegie Mellon University , Pittsburgh, PA, 15213, U.S.A.

E-mail: geetha@cs.cmu.edu ; whw@cs.cmu.edu

## ABSTRACT

This paper characterizes the temporal organization of speech within interstress intervals (ISI) for fast and normal rates of speech for two groups of English speakers: . normal speakers (NS), and a group of disfluent speakers (DS) whose speech rate was clinically rated as "slow". Duration measurements of stressed vowels, (SV) unstressed vowels (USV) and intervowel intervals (IVI) were obtained for ISIs ranging in length from one to five syllables for normal and fast rates of speech. Despite some differences, both groups were observed to show similar trends in the compression of SVs and IVIs for local (i.e., within ISIs) and global (i.e., fast speech) rate increase. Following the initial identification of these speech components as good predictors of speech rate, we investigated two measures of speech rate based on IVI as they can be computed without the linguistic knowledge of the utterance.

## 1. INTRODUCTION

The temporal dimension of speech provides information on the rate and rhythm of speech. For English speakers, the ISI plays a critical role in maintaining the local rate and the overall rhythm of speech. The ISI is defined as the duration between two major stresses in an utterance. One theory of rhythm claims that speakers of stress-timed languages, such as English, maintain equal duration between major stresses, regardless of the number of unstressed syllables that may be present within an ISI. This theory suggests that in an effort to maintain what is known as stress isochrony speakers will compress the duration of syllables for longer ISIs. In this paper, the term "length"(of ISI) refers to the number of syllables within an ISI and the term "duration" refers to the actual measurement of the interval in ms. Earlier work on rate and rhythm of speech in NS and DS (speakers who had a history of stuttering and their fluent speech was clinically diagnosed as "slow"), has shown no evidence of stress isochrony [3 ]. However, both groups of speakers showed compression of stressed vowels as a function of ISI length, thus providing evidence that ISIs serve as units of temporal planning for English speakers. It was hoped that a detailed characterization of rate control strategies for the group of DS might enable us to understand whether abnormally "slow" speakers differ in their rate control strategies from normal speakers. If normal and slow speakers are found to use certain common strategies for rate control, this information could be utilized in devising measures of speech rate that would be applicable to a wide population of English speakers.

Knowledge concerning rate control strategies is important for automatic speech recognition (ASR) systems as we strive to overcome the problem of increased word error rates for speech rates that are faster or slower than normal. Most ASR systems are trained on normal rates of speech and the word error rates increase if the system encounters speech that deviates from normal rate [4] [6]. If we were to adapt acoustic and language models to compensate for speech rate that deviates from normal, we need to have a robust measure that would provide accurate estimates of rate of speech. It would also be more economical to estimate the rate of speech prior to decoding. Recent studies have shown that durations of certain classes of phonemes are more affected by speaking rate than others [2 ] [7] [1]. Our earlier work [3] suggests that components such as SVs, USVs and the IVIs participate in rate compression to various degrees. This suggests that the duration of certain speech segments may be better indicators of rate than others.

In our first experiment we explore the possibility of identifying speech segments that would behave predictably with changes in speech rate for both groups of speakers. Specifically, we examine the compressibility of elements within ISIs of different lengths for a group of NS and DS for normal and fast rates of speech in an effort to understand how the temporal organization within ISIs may be affected by global changes in rate. The group of DS used in this study had a history of moderate stuttering. The fluent speech of this group was clinically rated as fluent, but "abnormally slow", especially for longer utterances. This group was included in this study to examine whether or not slow speakers differed from normal speakers in the use of strategies for rate control.

In the second experiment, we investigate two measures of speech rate based on IVI as they can be measured without the linguistic knowledge of an utterance.

## 2. EXPERIMENT I

### 2.1 Method

Speech samples were obtained from 20 subjects. Ten of the subjects were NS and ten were DS. They were all native speakers of midwestern American English. The subjects read two passages. One passage was an extract from an article in Scientific American and the other one was a dialogue from a radio play at two speech rates. One was at their normal speaking rate and the other was at a faster rate that was "still comfortable". The intent was to obtain sufficient number of ISIs of different lengths in continuous speech. ISIs of one to five syllable units were obtained from this sample. The ISIs spanned over word boundaries, but not over clause boundaries. Only two levels of stress were considered, primary stress and unstressed. Two linguists marked the stresses in the utterances. Only those utterances that were agreed upon by both experts were used for measurement. The ISIs of different lengths are referred to as $ISI_1$ to $ISI_5$. The numbers in the subscript refer to the number of syllables in the interval. A monosyllabic stress unit, for example, is $ISI_1$ and a two-syllable unit with one stress syllable followed by an unstressed syllable is $ISI_2$ . Likewise, $ISI_3$, $ISI_4$, and $ISI_5$ refer to three, four and five syllable units. Only fluent samples from the speakers were included in the analysis.

For each ISI, the SV, USV durations, the IVIs, and the overall duration of ISIs were measured. Vowel durations were measured from their onsets to their offsets. The IVI was measured as the duration between the offset of one vowel to the onset of the next vowel. The ISI was measured from the onset of the stressed vowel to the onset of the next stressed vowel. Both visual and audio play back were used to aid segmentation. A segmentation algorithm was written in Matlab environment for this purpose.

### 2.2. Results of Experiment I

#### 2.2.1 ISI duration

Figure 1a. illustrates the average syllable duration as a function of ISI length. The time taken per syllable decreases as the length of ISI increases. For the fast speech condition, the trend is essentially the same as for normal rate. The extent of compression is less for the longer units in fast speech. The DS showed compression up to $ISI_3$ for fast speech. But, for $ISI_4$ and $ISI_5$, no significant compression was noted compared to their speech at normal rate.

#### 2.2.2. SV duration

Figure 1b shows SV duration as a function of ISI length for the two groups and the two speech rates. The two groups showed a similar trend of compression with increase in ISI length and for fast speech. It is clear from these observations that the compression of SV within ISIs contribute greatly to local and global rate increase.

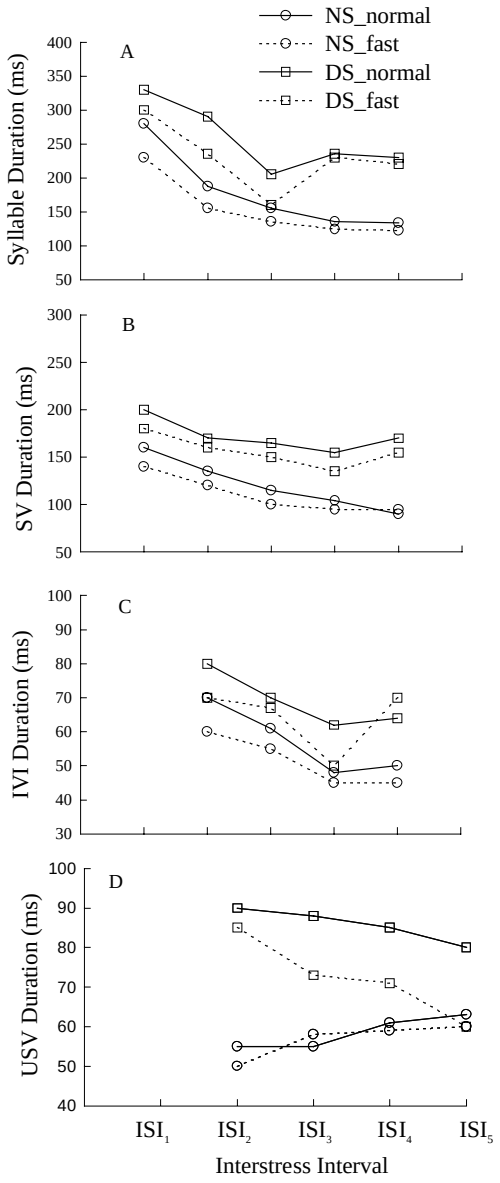


**Fig. 1.** Shows the compression of the different components within ISs for fast and normal rates of speech for the two groups of speakers as a function of ISI length. **A**. Average syllable duration . **B**. Stressed vowel durations **C.** Average intervowel intervals**. D.** Average unstressed vowel durations

**2.2.3. Intervowel Intervals**
Figure 1c. shows the IVI as a function of length of ISI for the two groups and two rates of speech. Similar to SVs, the IVIs also compress as a function of ISI length and for fast speech for both speakers.

**2.2.4. USV durations**
Figure 1d. shows the results for the USV durations. For normal rate of speech the trends for both the groups are similar. The two groups differed significantly for fast speech. Both groups show relatively less compression as a function of ISI length. However, for fast speech, the NS continue to display this trend, whereas, the DS show progressive compression from $ISI_2$ to $ISI_5$

## 2.3 Discussion

The results of experiment 1 showed that the increase in the ISI durations were less than linear as the number of syllables in the ISIs increased. This less than linear increase was achieved by the compression of SV and IVI. The speakers, in general, attempted a faster rate of speech for longer units than for shorter units. For the NS, the duration of the unstressed vowels were relatively unaffected by the length of ISI or by global rate increase, i.e., for fast speech. On the other hand, the DS compressed unstressed vowels when global rate increase was desired. It is possible that the DS can only produce a limited range of speech rates. The "chosen fast rate", combined with the rate increase needed for the longer units, makes it necessary to compress the unstressed vowels to achieve the desired rate. Thus, there appears to be a predetermined order in recruiting speech elements to effect a rate increase.

The two important findings from this experiment are: first, despite some differences between the two groups, both groups of speakers, in general, used similar rate control strategies to effect a local and a global rate increase. For both groups, the longer ISIs were essentially samples of fast speech. Second, for both groups, the SV and the IVI behave predictably during rate changes, whether it is locally or globally driven.

# 3. EXPERIMENT II

## 3.1. Method

From the first experiment it was clear that the ISI of varying lengths can be used to examine the effects of rate changes on segmental factors. Two normal speakers read transcripts from WSJ databases at their normal speaking rate and at a rate faster than normal, but at a pace that was still comfortable. ISIs of two to four syllable lengths were extracted from these samples. Using a combination of fast and normal speech rates and ISIs of varying lengths provided us with a wide range of speech rates. The primary objective of this experiment was to develop a measure for detecting rate of speech of a speech segment based on predictors we identified in the previous experiment. The choice of IVIs is more attractive than SVs for this purpose, as any measure based on IVIs can be computed without the linguistic knowledge of the utterance.

The speech waveforms of utterances were initially segmented into ISIs of varying number of syllables, so that samples of different rates could be obtained. The IVIs and durations of ISIs were measured as described in Experiment 1. The rate of speech was calculated as phones per second for these speech segments. Two measures of IVI were examined. The first one was the ratio of IVI to ISI. That is, the sum of the duration of IVIs divided by the duration of a speech segment. The second measure was the average IVI duration for a given speech segment.

## 3.2 Results and Discussion

Figure 2a. shows the correlation between the first measure (the ratio of IVI to ISI) and phone rate. A poor correlation coefficient of r = 0.41 was obtained. This measure is similar to a measure reported by Samudravijaya et al. [5]. They examined a measure termed mean transition duration which was defined as

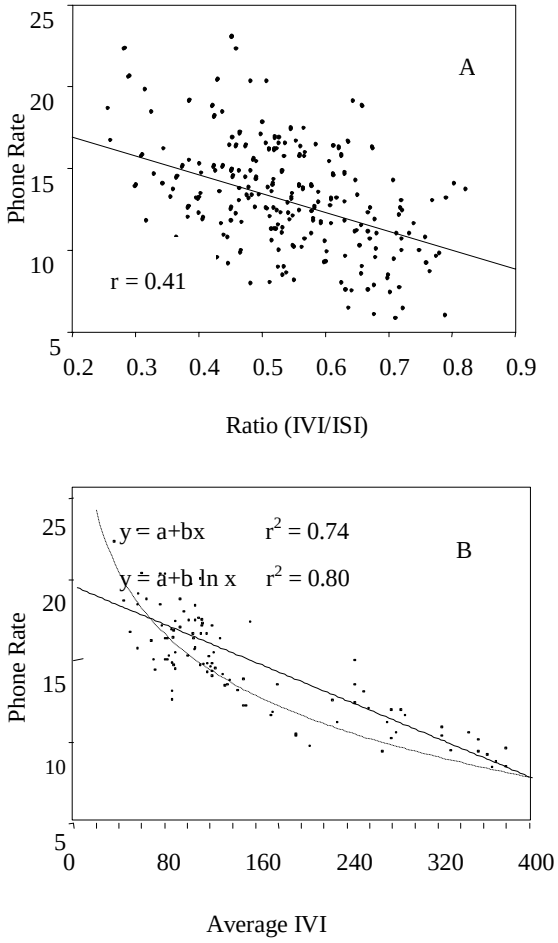


**Fig. 2. A**. Correlation of Phone rate with the ratio measure (IVI/ISI) **B**. Correlation of average IVI with Phone rate

the sum of durations of transitions in an utterance divided by the duration of the utterance. This was a measure based on nonstationary parts of speech. The correlation between their measure and phone rate was 0.42 which is very similar to our results for the ratio measure (IVI/ISI). This follows since IVIs are essentially nonstationary portions of speech between vowels.

Figure 2b. shows the correlation between average IVIs as a function of phone rate. A linear fit shows a significant correlation of r = 0.86 and a nonlinear fit shows a correlation of r = 0.90. These results were obtained on approximately one third of the total data points we had collected. The rest of the data was grouped into data from two, three and four syllable units and were used to examine how well they correlate with the estimated rate. For the estimations, we used the nonlinear function that provided the best fit in figure 2b. Table 1 shows the correlation between the estimated and the observed rates for the two, three and the four syllable units. The estimation part of the experiment resulted in a limited set of data for the two, three and four syllable units. These results are encouraging and suggest that speech rate may be better estimated for utterances that are at least three and four syllables in length than those with fewer syllables.

| Two-Syllable | Three-Syllable | Four -Syllable |
|---|---|---|
| r = 0.70 | r = 0.79 | r = 0.87 |

**Table 1.** Correlation between estimated and observed phone rates are shown for the two, three and four syllable speech segments.

## 4. CONCLUSIONS

In this study we examined rate control strategies used by two groups of speakers, normal speakers and a group of disfluent speakers whose fluent speech was clinically rated as "abnormally slow". The strategies used by the two groups for local (within ISIs) and global (for fast speech ) rate changes were examined. Despite some differences between the two groups, the results showed that the SVs and IVIs compressed systematically when local and global rate increase was desired by both types of speakers. In the second experiment we extended this finding to investigate two measures of speech rate based on IVI, as our intent was to find a segmental feature that could be measured without the linguistic knowledge of the utterance. The first measure, which is a ratio of IVI to the total duration of the speech segment, showed a poor correlation with phone rate. The second measure, which is the average IVI for each speech segment, showed a significant correlation with phone rate. Based on the best-fit curve, we estimated phone rates for two, three and four syllable segments. The correlation between the estimated and the observed rates were significant for segments of all lengths, but was considerably better for the four syllable segments. This preliminary study provides a basis for developing a measure of speech rate that can be implemented on speech waveforms, prior to the process of decoding in an ASR system and would be applicable to a wide population of speakers.

## 5. REFERENCES


1. Bronsted, T. and Madsen, J.P. " Analysis of speaking rate variation in stress-timed languages", *Proc. Eurospeech 1, 1997,* 481-484.
2. Gopal, H.S., "Effects of speaking rates on tense and lax vowel durations", *J. Of Phonetics. 18, 1990*, 497-518, 1990.
3. Krishnan, G. and Colson, K., **"**Temporal Organization of rhythmic units", *120th meeting of the Acoustical Society of America,* Sandiego, 1990.
4. Mirghafori, N, Foster, E., Morgon, N., "Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes". *Proc. Eurospeech '95*. 491-494.
5. Samudravijaya, K., Singh, S.K., Rao , P.V.S. "Pre-recognition measures of speaking rate", *Speech Communication, 24, 1998*, 73-84.
6. Siegler, M.A., and Stern, R.M., "On the effects of speech rate in large vocabulary speech recognition systems",. *Proc. IEEE. Internat. Conf. Acoust. Speech Signal Process, 1995,*. 612-615
7. Van Santen, J.P.H., "Contextual effects on vowel duration", *Speech Communication, 11, 513-546,* 1992.