PAUSE DURATION AND VARIABILITY IN READ TEXTS

Elena Zvonik and Fred Cummins

Department of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
{elena.zvonik,fred.cummins}@ucd.ie

ABSTRACT

Generating natural sounding synthetic speech from text requires a division of a text into IPs and assigning pauses between those phrases. A difficulty which faces attempts to model pauses quantitatively is high degree of variability exhibited by speakers in pause placement and duration. The present study seeks to investigate if Synchronous Speech (speech elicited when two speakers are asked to read a text together) can be used as a mean to reduce inter-speaker variability providing more reliable data for accurate modeling pause durations at IP breaks. We find reduced variability in pause duration when speakers read a text in synchrony. We also find an apparent dependence of pause duration on the length and/or syntactic complexity of the preceding phrase. The reduction in variability when reading synchronously is most evident for the one pause exhibiting markedly longer mean duration.

1. INTRODUCTION

The generation of natural sounding speech from text requires a division of the text into intonational phrases and the assignment of pauses between those phrases. Much work has been done on the prediction of phrase boundaries and some of the existing models show very good results [1, 2]. The generation of appropriate pause durations has received much less attention.

Some timing models dealing with generating segmental durations do not predict pause duration at all [3], while others assign a single standard duration [4].

One model which attempts to generate quantitative predictions for pause durations is the *z*-score model of Barbosa and Bailly [5], developed to generate rhythmically appropriate French speech. The model computes the silent duration that has to be assigned to each rhythmic group in the generation stage. The duration of the silence is predicted by computing the difference between the actual duration of a rhythmic group (inter P-centre group) and the sum of segmental durations of this group and optionally inserting a pause.

A difficulty which faces modeling attempts at the level of the intonational phrase (IP) is the high degree of variability exhibited by speakers in pause placement and duration. Pauses are known to vary depending on many factors such as speech rate, speaking style, discourse.

Thus, Fletcher, defining pauses as being silent intervals of at least 200 ms [6] finds that most speakers vary the number of pauses, but some vary pause length to alter speech rate, especially when speeding up. A series of experiments investigating the role of speaking rate on the level of speech intelligibility conducted by Uchanski et al.[7] showed very interesting results. Manipulation of pause structure—deletion of pauses in clear (slower speaking rate) speech and insertion of additional pauses into conversational sentences (faster speaking rate)—reduced intelligibility scores in both cases. Changes in pause duration had only a weak effect. This suggests that the role of pauses in not merely to provide a listener with processing time. Grossman and Lane [8] looked at the relative contributions of articulation rate and number of pauses to the perceived rate of speech. They found that when a speaker doubles her rate, she reduces the number of pauses, while change in articulation rate has a stronger perceptual effect on listener. Thus, an increase in articulation rate may amplify the effect of pause frequency on apparent rate.

An interesting attempt to relate prosody to syntax in Swedish sentences was made by Strangert [9]. She investigated how pause behavior depends on syntactic structure. In her research she varied both the complexity of NPs and VPs in a sentence and the length of the words immediately preceding the boundary. She found both factors to have significant influence on silent interval duration. Pause duration tended to increase when longer words preceded the boundary. As NP complexity increased, pause duration increased accordingly. Silent interval duration decreased when the NP had the simplest structure and the VP increased in complexity. Only a single speaker participated in the experiment reading a list of sentences, so these results remain to be tested on a larger amount of data.

Gustafson-Capkova and Magyesi [10] reported the effect of speaking style and discourse on pause behavior in Swedish. They mainly compared syntactic and discourse context in which speakers tend to make pauses in professional and non-professional readings and in spontaneous dialogues. Their results showed that in professional readings all the pauses appeared on strong boundaries, in non-professional readings speakers tended to locate pauses at sentence and clause boundaries and in front of the conjunctions, in spontaneous dialogues, however, silent intervals tended to appear at weak boundary positions.

Having analyzed a large corpus of read and spontaneous speech in five languages, Campione and Véronis [11] claim that pause durations also vary across languages. They found the average duration of pauses to be lower in Italian and higher in Spanish. The authors also described a trimodal distribution of pauses, categorizing them as brief (<200 ms), medium (200–1000 ms) and long (>1000 ms). All the data was log-transformed.

All the published studies on silent pause durations demonstrate clearly that pause behavior is highly variable, depending in complex fashion on both speaker and discourse situation. Synchronous Speech (SS) introduced by Cummins [12] may provide a means for reducing variability somewhat. SS is speech elicited when two speakers are asked to read a text together. In [12], pause placement was found to be considerably more predictable in synchronous speech than in unaccompanied readings. No quantitative analysis was done, so it is not clear whether pauses were also less variable in duration. Asynchrony at phrase onsets was found to be approximately 20 ms greater than phrase medially, suggesting that pause duration is not as predictable as the duration of substantial linguistic elements.

Speaking in synchrony with another person can only be achieved if the speakers manage to make their speech predictable for each other. This means agreeing on common temporal patterns, and suggests that speakers will have to exploit their shared knowledge of speech timing in order to achieve synchrony. A musical analogy suggests itself: to play together in an ensemble, musicians have to restrict their idiosyncratic variation while interpreting a score, which makes the resulting sound less variable than the performance of a soloist. Phrasing (phrasal prosody and pauses together) constitute a large part of the assumed shared knowledge among the musicians. Varying pause duration in speech is also an expressive mechanism, akin to the expressive variation of a soloist.

The present study investigated whether synchronous speech can be used to reduce inter-speaker variability in pause duration, thereby providing more reliable data for modeling pauses quantitatively.

2. METHODS

Twenty seven subject pairs participated in the experiment, but data from 22 pairs (44 speakers) is reported here. All were from the area of Dublin, Ireland. Each recording session provided a series of text readings which were made in the following order: one speaker from the pair first read the text alone, then both speakers read the text in synchrony and finally the second speaker read the text alone. Each speaker wore a head-mounted microphone, recordings were made onto the right and left channels of a stereo file. No control for familiarity of speakers within a pair was made.

The texts recorded were several small fables, a limerick, a short poem, and the first paragraph of the Rainbow Text. For this particular study the measurements of the duration of four pauses in the first paragraph of the Rainbow Text were made. The text is reproduced in Table 1. When reading synchronously, pauses were inevitably and exclusively made at sentence ends. Speakers familiarized themselves with the text, and provided a first set of solo-synchronous-solo readings. After six further practice readings each subject-pair was recorded again. There was a slight improvement in synchrony between speakers after practice, but no qualitative change was observed. Initial data analysis failed to reveal any significant differences between readings obtained before and after practice, and both data sets will be treated together in the following.

When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. [Pause 1]
The rainbow is the division of white light into many beautiful colors. [Pause 2]
These take the shape of a long round arch with its path high above, and its two ends apparently beyond the horizon. [Pause 3]
There is according to a legend a boiling pot of gold at one end.
People look, but no one ever finds it. [Pause 4]
When a man looks for something beyond his reach,
his friends say he is looking for a pot of gold at the end of the rainbow.

Table 1. *Text read in the experiment. The four pauses analyzed are indicated.*

3. RESULTS

3.1. Pause placement

Table 1 divides the first paragraph of Rainbow Text into 6 distinct phrases. In the synchronous condition pauses were

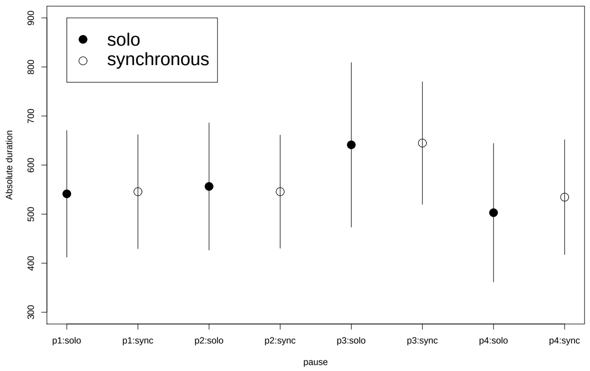

Fig. 1. *Mean and standard deviation of absolute pause duration.*

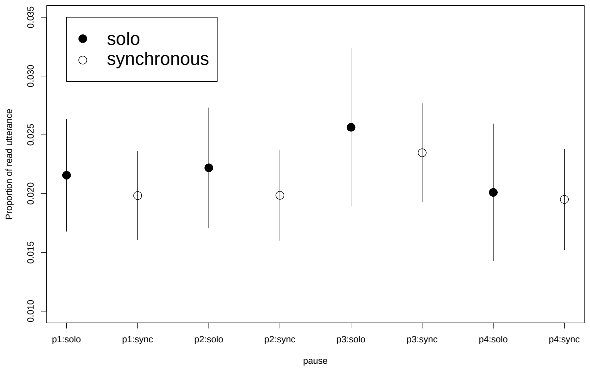

Fig. 2. *Mean and standard deviation of normalized pause durations.*

present at these phrase edges without exception. In earlier experiments on synchronous speech, with four subjects participating, [12] it was found that speakers made additional pauses at other points e.g. at major syntactic edges which were not sentence final. Pauses in non-sentence final positions occurred 48 times in 48 solo readings, and only 4 times in 24 paired readings. The present study based on a larger amount of data replicates the finding that speakers display agreement on pause placement when reading a given text in synchrony. No analysis of pauses in other locations was done in the present study.

3.2. Pause duration

Pause duration for each of 4 pauses was measured (n=704). The location of the pauses is given in Table 1. A two-way ANOVA with fixed factors of condition (solo, synchronous) and pause (pauses 1–4) showed no significant effect of condition: [F(1,696)=0.5, n.s.]. The effect of pause is significant: [F(3,696)=30.3, p<0.01]. Finally, the interaction of the two factors is not significant: [F(3,696)=0.8, n.s.].

Rate variation across readings may cause durational measurements to appear to be more variable than they are. We adopted a crude normalization scheme by expressing each pause duration as a proportion of the duration of the entire containing reading. A two-way ANOVA using normalized durations and with the same fixed factors of condition and pause was done. This time the effect of condition was significant: [F(1,696)=21.3, p<0.01]. The effect of pause remained significant: [F(3,696)=31.8, p<0.01]. The interaction was again insignificant: [F(3,696)=1.1, n.s.].

Post hoc Tukey HSD test on pause data using absolute durations and comparing all possible pairs of means showed that Pause 3 differed significantly from pauses 1, 2, and 4 (p<0.01). No other differences were significant. The third pause is longer in both solo and synchronous conditions. The sentence preceding Pause 3 is the longest and the most complex sentence in the text, while those preceding the other three pauses are both simpler and shorter. The sentence preceding pause four is shorter than all the others, but is associated pause duration is comparable to the first two.

3.3. Pause Variability

Variability across speaking condition was examined using both the absolute and normalized pause duration data. For each pair of pauses, an F-test was done with the directional hypothesis that variability would be reduced in the synchronous condition compared with the solo condition. Table 2 gives the results of the tests. When normalized durations were used, the variability in the synchronous condition was always significantly less than in the solo condition. Using absolute durations, only pause 3 was significantly less variable.

Pause	Absolute	Normalized
1	$F = 1.23$, n.s.	F=1.59, p <0.05
2	$F = 1.26$, n.s.	F=1.76, p <0.01
3	$F = 1.80$, p <0.01.	F=2.57, p <0.01
4	$F = 1.45$, p <0.05.	F=1.85, p <0.01

Table 2. F-test results for difference in variances, with directional hypothesis var(synch)<var(solo). All tests have (87,87) d.o.f.

4. DISCUSSION

The present study sought to demonstrate that synchronous speech (speech elicited when two people are asked to read a given text together) can be used to reduce inter-speaker variability in pause duration, thereby potentially providing more reliable data for modeling pauses quantitatively.

In this experiment the duration of pauses occurring between sentences were measured and analyzed. In order to avoid the possible influence of difference in speaking rate between subject-pairs, analysis of both absolute and normalized durations was done. In both cases a marked difference in pause duration was observed between pause three, which is preceded by a long, syntactically complex sentence, and the remaining pauses, which were preceded by simpler and shorter sentences.

Variability in pause behavior was reduced in the synchronous condition. This was seen most clearly when normalized durations were employed, even though our normalization procedure (division by the duration of the whole reading) was avowedly crude.

The present results expand on findings in Cummins [12] in demonstrating that not only is pause location more consistent in synchronous speech, but pause duration is also less variable. They also add to findings of Strangert [9] that syntactic complexity of a preceding sentence may be positively correlated with pause duration. The nature and strength of this relationship remains to be tested exhaustively, and the results of such an investigation will be of use in the quantitative modeling of pause duration in speech synthesis. Our findings suggest that such an investigation could usefully use synchronous speech as an elicitation device in order to obtain more consistent data.

The fact that the longest pause (pause three) also exhibited by far the greatest reduction in variability in the synchronous condition suggests that additional lengthening which results from greater than average complexity of the preceding sentence may be of indeterminate length. That is, speakers showed a high degree of agreement in pause duration for three of the four pauses in both conditions, and a comparable degree of agreement for pause three only in the synchronous condition, where they are required to maximize the predictability of their speech timing. Neither the timing literature nor our findings allows us yet to hold strong views on the possibility of independent contributions to pause duration, but the possibility of independent contributions with different characteristic properties, including inherent variability, needs to be explored.

In summary, little is known about the determining influences on the length of silent intervals at IP boundaries and no current models accurately predict their durations. A controlled study of the factors determining pause behavior, which elicits IPs of varying length and complexity, would significantly future quantitative modeling efforts. Using synchronous speech as a possible mean to reduce inter-speaker variability would also provide more reliable results.

5. REFERENCES

[1] Wang M. and Hirshberg J., “Predicting intonational boundaries automatically from text. the ATIS domain,” in *Proceedings of DARPA Speech and Natural Language Workshop*, 1992, pp. 378–383.

[2] Sanders E. and Taylor P., “Using statistical models to predict boundaries for speech synthesus,” in *Proceedings of Eurospeech’95*, 1995, pp. 1811–1814.

[3] W. Campbell, *Multi-level Timing in Speech*, Ph.D. thesis, University of Sussex, Sussex, UK, 1992.

[4] Klatt D. H., “The KLATTalk text-tospeech conversion system,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 1982, pp. 1589–1592.

[5] Plínio Almeida Barbosa and Gérard Bailly, “Generation of pauses within the z-score model,” in *Progress in Speech Synthesis*, Jan P. H. Van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, Eds., pp. 365–381. Springer Verlag, New York, 1997.

[6] Janet Fletcher, “Some micro and macro effects of tempo change on timing in French,” *Linguistics*, vol. 25, pp. 951–967, 1987.

[7] Uchanski Rosalie M., Sunkyung S. Choi, Louis D.and Reed Charlotte M. Braida, and Nathaniel I. Durlach, “Speaking clearly for the hard of hearing IV: Further studies of the role of speaking rate,” *Journal of Speech and Hearing Research*, vol. 39, pp. 494–509, 1996.

[8] François Grosjean and Harlan Lane, “Effects of two temporal variables on the listener’s perception of reading rate,” *Journal of Experimental Psychology*, vol. 102, no. 5, pp. 893–896, 1974.

[9] Eva Strangert, “Relating prosody to syntax: boundary signalling in Swedish,” in *Proceedings of the 5th European Conference on Speech Communication and Technology*, 1997, vol. 1, pp. 239–242.

[10] Sofia Gustavson-Čapkova and Beáta Megyesi, “Silence and discourse context in read speech and dialogues in Swedish,” in *Proceedings of Prosody 2002*, 2002, to appear.

[11] Estelle Campione and Jeand Véronis, “A large-scale multilingual study of silent pause duration,” in *Proceedings of Prosody 2002*, 2002, to appear.

[12] Fred Cummins, “On synchronous speech,” *Acoustic Research Letters Online*, vol. 3, no. 1, pp. 7–11, 2002.