ISCA Archive
http://www.isca-speech.org/archive

Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 123-126.

Influence of manipulation of short silent pause duration on speech fluency

Tobias Lövgren & Jan van Doorn

Umeå University, Umeå, Sweden

Abstract

Ordinary speech contains disfluencies in the form of hesitations and repairs. When listeners make global judgements on speech fluency they are influenced by the frequency and nature of the individual disfluencies contained in the speech. The aim of this study was to investigate a single dimension, pause duration, in the perception of speech fluency. The method involved simulation of pause duration within naturally fluent speech by manipulating existing acoustic silences in the speech. Four conditions were created: one for the natural speech and three with step wise increases in acoustic silence durations (average x2, x4 and x7.5 respectively). In a forced choice task listeners were asked to judge the speech samples as fluent or non fluent. The results showed that the percentage of judgements of disfluency increased as the pause durations increased, and that the difference between the unmanipulated speech condition and the two conditions with the longest pause durations were statistically significant. The results were interpreted to indicate that the individual dimension of pause duration has an independent influence on the judgement of fluency in ordinary speech.

1. Introduction

Speech fluency is generally characterised by the presence or absence of disruptions to speech flow [8]. Studies of naturally disfluent speech have identified the types of disfluency that can be found in normal speech [7] and have reported their frequency of occurrence [12]. Those speech disruptions which are generally considered to influence speech fluency include word or part word repetition, phrase repetition or revision, prolongation of sounds, filled pauses (uh, um etc) and silent pauses [7].

Various methodologies have been used to study the influence of individual factors on speech fluency. In particular, digital simulation of speech features opens possibilities for investigation into how factors that have been identified as features of disfluent speech affect listener perception. It enables independent manipulation of factors that could not otherwise be controlled in natural speech. Recent studies have taken advantage of accessible high quality speech technology to simulate disfluent speech. For instance, Amir and Yairi [1] manipulated vowel durations and between-word pauses in part word and full word repetitions in order to investigate features that distinguish stuttered speech from normally disfluent speech.

While the presence of pauses has been identified as a contributor to the perception of disfluent speech, only a subset of all pauses found in natural speech influence the perception of speech fluency. Those pauses which perceptually disrupt the smooth flow of speech, sometimes referred to as hesitation pauses [11] frequently occur at non syntactic boundaries, and have been referred to by Duez [4, 5] as within-constituent pauses i.e. pauses occurring between strongly connected elements. Hesitation pauses can be signaled by a filler such as 'um' (filled pauses), a period of silence (silent pauses) or prepausal lengthening [13].

For silent pauses a duration of about 200 ms silence has been considered as a threshold before the silence is consistently perceived as a pause [17]. Thus, in investigations on pausing it is common to eliminate periods of silence less than 200-300 ms [9]. However, it should be noted that shorter silent periods have also been found to be associated with a perceived pause. For instance intervals as short as 60 ms could be associated with perceived pauses [2], while it has also been reported that in certain acoustic environments (such as prepausal lengthening) a pause could be perceived in the absence of any silent period [6, 13]

Martin & Strange [10] reported that listeners found hesitation pauses (in the form of silent pauses) to be less salient than syntactic pauses, in that they were frequently disregarded in tasks where listeners were specifically asked to identify pauses. However, Duez found that such pauses were not entirely disregarded by listeners because the identification rate for silent within-constituent pauses was higher when the pause duration was longer [5].

Studies on speech fluency have looked at some individual contributing factors to the perception of fluency, but have not specifically looked at pause characteristics. On the other hand studies on pauses have looked at the influence of pause characteristics on the perception of the pauses themselves, but not on the perception of fluency. The purpose of this study was to investigate the link between duration of silent hesitation pauses and the perception of speech fluency.

Specifically, the aim of this study was to investigate the influence of artificial manipulation of the duration of silent intervals that occurred at non syntactic boundaries in ordinary speech on the global perception of speech fluency. The study manipulated duration of existing acoustic speech silences in natural speech. Naïve listeners judged samples of original and manipulated speech as fluent or disfluent in a forced choice task.

2. Method

2.1. Source speech material

The speech material for this experiment was required to have duration of sufficient length for listeners to obtain a global impression of fluency (15s according to Dalton & Hardcastle [3]), and needed to have disfluencies only in the form of natural silent pauses imbedded in it. The planned method involved manipulating the duration of natural silent pauses within the speech rather than the insertion of pauses in order to minimise disruption of other prosodic features in the vicinity of the pause.

Material from professional newsreaders met the specific criteria for the purposes of this study. It has been established that professional newsreading can be considered highly fluent speech, where natural pauses are used at syntactic boundaries for linguistic emphasis [14]. Acoustic silences at non syntactic boundaries e.g. those that formed occlusions in unvoiced word or syllable initial plosives were present in the material to enable simulation of pauses within syntactical units. It was also possible to select material that did not contain other types of disfluency.

Speech material from 10 news programs from Swedish national television was recorded using VHS video recording (hi-fi audio quality with a dynamic range 20 Hz-20 kHz) and reviewed for selection of suitable excerpts. The criteria for selection were that the speech must come from presenters who were in the recording studio and excluded material that could be offensive to the listeners in the experiment (e.g. reports of violent crime).

From this material four speech excerpts were selected. They were each 15-20 s long, read by four different professional newsreaders (two male, two female), and each containing three or four clear silent periods at occlusions of word or syllable initial unvoiced plosives or natural non syntactic silent pauses. The silent periods were spaced at a frequency within the range found for normal disfluencies in ordinary speech [12]. A description of the original speech material, giving number of pauses and distribution of their duration for each sample is shown in Table 1.

Table 1: Description of the original speech excerpts selected for silent pause manipulation.

Sample	# words	# silent pauses	Duration of silent pauses
1	37	4	27,31,51,110 ms
2	42	3	41,53,53 ms
3	43	4	40,46,48,102 ms
4	36	3	44,73,73 ms

2.2. Manipulation of pauses

The audio signals from the VHS video recordings of the selected speech excerpts were recorded directly into the speech analysis program Praat version 4.2. Praat was selected as the program with the most suitable function for duration manipulation.

The boundaries of each acoustic silence were carefully identified within Praat, and in the first instance each silence was stretched x2.5, x5, and x10 respectively to produce three conditions with step wise increase in pause duration. Using the stretch function in Praat allowed the exact acoustic environment of the silence to be preserved, and thus avoided any acoustic disjunction that could be perceived as artificial by listeners. A multiplicative increase in silence duration was chosen in an attempt to maintain the relativity of the pause lengths, and thus improve the naturalness of the speech samples that contained simulated silences.

The simulated material was then reviewed by two independent listeners who were asked to comment on the naturalness of each sample. For the speech samples that were judged as unnatural, those acoustic silences that were perceived as artificial were reduced in absolute duration, yet still maintained step-wise increases from one condition to the next. The revised material was re-judged for naturalness by a further two independent listeners. This process was repeated until there was consensus that the speech did not sound artificial. This process produced a spread of multiplicative factors for each pause-lengthened condition, but for each individual silent pause there remained a step-wise increase from one condition to the next. The actual range of pause durations and corresponding multiplicative factors for each condition can be seen in Table 2. The speech material (in Swedish, with translation to English) showing the location of each manipulated pause, along with the actual pause durations in each condition can be seen in Appendix 1.

Table 2: Description of manipulated pauses

Condition	Mean & range of pause duration (ms)	Mean & range of % increase
C1	57 (27 – 110)	-
C2	117 (65 – 237)	214 (120 - 298)
C3	212 (98 – 479)	394 (149 - 633)
C4	403 (184 – 975)	744 (300 - 1263)

2.3. Listening speech material

The eventual listening material was prepared from the 16 samples (4 samples x 4 silent pause conditions per sample). Specific sequences of these samples were prepared for presentation to the listeners. A Latin squares construction was selected so that any one listener would hear four utterances: one utterance from each speaker in one of the four conditions. Thus 16 different listening sequences were prepared, each consisting of four excerpts. Each sequence was placed on a single track on a listening CD, with a 3 second gap between each of the four samples to allow time for listeners to record their judgments.

2.4. Listeners

Thirty two listeners were used, 14 males and 18 females aged between 20 and 32 years. All were students of medical and allied health programs at Umeå University in Sweden. The listeners had no previous experience in assessing speech, had not studied phonetics, and had no personal experience of stuttering in close friends or family.

2.5. Procedure

Each listener completed a simple questionnaire to establish basic listener information (age, gender, university education). They were then given a verbal description of the concept of speech fluency, and instructions regarding the required task - a forced choice judgement on whether they considered each speech sample to be fluent or not. They were then presented with two practice examples to familiarise them with the nature of the task prior to the commencement of the task. Each person listened to the four samples on a specific track on the CD, using a portable CD player and a set of AKG K130 headphones. All listeners listened from the same portable CD player, using the same headphones, and in the same listening environment. The listeners listened once to each sample and wrote down their spontaneous judgement (fluent/non fluent) at the end of each sample. At the completion of the task the listeners were then asked if they had a close association (family or friends) with stuttering. No listeners reported such association.

2.6. Data analysis

For any one speech sample there were two listener judgements giving eight judgements per pause condition. The total number of non-fluent judgements was counted across speech samples for each condition. A non parametric Kruskall-Wallis test was used to look for overall statistical significance of the pause condition on the number of disfluent judgements. In the case of overall statistical significance from the Kruskall-Wallis test, post hoc testing between individual pause conditions was then planned using Mann-Whitney U tests with Bonferroni correction.

3. Results

3.1. Disfluency judgements

The percentage of judgements of disfluent speech for each condition can be seen in Table 3. There is a steady increase in the percentage of disfluent judgements from the condition C1 (unmanipulated) to Condition C4 (longest durations).

Table 3: Percentage of judgements of disfluent speech for each pause duration condition.

Condition	Mean duration of silent periods	% of disfluent judgements
C1	57 ms	12.5
C2	117 ms	34.4
C3	212 ms	71.9
C4	403 ms	93.8

A Kruskall-Wallis test showed that there was an overall significance effect of pause duration condition H(3) = 12.23, p<0.01. Post hoc testing with Mann-Whitney U tests using BonFerroni correction for multiple testing showed that the differences were significant between Condition C1 (no manipulation) and Conditions C3 and C4 respectively. There was also significant difference in the number of judgments of disfluency between Conditions C2 and C4

The percentage of samples that were judged disfluent within each condition as a function of the mean duration of the silent interval in each condition can be seen in Figure 1.

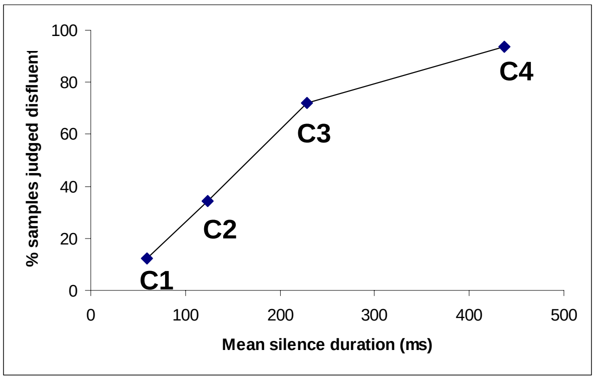

Figure 1: Percentage of speech samples judged disfluent vs. mean duration of silent interval for each of four pause conditions.

The statistically significant increase in the percentage of samples judged to be disfluent in Condition C3 (71.9%) compared with Condition C1 (12.5%) corresponds to an increase in the mean silence duration from 57 to 212 ms.

4. Discussion

In a forced choice judgement of global speech fluency the percentage of disfluent judgements increased when silent pause duration was artificially increased in increments from the original speech condition C1 (with pause durations from 27–110 ms). This increase reached significance for conditions C3 (pause durations 98–479 ms) and C4 (184– 975 ms).

Interpretation of these results must be made within the experimental limitations of the study. Using a design of forced-choice categorical judgements by naïve listeners presents difficulties for those samples where listeners are equivocal about categorising in a dichotomous task. For fluency judgements listeners have been shown to be able to use a continuum scale [1]. However, for naïve listeners the use of a forced choice of presence or absence of a global speech feature is a simple, effective and reliable task in experimental conditions. The design has been able to detect a clear distinction between each condition, and demonstrate a steady increase in the percentage of disfluent judgements from Conditions C1 to C4.

The changes in fluency judgement that occurred in the present study can be attributed to the artificial elongation of existing pauses in the naturally fluent speech of professional newsreaders. It would have been easier to generalise the result if spontaneous speech had been used. However, in order to achieve the aim of isolating the influence of pause duration on speech fluency, it was essential to use material that did not contain any other disfluencies besides silent pauses. The use of newsreading material rather than spontaneous speech needs to be considered in the interpretation of the results. The material was clearly recognizable as excerpts from mass media news, and spoken by well known media personalities. It is thus likely that listeners expected high levels of fluency from these professional readers, and would have judged any hesitations more harshly than they would have if it were another reading situation, and probably also for hesitations in spontaneous speech.

Using artificial manipulation means that it is not possible to be certain that the fluency judgements were not influenced by an artifact of such manipulation. There are two possible sources of artifact - either that the actual process of electronic manipulation of the pauses themselves, or that those manipulations introduced an unnatural distribution of pause durations. Experimentally both those factors were addressed. First, the increase in pause duration was carried out using a stretching function which avoided introducing any electronic boundaries into the acoustic signal. Secondly a multiplicative increase was employed to preserve the relativity of the distribution of pause durations in the samples. Finally, the preparation of the eventual listening material involved naturalness judgements by listeners who were independent of those who participated in the fluency judgements.

An interesting speculation can be made from these results when they are considered in conjunction with findings by Duez [5] for identification rate of within-constituent pauses. The present study found that listeners perceived speech to be disfluent when there were pause durations of 98–479 ms, while Duez found only low rates of pause identification for the same types of pauses (within-constituent) that had durations in the range 250-400 ms. It appears that the global judgement of speech disfluency has occurred for pause durations that were in the range where identification rates of the pauses themselves were only low.

It is thus feasible that this experiment has shown that pause durations required for speech to be judged disfluent are shorter than those when the listening task is to identify the presence of a pause. Such an interpretation would be consistent with the notion that listeners are able to judge global features of speech better than they can perceive individual features for multifactorial features of speech such as voice and speech quality, and speech nasality. Within the speech pathology literature there are numerous reports that show that the reliability of perceptual judgements for individual factors that contribute to a specific speech disorder is found to be inferior to judgements of more global features.

In conclusion, the findings of this experiment demonstrate a relationship between the single factor of pause duration and judgement of global speech fluency when newsreading speech is artificially manipulated. The range of pause duration for which there is a significant increase in the number of disfluency judgements should be further investigated to make a direct comparison with the durations for which such pauses themselves can be identified.

5. Acknowledgements

Technical assistance from Anders Asplund and Thierry Deschamps is gratefully acknowledged. The authors would also like to thank Eva Strangert for her constructive suggestions. Finally, thanks must go to the listeners who took the time to participate in this study.

6. References

[1] Amir, O. & Yairi, E. 2000. The effect of temporal manipulation on the perception of disfluencies as normal or stuttering. *Journal of Communication Disorders,* 35, 63-82.

[2] Campione, E. & Véronis, J. 2002. A large-scale multilingual study of silent pause duration. *SP-2002,* 199-202

[3] Dalton, P. & Hardcastle, W. J. 1977. Disorders of fluency. New York: Elsevier.

[4] Duez, D. 1982. Silent and non-silent pauses in three speech styles. *Language and Speech,* 25, 11-28.

[5] Duez, D. 1985. Perception of silent pauses in continuous speech. *Language and Speech*, 28, 377-389.

[6] Duez, D. 1993. Acoustic correlates of subjective pauses. Journal of Psycholinguistic Research, 22, 21-39.

[7] Guitar, B. 1998. *Stuttering: An integrated approach to its Nature and Treatment.* 2nd edition. Baltimore: Williams and Wilkins.

[8] Hegde, M. N. 1978. Fluency and fluency disorders: their definition, measurement and modification. *Journal of Fluency Disorders,* 3, 51-71.

[9] Kirsner, K., Dunn, J., Hird, K., Parkin, T. & Clark, C. 2002. Time for a pause. *Proceedings of the 9th Australian International Conference on Speech Science and Technology*. Melbourne, December 2-5 2002. Australian Speech Science and Technology Association Inc.

[10] Martin, J. G. & Strange, W. 1968. The perception of hesitation in spontaneous speech. *Perception & Psychophysics,* 3, 427-438.

[11] Ruder, K. F. & Jensen, P. J. Fluent and hesitation pauses as a function of syntactic complexity. *Journal of Speech and Hearing Research,* 15, 49-60.

[12] Searl, J. P., Gabel, R. M. & Fulks J. S. 2002. Speech disfluency in centenarians. *Journal of Fluency Disorders,* 35, 383-392.

[13] Strangert, E. 1990. Perceive pauses, silent intervals, and syntactic boundaries. PHONUM 1, 35-38. University of Umeå: Department of Phonetics.

[14] Strangert, E. 1993. Speaking style and pausing. PHONUM 2, 121-137. University of Umeå: Department of Phonetics.

[15] Susca, M. & Healy, E. C. 2001a. Listeners' perceptions along a fluency-disfluency continuum: a phenomenonological analysis. *Journal of Fluency Disorders*, 27, 135-161.

[16] Susca, M. & Harley, E. C. 2001b. Perceptions of simulated stuttering and fluency. Journal of Speech, Language and Hearing Research, 44, 61-72.

[17] Zellner, B. 1994. Pauses and the temporal structure of speech. In E. Keller (Ed.) *Fundamentals of speech synthesis and speech recognition*. Pp 41-62.

7. Appendix

Excerpt 1.
En upptäckt vid Karolinska Institutet kan få stor betydelse för patienter ***med /27, 65, 121, 247/ typ****-två-diabetes. Ett forskarlag* ***har /51, 120, 197, 349/ kartlagt*** *den insulinproducerande betacellen och har* ***upp /110, 132, 164, 330/ täckt*** *hur* ***insulinproduktionen /31, 67, 98, 184/ kan*** *regleras genom att man tar bort en del av cellen*

English translation.
A discovery at the Karolinska Institute could have great significance for patients with ***[P1]*** *type two diabetes. A research team has* ***[P2]*** *investigated insulin producing beta cells and has dis-* ***[P3]*** *covered how insulin production* ***[P4]*** *can be regulated by removing some of the cells.*

Excerpt 2.
Och först ska vi berätta att den kraftigaste jordbävningen på fem år nu på morgonen skakade stora delar av Taiwan. Jordbävningen mätte sju komma ***noll /53, 105, 184, 340/ på*** *richterskalan vilket är* ***förhållandevis /41, 93, 180, 339/ kraftigt****, men epicentrum låg under havet, elva mil öster om Taiwans* ***öst /53, 114, 237, 450/ kust****.*

English translation.
And first we will report that the strongest earthquake for five years this morning shook extensive regions of Taiwan. The earthquake measured seven point zero ***[P1]*** *on the Richter scale which is relatively* ***[P2]*** *powerful, but the epicentre lay under the ocean eleven thousand kilometres east of Taiwan's east* ***[P3]*** *coast.*

Excerpt 3.
Den svenska säkerhetspolisen ägnar allt mer resurser åt att spana på grupper av islamistiska extremister som man hävdar ***finns /102, 237, 479, 975/ i*** *Sverige. Säpo tror* ***att /48, 112, 241, 444/ grupperna*** *vistas i Sverige för att* ***planera /46, 82, 174, 254/ nya*** *terrordåd och för att rekrytera* ***nya /40, 119, 253, 505/ terrorister*** *bland unga svenska muslimer.*

English translation.
The Swedish security police are allocating further resources to investigate groups of Islamic extremists that are claimed to exist ***[P1]*** *in Sweden. SÄPO believes that the* ***[P2]*** *groups are staying in Sweden to plan* ***[P3]*** *new terror attacks and to recruit new* ***[P4]*** *terrorists from among young Swedish Moslems.*

Excerpt 4.
Brittiska forskare har utvecklat en metod för att förvara vaccin så att det inte behöver ***vara /73, 139, 226, 416/ kallt****. Det här kan* ***leda /73, 169, 287, 559/ till*** *att hanteringen av vaccin blir billigare och att fler barn i* ***u-länderna /44, 82, 133, 251/ kan*** *vaccineras.*

English translation.
British researchers have developed a method of storing vaccine so that it does not need to be ***[P1]*** *cold. This can lead* ***[P2]*** *to handling being cheaper and more children in underdeveloped countries* ***[P3]*** *can be vaccinated.*

.