Modelling Hesitation for Synthesis of Spontaneous Speech

*Rolf Carlson[1], Kjell Gustafson[1,2] and Eva Strangert[3]**

[1]CSC, Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
{rolf;kjellg}@speech.kth.se
[2]Acapela Group Sweden AB, Solna, Sweden
[3]Department of Philosophy and Linguistics, Phonetics, Umeå University, Sweden
strangert@ling.umu.se

*Names in alphabetical order

Abstract

The current work deals with the modelling of one type of disfluency, hesitations. A perceptual experiment using speech synthesis was designed to evaluate two duration features found to be correlates to hesitation, pause duration and final lengthening. A variation of F0 slope before the hesitation was also included. The most important finding is that it is the total duration increase that is the valid cue rather than the contribution by either factor. In addition, our findings lead us to assume an interaction with syntax. The absence of strong effects of the induced F0 variation was unexpected and we consider several possible explanations for this result.

1. Introduction

Human spontaneous speech typically includes disfluencies. These can occur for reasons such as problems in lexical access or in the structuring of utterances or in searching feedback from a listener. The current work deals with the modelling of one type of disfluency, hesitations. The aim is to gain a better understanding of what features contribute to the impression of hesitant speech on a surface level. The work has been carried out through a sequence of experiments using Swedish speech synthesis. One of our long term research goals is to build a synthesis model which is able to produce spontaneous speech including disfluencies. In addition to the general goal to understand and model the features of spontaneous speech, such a model can be explored in spoken dialogue systems to increase the naturalness of the synthesized speech [1] and also to facilitate the signalling of for example uncertainness in a dialogue.

Up till know much less work has been done on the perception than on the production of hesitant speech. A few studies have shown that hesitations (and other types of disfluencies) very often go unnoticed in normal conversation, even during very careful listening, as in labelling tasks [14,15]. We need to know more about what phonetic features contribute to the impression that a speaker is being hesitant. Previous analyses have shown pauses and retardations to be among the acoustic correlates of hesitations [9]. Focused studies on function word behaviour before hesitations showing significant patterns of retardation have been reported [12]. A recent perception study [16] confirms that pause insertion is a salient cue to the impression of hesitation, and the longer the pause, the more certain the impression of hesitance. The intonation has also been studied as an important feature signalling hesitation in addition to possible feedback or turn-taking [8].

With a few exceptions, relatively little effort has so far been spent on research on spontaneous speech synthesis with a focus on disfluencies. The introduction of the VoiceFont [2] pointed to the need to include extralinguistic features of this type in speech synthesis. Methods based on unit selection can naturally include some of the hesitation features present in spontaneous speech but this is mostly by accident. In recent work [20] new steps are taken to predict and realize disfluencies as part of the unit selection in a synthesis system.

In [19] an attempt to synthesize hesitation using parametric synthesis was presented. The current work is a continuation of this effort.

2. Data from analysis and perception

As a background for our effort we use a database developed in the GROG project [7]. The data including prosodic labelling, acoustic and linguistic analyses were selected from an interview given by a female politician, which was originally broadcast on public Swedish radio. A detailed description of the procedure and the database properties is given in [11].

An analysis combining the prosodic labelling, duration data and linguistic feature analysis was undertaken [18]. The analysis revealed a great proportion (20%) of perceived boundaries in syntactically unmotivated positions. A frequent pattern was a perceived boundary after rather than before an initial function word and a slowing down manifested as a lengthening of the function word and even one or two of the preceding words.

A reanalysis of the results revealed that some of the perceived boundaries, and most frequently those following clause initial function words, could be interpreted as instances of hesitation. Thus, for the current study we made a complementary analysis collecting data from three independent listeners who labelled the interview material for perceived hesitation disfluencies. These labellings and the corresponding acoustic data from the database (pause durations and lengthening patterns before perceived hesitations) were used as a basis when designing the synthesis experiment reported on here.

We also used results from a perception experiment designed to study how well a listener can predict an upcoming break [3]. The experiment, based on GROG speech data, revealed a small but significant effect of F0 before the break. There was a correlation between perceived boundary strength and the median F0 value of the last voiced 50 ms of the word before the perceived boundary (r=0,62). Phrase-final F0 slope during the same interval turned out to have a significant effect, too, even if it had less predictive power (r=0,51).

3. Experiment

In the perception experiment presented here, synthetic stimuli were manipulated with respect to two duration features found to be correlates to hesitation, pause duration and final lengthening. A variation of F0 slope before the hesitation was also included. As we hypothesized that syntactic structure might play a role, too, the parameter manipulation was done in two different positions, see below. We refer to this parametric manipulation as "inserting a hesitation". A number of stimuli covering all feature combinations in the two positions were presented to listeners who had to evaluate if and where they perceived a hesitation. The subjects were 16 students of linguistics or logopedics from Umeå University, Sweden. They can be regarded to be naive users of speech synthesis but had experience in perceptual evaluations. The subjects were paid a small amount for their participation.

3.1. Stimuli

A Swedish utterance was synthesized using the KTH formant based synthesis system [6] giving full flexibility for prosodic adjustments, Figure 1. A hesitation was placed either in the first part of the utterance (**F**) or in the middle (**M**):

"I sin **F** trädgård har Bettan **M** tagetes och rosor."

English word-by-word translation: "In her **F** garden has Bettan **M** tagetes and roses." (Tagetes is, like roses, a type of flower.) In addition, there were stimuli without inserted hesitations.

The two positions were chosen to be either inside a phrase (F) or in-between two phrases (M). In designing the experiment, we also took into account the fact that disfluencies are more likely to occur in the first part of an utterance rather than the last [17]. The hesitation points F and M were placed in the unvoiced stop consonant occlusion and were modelled using three parameters: pause duration; retardation; intonation.

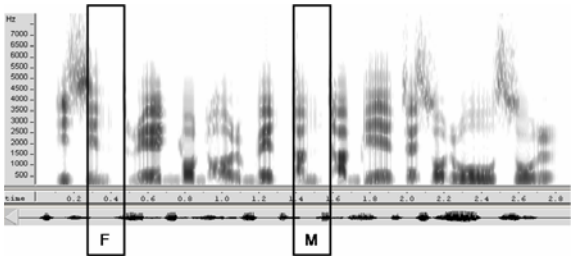

Figure 1. *Default synthesis with the two possible positions for hesitation marked with* **F** *and* **M**.

3.1.1. Duration adjustments

The pausing parameter (Pause) was a simple lengthening of the occlusion in the unvoiced stop. The segment durations in our test stimuli were set according to the default duration rules in the TTS system. The rules have the same basic structure as presented in [5, 13] and have as input: inherent duration (inh); minimum duration (min); and a factor depending on stress and context (P). In the experiments we explored a retardation factor (Ret), which modified segment duration according to the rule below. The retardation factor was applied on the VC sequence /in/ in "sin" and /an/ in "Bettan" before the hesitation points F and M, respectively.

$$\text{Segment duration} = (\text{inh}-\text{min})*P*\text{Ret} + \text{min}$$

The parameter settings were based on preliminary perceptual experiments and made slightly different depending on the position of the disfluency, Table 1. Note that the two tables present the duration extension of the full interval including both retardation and pause insertion.

Table 1: *Duration of* /in + t occlusion/ *(ms) in F position and /an +t occlusion/ (ms) in M position depending on combined pause insertion and retardation.*

F pos	Pause insertion				
Retardation	0	20	40	60	80
	20	40	60	80	100
	40	60	80	100	120
	60	80	100	120	140
	80	100	120	140	160
	100	120	140	160	180

M pos	Pause insertion				
Retardation	0	40	80	120	160
	30	70	110	150	190
	60	100	140	180	220
	90	130	170	210	250
	120	160	200	240	280

3.1.2. Intonation gestures

The intonation was modelled by the default rules in the TTS system. At the hesitation point the F0 was adjusted to model a rising contour (+25 Hz), a flat contour (0) and a falling (-25 Hz) contour, Figure 2. The pivot point before the hesitation was placed at the beginning of the last vowel before the hesitation.

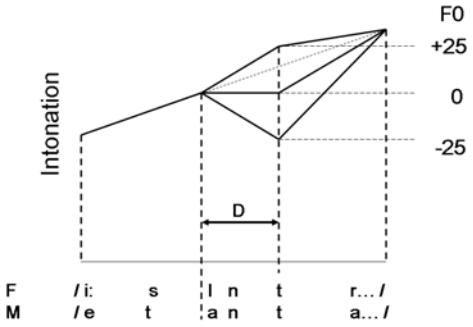

Figure 2. *Intonation contours for the two possible hesitation positions F and M. D=Pause + Ret.*

3.2. Experimental procedure

The subjects were presented with a written description of the experiment and a few examples of the stimuli were played to familiarize them with the synthesis quality. During the test the subjects listened to an individually randomized list of 165 stimuli. The subjects evaluated each stimulus, noting whether they perceived a hesitation, and, if so, where it was positioned, see Figure 3. Each stimulus could be repeated until they were satisfied with their judgement.

Figure 3*: User interface with buttons corresponding to judgment of hesitation, repeat and next stimulus.*

4. Results

The results, presented in Figures 4, 5 and 6, are pooled over the three F0 conditions. In Figures 4 and 5 the two duration features are plotted according to their retardation and pause values. A retardation of 100 ms means that the total duration of VC before the hesitation point is increased by 100 ms. Please note that the parameter values are dependent on the position in the utterance. As expected, the retardation and pause duration are significant features in modelling hesitation. The effect is stronger in the F than in the M position. An 80 ms pause insertion gives 60% hesitation responses in the F position compared to only 42% in the M position. The retardation, too, has a much stronger impact in the F position.

Clearly the results are influenced by both duration parameters (Pause + Ret). Plotting the data according to the total duration change, Table 1, on a logarithmic scale suggests a strong linear dependency, Figure 6. A regression analysis supports this view r = .86 for the F position data and r = .88 for the M position data. The M stimuli, moreover, need a larger duration increase before they are perceived as hesitant.

The F0 variation produced only minor effects which are discussed below.

5. Discussion

The perceptual results are in keeping with the generally accepted understanding that both pause duration and retardation signal hesitation. The most important finding of the present study is that it is the total duration increase that is the valid cue rather than the contribution by either factor. Similar compensatory duration effects were previously observed between individual segment durations in consonant clusters [4, 10].

An interesting observation is that the subjects are less sensitive to modifications in the middle position (M) than in the first position (F). This may be related to the difference in syntactic structure between the two positions: in the F position the hesitation occurs in the middle of a noun phrase ("I sin F trädgård"), whereas in the M position it occurs between two noun phrases, functioning as subject and object respectively. It is possible that the subjects expected some kind of prosodic marking in this latter position and that therefore a greater lengthening was required in order to produce the percept of hesitation. Another possible explanation could be that the duration change is related to a timing unit that would favour the F position.

We considered several alternative units. For example, the default duration of the adjusted segments (in our synthesis rules) are 220 ms for /in + t occlusion/ and 200 ms for /an + t occlusion/, which seems to rule out this unit size as a candidate. Another possible unit is the distance between the preceding stressed vowel (the initial /i:/ in "I" and /e/ in "Bettan", respectively) and the end of the /t/ occlusion in F

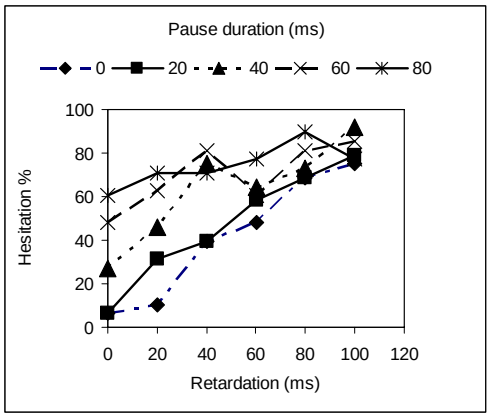

Figure 4*: Perception of hesitation in the first position* **F** *in the utterance separated according to retardation and pause insertion.*

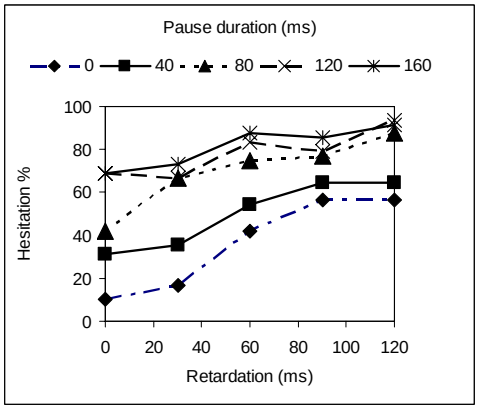

Figure 5*: Perception of hesitation in the middle position* **M** *in the utterance separated according to retardation and pause insertion.*

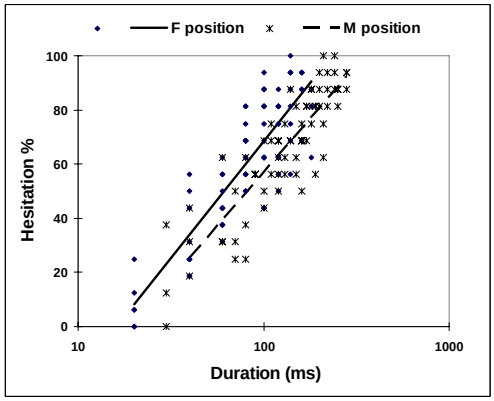

Figure 6*: Perception results plotted according to the total duration increase (Pause + Ret). The regression lines correspond to the two positions F and M.*

and M. The durations here, 380 ms and 400 ms respectively, reveal a very small difference which cannot explain the result. Although we also considered other alternative time units, we favour at present the syntactic explanation, that is, that the

inter-phrasal M position demanded a greater lengthening than the intra-phrasal F position.

In addition to the expected – and observed – effects of the duration features, we had expected to find marked effects of the F0 variation. However, our perceptual data reveal only minor effects and there are no very obvious tendencies, although there are some weak trends. Thus, an F0 fall could in itself (with no contribution of duration cues) sometimes be perceived as a hesitation. This occurred in 20-30% of the total number of responses in both positions. Thus the endpoint (0% perceived hesitation) is only reached when the contour is either flat (0) or rising (+25 Hz). There are also interaction effects of F0 and the other acoustic correlates, in particular in the F position, where the flat and rising contours differ from the falling contour. The flat and rising contours also appear to be responsible (in combination with the duration cues) for the marked jumps in the curves in Figure 4.

We can at present only speculate on the reasons for the absence of F0 effects. It may be that F0 is a very weak cue for hesitation and therefore does not show up in the results. Absence of any significant mean F0 differences between fluent and disfluent function words reported in [12] possibly point in this direction. However, other studies [3] and [8] indicate that F0 might play a role. If so, a possible explanation for the present results could be inadequacies in the F0 modelling: either the variation might have been too small, or the shape of the contours as modelled by the rules of the TTS system was inadequate.

6. Conclusion

Our results indicate clear effects of pausing and retardation in the perception of hesitation disfluencies. In addition, our data revealed a compensatory pattern: it is the total duration increase that counts rather than the contribution by either factor. In addition, our findings lead us to assume an interaction with syntax; our results differed depending on the syntactic position of the hesitation.

The absence of strong effects of the induced F0 variation was unexpected and we have considered several possible explanations for the observations.

In addition to refining our modelling as far as F0 and duration cues are concerned, our future plans include experimental studies of other correlates to hesitation disfluencies, such as for example creaky voice, filled pauses etc. This is part of our long-term goal, to build a synthesis model which is able to produce spontaneous speech including disfluencies.

7. Acknowledgements

We thank Jens Edlund, CTT, for designing the test environment, and Thierry Deschamps, Umeå University, for technical support in performing the experiments. The work was partially carried out at the department of Speech, Music, and Hearing (TMH) and the Centre of Speech Technology (CTT), KTH, Stockholm and partially at Umeå University, Umeå. This work was supported by The Swedish Research Council (VR) and The Swedish Agency for Innovation Systems (VINNOVA).

8. References

[1] Callaway, C. 2003. Do we need deep generation of disfluent dialogue? In: *AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue, Tech. Rep.* SS-03-07. AAAI Press, Menlo Park, CA.

[2] Campbell, N. 1998. Where is the information in speech? in: *Proc. Third ESCA/COCOSDA Workshop on Speech Synthesis*.

[3] Carlson R, Hirschberg J. and Swerts, M. 2005. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication* 46: 326-333.

[4] Carlson, R. and Granström, B. 1975. Perception of segmental duration. In: *Structure and Process in Speech Perception*, Springer Verlag, Berlin, pp. 90-106.

[5] Carlson, R., and Granström, B. 1989. Modeling duration for different text materials. In: *Proc. Eurospeech 89, Paris, September 26-28,* Vol 2 pp. 328-331.

[6] Carlson, R., and Granström, B. 1997. Speech synthesis. In: Hardcastle W. J. and Laver J. *The Handbook of Phonetic Science*. Oxford: Blackwell Publ., pp 768-788.

[7] Carlson, R., Granström, B., Heldner, M., House, D., Megyesi, B., Strangert, E. and Swerts, M. 2002. Boundaries and groupings – the structuring of speech in different communicative situations: A description of the GROG project. *Proc. Fonetik 2002,* 65-68.

[8] Edlund, J. and Heldner, M. 2005. Exploring prosody in interaction control. *Special issue of Phonetica: Progress in Experimental Phonology*, 62 (2-4)

[9] Eklund, R. 2004. *Disfluency in Swedish human-human and human-machine travel booking dialogues.* Dissertation 882, Linköping Studies in Science and Technology.

[10] Fant, G. and Kruckenberg, A. 1989: Preliminaries to the study of Swedish prose reading and reading style, STL-QPSR 2/1989, pp. 1-83.

[11] Heldner, M. and Megyesi, B. 2003. Exploring the prosody-syntax interface in conversations. In: *Proc. 15th ICPhS, Barcelona*, pp. 2501-2504.

[12] Horne, M., Frid, J., Lastow, B., Bruce, G. and Svensson, A. 2003. Hesitation disfluencies in Swedish: Prosodic and segmental correlates. In: *Proc. 15th ICPhS, Barcelona*, pp. 2429-2432.

[13] Klatt, D. H., 1979. Synthesis by rule of segmental durations in English sentences. In: Lindblom, B. and Öhman, S. *Frontiers of Speech Communication Research*, Academic Press, London.

[14] Lickley, Robin J. 1994. *Detecting Disfluency in Spontaneous Speech*. PhD. Thesis, University of Edinburgh.

[15] Lickley, Robin J. 1995. Missing Disfluencies. In: *Proc. ICPhS, Stockholm*, Vol. 4. pp. 192-195.

[16] Lövgren, T. and van Doorn, J. 2005. Influence of manipulation of short silent pause duration on speech fluency. In: *Proc. DISS2005*, pp. 123-126.

[17] Shriberg, E. 1994. *Preliminaries to a theory of speech disfluencies*. PhD thesis, University of California at Berkeley.

[18] Strangert, E. 2004. Speech chunks in conversation: Syntactic and prosodic aspects. In: *Proc. Speech Prosody 2004, Nara*, pp. 305-308.

[19] Strangert, E., and Carlson, R. (forthcoming). On modelling and synthesis of conversational speech. To appear in proceedings of Nordic Prosody IX. Lund.

[20] Sundaram S. and Narayanan S. 2003. An empirical text transformation method for spontaneous speech synthesizers, In: *Proc. Interspeech 2003, Geneva, Switzerland*.