Cues for Hesitation in Speech Synthesis

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 Comparative Literature and Scandinavian Languages, Umeå University, Sweden
strangert@nord.umu.se
[*]Names in alphabetical order

Abstract

The current study investigates acoustic correlates to perceived hesitation based on previous work showing that pause duration and final lengthening both contribute to the perception of hesitation. It is the total duration increase that is the valid cue rather than the contribution by either factor. The present experiment using speech synthesis was designed to evaluate F0 slope and presence vs. absence of creaky voice before the inserted hesitation in addition to durational cues. The manipulations occurred in two syntactic positions, within a phrase and between two phrases, respectively. The results showed that in addition to durational increase, variation of both F0 slope and creaky voice had perceptual effects, although to a much lesser degree. The results have a bearing on efforts to model spontaneous speech including disfluencies, to be explored, for example, in spoken dialogue systems.

Index Terms: hesitation, perception, speech synthesis

1. Introduction

Disfluencies are a recurrent feature in spontaneous speech and occur for reasons such as problems in lexical access or in the structuring of utterances or in searching feedback from a listener. The focus in the current work is one type of disfluency, hesitations.

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. Current work shows that, on the one hand, people's interactions with computers are fundamentally social; users apply "overlearned" social rules to computers such as politeness. Users further appear to prefer synthesized speech with a "prosodic personality" similar to their own [12, 13]. On the other hand, as argued by [1], the naturalness of synthesized speech needs further improvement in order to meet higher demands on human-computer interaction.

With a few exceptions, relatively little effort has so far been spent on research on spontaneous speech synthesis with a focus on disfluencies. 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 [16] new steps are taken to predict and realize disfluencies as part of the unit selection in a synthesis system.

The current study leans on previous work in the area of spontaneous speech with an emphasis on hesitations and more generally on boundaries and boundary signalling (see [3] for an overview). 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.

A background for our effort is a database developed in the GROG project [5, 8] and analyses made on the data collected there [2, 14]. In [15], an attempt to model hesitation using parametric synthesis was presented. This effort was followed by a study exploring the perceptual signalling of hesitation [3]. The manipulations were made in two positions (phrase-internal and phrase external) in a short utterance and were based on observations of the distribution and acoustic manifestations of perceived hesitations in spontaneous speech. This study showed pause duration to be a strong cue to perceived hesitation together with final lengthening. This is a result supported by previous studies showing pauses and retardations to be among the acoustic correlates of hesitations [7, 9] and also that pause insertion is a salient cue to the impression of hesitation [11]. Most important, however, was the result that it was the *total increase*, the combination of both pause duration and final lengthening, that was the valid cue rather than the contribution by either factor.

Variation of F0 slope, which was also investigated, had almost negligible effects. Whether F0 in effect plays a role can be considered an open question in the light of what is known up till now, as there are results going in both directions. [2, 6] point to F0 as a cue for hesitation, while the results in [9] reveal no significant F0 differences between fluent and disfluent contexts. A factor, on the other hand, that appeared to have an influence in [3] was syntactic position; different results were obtained in phrase-internal and phrase-external positions.

In the present study, we aim at a deeper understanding of the acoustic/prosodic basis for perceiving hesitation. Starting out from the findings in [3] as described above, we include new potential cues to hesitation (creaky voice) as well as modifications of earlier parameter settings (temporal parameters and F0 slope variation).

2. Experiment

Synthetic stimuli were manipulated with respect to duration features, F0 slope and presence versus absence of creaky voice. Since the total increase in duration is the most important cue

rather than each factor separately [3], pause and final lengthening were combined in one "total duration increase" feature. The parameter manipulation was done in two different positions as in the previous study. However, in the current experiment the manipulations were similar in the two positions, whereas different parameter settings were used in the previous one.

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 14 students of linguistics or literature from Umeå University, Sweden. They can be regarded to be naive users of speech synthesis. The subjects were paid a small amount for their participation.

2.1. Stimuli

A Swedish utterance was synthesized (the same as in [3]) using the KTH formant based synthesis system [4], giving full flexibility for prosodic adjustments, see 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."[1] In addition, there were stimuli without inserted hesitations.

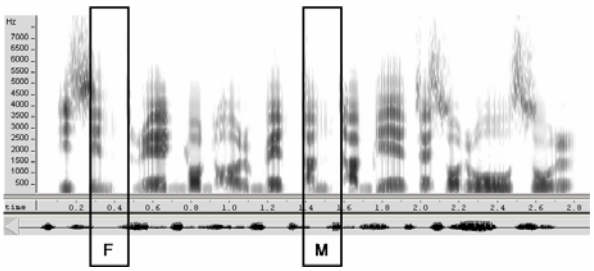

Figure 1. *Default synthesis with the two possible positions for hesitation marked with F and M.*

The two positions were chosen to be either inside a phrase (F) or between two phrases (M). The hesitation points F and M were placed in the unvoiced stop consonant occlusion and were modelled using three parameters: a) total duration increase combining retardation before the hesitation point and pause, b) F0 slope variation and c) presence/absence of creak.

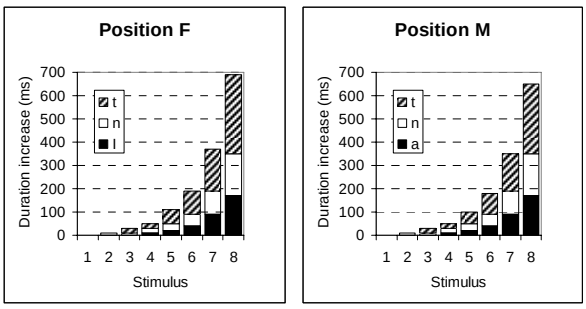

Figure 2. *Total duration increase before hesitation point in phrase-internal (F) and phrase-external (M) position.*

[1] English word-by-word translation: "In her F garden has Bettan M tagetes and roses." (Tagetes is, like roses, a type of flower.)

2.1.1. Retardation and pause

The segment durations in our test stimuli were set according to the default duration rules in the TTS system. The retardation adjustment was applied on the VC sequence /in/ in "sin" and /an/ in "Bettan" before the hesitation points F and M, respectively and the pausing was a simple lengthening of the occlusion in the unvoiced stop. All adjustments were done with an equal retardation and pause contribution. Figure 2 pictures the total duration increase in 7 steps in the two positions. Except for minor deviations, the adjustments were identical for the two positions. (The /t/ segments include pause durations).

2.1.2. F0 slope variation

The intonation was modelled by the default rules in the TTS system. At the hesitation point the F0 was adjusted to model slope variation in 5 shapes with rising contours (+20, +40 Hz) a flat contour (0) and falling contours (-20, -40 Hz) as shown in Figure 3. The pivot point before the hesitation was placed at the beginning of the last vowel before the hesitation. See Figure 4 for an illustration of maximally falling and maximally rising contours.

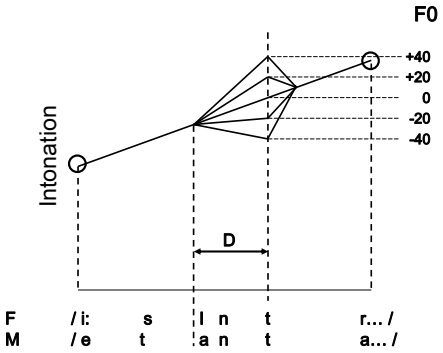

Figure 3. *F0 shapes for the two possible hesitation positions F and M. D=Retardation + Pause.*

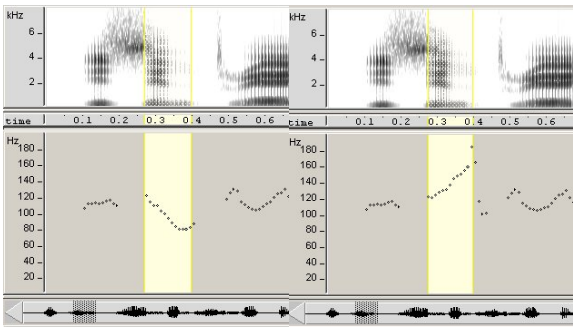

Figure 4. *Illustration of intonation contours for the two extreme cases in position F.*

2.1.3. Creak

Creaky voice was set to start three quarters into the last vowel before the hesitation and to reach full effect at the end of the vowel. The creak was modelled by changing every other glottal pulse in time and amplitude [10], see Figure 5.

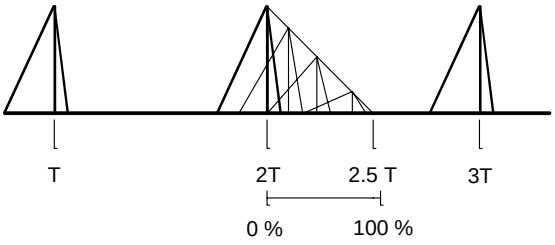

Figure 5. *Modelling presence of creaky voice. Every other glottal pulse is changed in time and amplitude.*

2.2. 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 160 stimuli. The subjects evaluated each stimulus, noting whether they perceived a hesitation, and, if so, where it was positioned. Each stimulus could be repeated until they were satisfied with their judgment.

3. Results

In Figure 6 hesitation perception is plotted as a function of total duration increase (across variation of F0 slope and creak). The strong effect is similar to and confirms the previous result that the combined effect of pause and retardation is a very strong cue to hesitation [3]. The perception of hesitation further depends on syntax, as reflected in the difference between the two curves, a difference of about 25% in the cross-over area. That is, it is easier to detect hesitation in the first (phrase-internal) position than in the second (phrase-external) position.

Similarly, the effects of F0 slope and creaky voice on hesitation detection vary with syntax. Presence of creak, and in particular when combined with a flat or negative F0 slope, tends to make detection easier but only in the first position (within a phrase), see Figure 7.

Finally, the change in hesitation perception due to the presence of creak is plotted against total duration increase (Figure 8a). The two curves represent the phrase-internal and phrase-external positions, respectively. Here, a compensatory pattern is revealed, in particular in the phrase-internal (F) position; in the cross-over area (at, or close to, a total duration increase of about 100 ms, cf. Figure 6), creak has a strengthening effect, making detection easier. In a similar way, F0 had the capacity to compensate for weak duration cues. Falling F0 contours made perception of hesitation easier in the cross-over area for duration (Figure 8b).

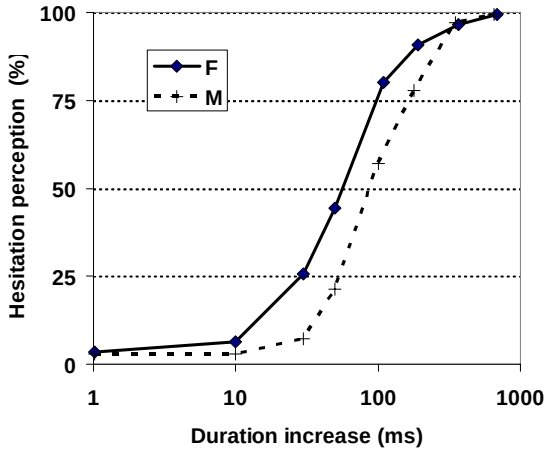

Figure 6. *Detection of hesitation as a function of duration increase across variation of F0 slope and creak. Data separated depending on position of hesitation (phrase-internally=F, phrase-externally=M).*

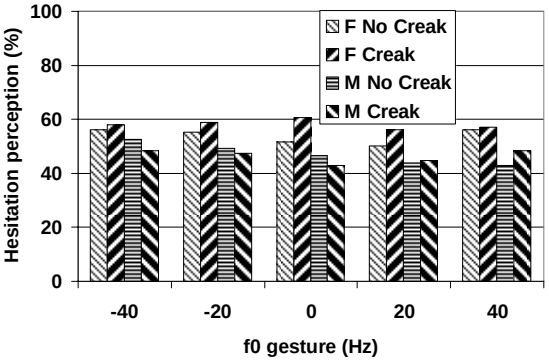

Figure 7. *Detection of hesitation depending on F0 slope and creak phrase-internally (F) and phrase-externally (M).*

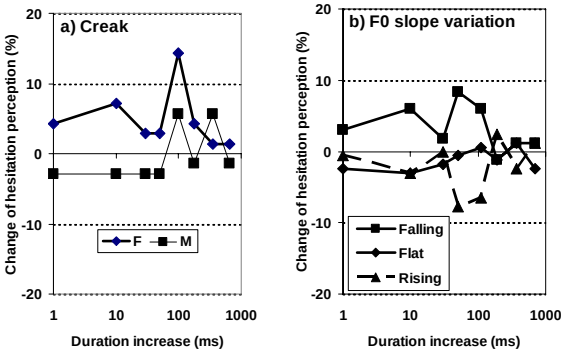

Figure 8. *Change of hesitation perception depending on presence of creak (a) and F0 slope variation (b). Phrase-internal (F) and phrase-external (M) data separated in (a).*

4. Discussion and concluding remarks

The results support the conclusion that duration increase, achieved by the combined effects of retardation and pause, is an extremely powerful cue to perceived hesitation. People apparently have expectations on the temporal structuring of an utterance and react to even modest deviations from this structuring.

F0 slope variation and creak play a role, too, but both are far less powerful, being more of supporting cues. Their greatest effects apparently occur in the cross-over area, when the decision hesitation/no hesitation is the most difficult.

The assumption that subjects are less sensitive to modifications in the middle position (M) than in the first position (F) is also borne out. We relate this to the difference in syntactic structure; 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. A reasonable assumption is that the subjects expected some kind of prosodic marking in the latter position and that therefore a greater lengthening was required in order to produce the percept of hesitation.

This assumption and even more the corollary, that subjects do not expect boundary signalling cues within phrases, get support from how the subjects reacted to the other two features investigated. Both intonation and creaky voice have the capacity to signal an upcoming boundary. Therefore it is not surprising that both intonation (negative F0 slopes) and presence of creak made detection of hesitance easier in the phrase-internal than in the phrase-external position. This dependence on syntax is not unexpected in the light of vast numbers of production studies showing the strength of prosodic signalling to depend on the strength of the syntactic boundary.

In conclusion, our results indicate that the perception of hesitation is strongly influenced by deviations from an expected temporal pattern. In addition, different syntactic conditions have an effect on how much changes in prosodic features like the F0 contour and retardation and the presence of creaky voice contribute to the perception of hesitation. In view of this, the modelling of hesitation in speech technology applications should take account of the supporting roles that F0 and creak can play in achieving a realistic impression of hesitation. In the future also, more global aspects of hesitant speech should be covered. It might be worth while to consider other modifications than those occurring at or before the hesitation point. Also, hesitation is not a binary feature; therefore, a generic model of hesitation should include different degrees of hesitance.

An important step in the modelling of spontaneous speech would be to include predictions of hesitations depending on the utterance structure. To do this, data are required, not only of the realization of hesitant speech, but also of the distribution of hesitations, see e.g. [14]. Our long-term goal is to build a synthesis model which is able to produce spontaneous speech on the basis of such data. An even more long-term goal is to include other kinds of disfluencies as well, and to integrate the model in a conversational dialogue system, cf. [1].

5. 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).

6. 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. AAAI Press, Menlo Park, CA.

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

[3] Carlson, R., Gustafson K. and Strangert E. 2006. Modelling Hesitation for Synthesis of Spontaneous Speech. Proc. Speech Prosody 2006. Dresden, Germany

[4] 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.

[5] 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.

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

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

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

[9] 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.

[10] Klatt, D. and Klatt, L. 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. JASA 87, pp 820-857.

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

[12] Nass, C. and Lee, K. M. 2000. Does computer-generated speech manifest personality? CHI2000, 329-336.

[13] Nass, C. and Moon, Y. 2000. Machines and mindlessness: Social responses to computers. Journal of Social Issues, 60(1):81-103.

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

[15] Strangert, E., and Carlson, R. 2006. On modelling and synthesis of conversational speech. In Proc. Nordic Prosody IX. Lund.

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