would not use what we referred to as gårdstenska in class because his classmates would laugh at him. He refused to name any typical words or features of gårdstenska in class but volunteered to hand in a word list, which only we as researchers were allowed to look at. This student made it clear that "invandriska" was a language he used with his friends outside his class and never in the classroom. Interestingly this was the same class we mentioned above, where students talked about gårdstenska as an identity marker, whereas some students were quite determined in their opinion that this kind of language use was due to a low proficiency in Swedish. Within the other classes the subject seemed less controversial. We can, of course, only speculate about the cause for these differences between the classes. One impression was that there was less controversy about the issue in those classes where more students seemed to identify with speakers of gårdstenska, which were also the more heterogeneous regarding the students' linguistic and cultural background.

3.5 Listeners' awareness of sociolinguistic variation

After visiting the five different school classes in two of Göteborgs multi-lingual areas the overall impression was that a lot of the students showed at least some awareness of sociolinguistic aspects in language use. Some students, as mentioned above, explicitly discussed aspects of language and identity, showing great insight and strong opinions on the issue. Overall most students seemed to acknowledge that gårdstenska is spoken in certain groups (i.e. among friends but not with teachers or parents) and in certain situations and not in others. Thus the listeners showed some awareness of register variation, even though there were different opinions on the question to what extent speakers make a conscious linguistic choice or unconsciously adapt their language when code-switching between gårdstenska and other varieties of Swedish. There was, however, a minority of listeners who categorized what they heard in some of the stimuli as interlanguage of individuals lacking proficiency in Swedish.

4 Future work

The monolingual speakers of SMG support the hypothesis of SMG being a variety of Swedish rather than foreign accent. From discussions with adolescents we have learnt that SMG is primarily used among friends and not with e.g. teachers and parents. Therefore it is interesting that some speakers in the experiment were perceived as speaking SMG (albeit to a lesser degree) even in dialogues with adults. Future work includes investigating if some features of SMG (e.g. the foreign-sounding pronunciation) are kept even in situation where other features (e.g. the SMG vocabulary) are not used, and if these features possibly are kept also later in life when the speakers no longer use a youth language.

Acknowledgements

The research reported in this paper has been financed by the Bank of Sweden Tercentenary Foundation. The authors would like to thank Roger Källström for fruitful discussions on the experiment's design and much appreciated help!

References

Hansson, P. & G. Svensson, 2004. Listening for "Rosengård Swedish". *Proceedings FONETIK 2004*, 24-27.

Lindberg, I., 2006. *Språk och språkbruk bland ungdomar i flerspråkiga storstadsmiljöer 2000–2006.* Institute of Swedish as a Second Language, Göteborg University. http://hum.gu.se/institutioner/svenska-spraket/isa/verk/projekt/pag/pg_forsk2

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 21–24

Prosodic Cues for Hesitation

Rolf Carlson[1], Kjell Gustafson[1,2], and Eva Strangert[3*]
[1]Department of Speech, Music and Hearing, KTH
{rolf|kjellg}@speech.kth.se
[2]Acapela Group Sweden AB, Solna
kjell.gustafson@acapela-group.com
[3]Department of Comparative Literature and Scandinavian Languages, Umeå University
eva.strangert@nord.umu.se
*names in alphabetical order

Abstract

In our efforts to model spontaneous speech for use in, for example, spoken dialogue systems, a series of experiments have been conducted in order to investigate correlates to perceived hesitation. Previous work has shown that it is the total duration increase that is the valid cue rather than the contribution by either of the two factors pause duration and final lengthening. In the present experiment we explored the effects of F0 slope variation and the presence vs. absence of creaky voice in addition to durational cues, using synthetic stimuli. The results showed that variation of both F0 slope and creaky voice did have perceptual effects, but to a much lesser degree than the durational increase.

1 Introduction

Disfluencies of various types are a characteristic feature of human spontaneous speech. 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 aim of the current work is to gain a better understanding of what features contribute to the impression of hesitant speech on a surface level. One of our long term research goals is to build a synthesis model which is able to produce spontaneous speech including disfluencies. Apart from increasing our understanding of the features of spontaneous speech, such a model can be explored in spoken dialogue systems, both to increase the naturalness of the synthesized speech (Callaway, 2003) and as a paralinguistic signalling of for example uncertainness in a dialogue. The current work deals with the modelling of one type of disfluency, hesitations. The work has been carried out through a sequence of experiments using Swedish speech synthesis.

If we are to model hesitations in a realistic way in dialogue systems, we need to know more about what phonetic features contribute to the impression that a speaker is being hesitant. A few studies have shown that hesitations (and other types of disfluencies) very often go unnoticed in normal conversation, even during very careful listening, but scientific studies have in the past concentrated much more on the production than on the perception of hesitant speech. Pauses and retardations have been shown to be among the acoustic correlates of hesitations (Eklund, 2004). Significant patterns of retardation in function words before hesitations have been reported (Horne et al., 2003). A recent perception study (Lövgren & van Doorn, 2005) 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.

With a few exceptions, relatively little effort has so far been spent on research on spontaneous speech synthesis with a focus on disfluencies. In recent work (Sundaram &

Narayanan, 2003) new steps are taken to predict and realize disfluencies as part of the unit selection in a synthesis system. In Strangert & Carlson (2006) an attempt to synthesize hesitation using parametric synthesis was presented. The current work is a continuation of this effort.

2 Experiment

Synthetic versions of a Swedish utterance were presented to listeners who had to evaluate if and where they perceived a hesitation. The subjects, regarded as naive users of speech synthesis, were 14 students of linguistics or literature from Umeå University, Sweden.

The synthetic stimuli were manipulated with respect to duration features, F0 slope and presence vs. absence of creaky voice to invoke the impression of a hesitation. A previous study (Carlson et al., 2006) showed the total increase in duration at the point of hesitation to be the most important cue rather than each of the factors pause and final lengthening separately. Therefore, pause and final lengthening were now combined in one "total duration increase" feature. The parameter manipulation was done in two different sentence 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.

The stimuli were synthesized using the KTH formant based synthesis system (Carlson & Granström, 1997), giving full flexibility for prosodic adjustments. 160 versions of the utterance were created covering all feature combinations in two positions: A hesitation was placed either in the first part (F) or in the middle (M) of the utterance "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.") In addition, there were stimuli without inserted hesitations.

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.

2.1 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 following our earlier results in Carlson et al. (2006).

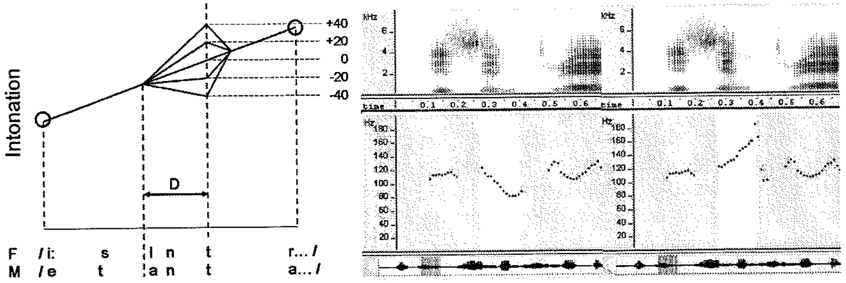

Figure 1. a) F0 shapes for the two possible hesitation positions F and M. D=Retardation + Pause. **b)** Illustration of intonation contours for the two extreme cases in position F.

2.2 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). The pivot point before the hesitation was placed at the beginning of the last vowel before the hesitation, see Figure 1a. Figure 1b shows spectrograms with intonation curves for the two extreme cases in the F position.

2.3 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 (Klatt & Klatt, 1990).

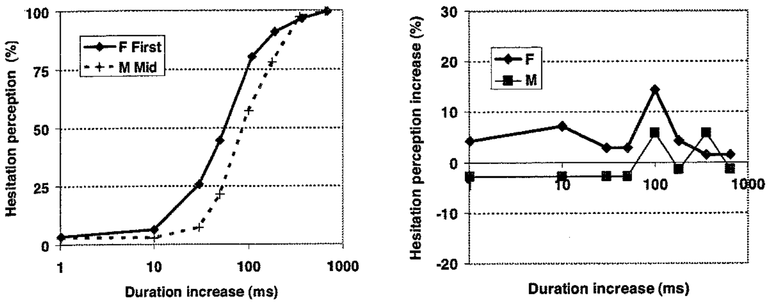

Figure 2. a) Distribution of hesitation responses **b)** Distribution of hesitation perception increase due to addition of creak. Data separated according to position of hesitation.

3 Results and discussion

The results of the experiment are summarized in Figure 2. In Figure 2a, hesitation perception is plotted as a function of total duration increase. 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. In Figure 2b, the increase in hesitation perception due to the addition of creak is plotted against total duration increase. Here, a compensatory pattern is revealed, in particular in the first position; when the duration adjustment is at the categorical border (at a total duration increase of about 100 ms, cf. Figure 2a), creak has a strengthening effect, favouring the perception of a hesitation. In a similar way, falling F0 contours made perception of hesitation easier at the categorical border for duration, compensating for weak duration cues.

These results support the conclusion that duration increase, achieved by the combined effects of retardation and pause, is an extremely powerful cue to perceived hesitation. F0 slope variation and creak play a role, too, but both are far less powerful, functioning as supporting rather than as primary cues. Their greatest effects apparently occur at the categorical border, when the decision hesitation/no hesitation is the most difficult.

The results further indicate that subjects are less sensitive to modifications in the middle position (M) than in the first position (F). 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 is strengthened by the subjects' reaction to the other two features investigated. Both intonation and creaky voice have the capacity to signal an upcoming boundary so that they are more likely to facilitate the detection of a hesitation in a phrase-internal position, where a boundary is unexpected, than between two grammatical phrases. This dependence on syntax is not unexpected: vast numbers of production studies have shown 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.

An important step in the modelling of spontaneous speech would be to include predictions of different degrees of hesitations depending on the utterance structure. To do this, data are required of the distribution of hesitations, see e.g. Strangert (2004). 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. Callaway (2003).

Acknowledgements

We thank Jens Edlund, CTT, for designing the test environment, and Thierry Deschamps, Umeå University, for technical support in performing the experiments. This work was supported by The Swedish Research Council (VR) and The Swedish Agency for Innovation Systems (VINNOVA).

References

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*. Menlo Park, CA: AAAI Press.

Carlson, R. & B. Granström, 1997. Speech synthesis. In W.J. Hardcastle & J. Laver (eds.), *The Handbook of Phonetic Science*. Oxford: Blackwell Publ., 768-788.

Carlson, R., K. Gustafson & E. Strangert, 2006. Modelling Hesitation for Synthesis of Spontaneous Speech. *Proc. Speech Prosody 2006*, Dresden.

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

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

Klatt, D. & L. Klatt, 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. *JASA 87*, 820-857.

Lövgren, T. & J. van Doorn, 2005. Influence of manipulation of short silent pause duration on speech fluency. *Proc. DISS2005*, 123-126.

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

Strangert, E. & R. Carlson, 2006. On modelling and synthesis of conversational speech. *Proc. Nordic Prosody IX, 2004*, Lund, 255-264.

Sundaram, S. & S. Narayanan, 2003. An empirical text transformation method for spontaneous speech synthesizers. *Proc. Interspeech 2003*, Geneva.

Working Papers 52 (2006), 25–28

F-pattern Analysis of Professional Imitations of "hallå" in three Swedish Dialects

Frantz Clermont and Elisabeth Zetterholm
Dept. of Linguistics and Phonetics, Centre for Languages and Literature, Lund University
{frantz.clermont|elisabeth.zetterholm}@ling.lu.se

Abstract

We describe preliminary results of an acoustic-phonetic study of voice imitations, which is ultimately aimed towards developing an explanatory approach to similar-sounding voices. Such voices are readily obtained by way of imitations, which were elicited by asking an adult-male, professional imitator to utter two tokens of the Swedish word "hallå" in a telephone-answering situation and three Swedish dialects (Gothenburg, Stockholm, Skania). Formant-frequency (F1, F2, F3, F4) patterns were measured at several landmarks of the main phonetic segments ('a', 'l', 'å'), and cross-examined using the imitator's token-averaged F-pattern and those obtained by imitation. The final 'å'-segment seems to carry the bulk of differences across imitations, and between the imitator's patterns and those of his imitations. There is however a notable constancy in F1 and F2 from the 'a'-segment nearly to the end of the 'l'-segment, where the imitator seems to have had fewer degrees of articulatory freedom.

1 Introduction

It is an interesting fact but all the same a challenging one in forensic voice identification, that certain voices should sound similar (Rose & Duncan, 1995), even though they originate from different persons with differing vocal-tract structures and speaking habits. It is also a familiar observation (Zetterholm, 2003) that human listeners can associate an imitated voice with the imitated person. However, there are no definite explanations for similar-sounding voices, and thus there is still no definite approach for understanding their confusability. Nor are there any systematic insights into the degree of success that is achievable in trying to identify an imitator's voice from his/her imitations. Some valiant attempts have been made in the past to characterise the effects of disguise on voice identification by human listeners. More recently, there have been some useful efforts to evaluate the robustness of speaker identification systems (Zetterholm et al., 2005). The results are however consistent in that "it is possible to trick both human listeners and a speaker verification system" (Zetterholm et al., 2005: p. 254), and that there are still no clear explanations.

Overall, the knowledge landscape around the issue of similarity of voices appears to be quite sparse, yet this issue is at the core of the problem of voice identification, which has grown pressing in dealing with forensic-phonetic evaluation of legal and security cases. Our ultimate objective, therefore, is to use acoustic, articulatory and perceptual manifestations of imitated voices as pathways for developing a more explanatory approach to similar-sounding voices than available to date.

The present study describes a preliminary step in the acoustic-phonetic analysis of imitations of the word "hallå" in three dialects of Swedish. The formant-frequency patterns obtained are enlightening from a phenomenological and a methodological point of view.