Phonological Reduction in Swedish

Per-Anders Jande

Centre for Speech Technology (CTT)
Department of Speech, Music and Hearing, KTH
Drottning Kristinas väg 31, S-100 44 Stockholm, Sweden
jande@speech.kth.se

ABSTRACT

In this paper, the importance of pronunciation variation modelling is discussed. As a first step in developing a model of Swedish pronunciation variation due to speaking style and speech rate, a tentative reduction rule system has been developed. An assessment experiment testing the impact of phonological reduction, as defined by this system, on the perceived naturalness of speech synthesis was conducted. Canonical and reduced synthetic speech stimuli with three different speech rates were presented to naïve subjects. The reduced pronunciations were significantly more often perceived as more natural than the canonical pronunciations at the higher speech rates, while there was no significant general difference in perceived naturalness depending on reduction level for the lowest rate. The dependence on speech rate for perceived naturalness was significant. A possible cause for some observed differences in perceived naturalness depending on the nature of specific stimuli is discussed.

1 INTRODUCTION

Differences in speaking style influences many aspects of the linguistic and acoustic message. In informal, casual speech, the speaker often assumes that the listener has more background knowledge than in more formal conversations. Informal speech is thus often less explicit. The speaking style also affects the choice of words, the speech rate etc. The amount of phonological and phonetic detail in the realisation of an utterance is also affected, so that fast and informal speech is often more reduced that slow and formal speech. To have knowledge about how pronunciation varies in a language can be beneficial e.g. for improving the accuracy of automatic speech recognition (ASR) systems and for improving the naturalness of speech synthesis and for synthesising different speaking styles.

The ultimate goal of the research, of which the experiment reported in this paper is a part, is to construct a non-application specific model of Swedish pronunciation variation due to speaking style and speech rate. For a general description of speaking style dependent pronunciation variation in Swedish, both phonological and phonetic level rules will have to be developed. For this purpose, data-driven methods will be used on annotated spontaneous speech corpora. The focus will be on general aspects of pronunciation variation, rather than on variation due to dialect or individual factors.

As a starting point, a tentative knowledge-based phonological level rule system has been developed. The purpose of this rule system is to form a base from which a more elaborate rule system can be built. The rule system has some empirical grounds, since it is based partly on empirical results reported by Gårding [1]. The result of applying the rule system to canonical (maximally detailed) transcriptions can be compared to what can be observed in spontaneous speech corpora and the rules updated, if systematic deviations are found. The phonological rules can also serve as a skeleton to which more detailed phonetic rules can be added.

In this paper, an assessment experiment testing the impact on the perceived naturalness of speech synthesis of the tentative rule system is reported. The specific questions asked are: 1) What is the impact on the perceived naturalness of synthesis output of applying the rule system to the input transcriptions, 2) What is the correlation between the perceived naturalness of the rule-processed stimuli and synthesis speech rate and 3) Are there differences in perceived naturalness depending on the rules applied and, if so, how can these differences be explained?

1.1 BACKGROUND

Work on reduction rules for Swedish (focusing on sandhi rules) has been reported by Gårding [1] and Eliasson [2]. Gårding [1] studied consonant clusters and recorded a database consisting of lists of all possible consonant clusters at word boundaries. The lists were read at different rates and reduction rules for describing the fastest rate in relation to the slowest rate were compiled. More recently, Bannert and Czigler [3] published a status report on a project with one of its goals to describe reduction patterns in consonant clusters of spontaneous speech with phonetic detail. The

ISBN 1-876346-48-5

status report contains data from corpus studies, but no developed rule system. Bruce [4] has built a reduction rule system from observations and phonetic knowledge and evaluated it with preliminary empirical studies. This rule system is mainly concerned with vowel and syllable elisions and uses stress patterns and the alternation between strong and weak syllables to predict elision. Bruce's [4] work thus complements e.g. Gårding's [1] work on consonant cluster reduction. There are also studies of reduction in Swedish at the phonetic level. For example, Engstrand [5] has studied the phonetic variation in natural Swedish discourse.

The use of data-driven methods to explore pronunciation variation has been shown to be beneficial in many studies concerning other languages than Swedish, especially for expanding ASR lexica. For example, Kessens, Strik and Cucchiarini [6] used forced recognition to select the most often applied rules from a very general system. Kipp, Wesenick and Schiel [7] describe a data-driven approach that outperforms a knowledge-based approach for German continuous speech segmentation. Nakajima, Hirano, Sagisaka and Shirai [8] derive general and part of speech specific reduction rules using a Japanese speaking style parallel corpus. Pastor and Casacuberta [9] automatically train finite state automata describing different transcription variants of words using Spanish speech data.

Research on ASR pronunciation modelling explicitly addressing speaking style and speech rate has also been reported, e.g. by Fosler-Lussier and Morgan [10], who investigated the relationships between word predictability, speaking rate and pronunciation. It was shown that reduction is greater for highly predictable words and when speech rate is high. Further, Finke and Waibel [11] describe a framework in which the probability of encountering a certain pronunciation variant is a function of speaking style. Zheng, Franco and Stolcke [12] report using parallel, rate specific acoustic models. This approach improved recognition accuracy compared to a model with a collapsed rate acoustic model. This study is an example pronunciation variation modelling on a higher level. The acoustic models used are trained in the usual way, but on different data sets (slow and fast speech, respectively) and it is shown that separating different types of speech can improve recognition.

2 A TENTATIVE RULE SYSTEM

Building partly on the work of Gårding [1] and Eliasson [2], a tentative rule system for reduction of Swedish words has been constructed. The input to the system are phonological lexicon transcriptions corresponding to canonical pronunciations. The tentative rule system thus only concerns phonological level reduction, which means that only reduction phenomena that changes the number of segments or the segment identities in a segment string are considered. Vowel length and word stress and accent are features that are provided in the input transcriptions and reduction involving these features were also included in the phonological level reduction rules. Some vowel and syllable deletion rules similar to those described by Bruce [4] were thus possible to include in the system.

The rules can be divided into ten major categories; rules for 1) Haplology, 2) General forward assimilation, 3) Recursive retroflex assimilation, 4) Backward assimilation, 5) /r/ elision, 6) /h/ elision, 7) common suffixes (rules for specific suffixes or groups of suffixes), 8) common stems (rules for specific stems), 9) Vowel reduction and 10) Syllable elision.

A haplogogy rule deletes the first of two identical consonant clusters at each side of a compound boundary. The forward assimilation rules transfer a phonological feature or a set of features from a segment in a phonological string to the succeeding segment, so that the segments come articulatory closer. Only features that change the phonological identity of a segment, e.g. voicing, place of articulation and manner of articulation, are considered in the system. The forward assimilation rules can result in a merging of two segments to one. Backward assimilation works in the same way as forward assimilation, but with a segment affecting its preceding neighbour instead of its succeeding neighbour. The /r/ elision and /h/ elision rules delete /r/ and /h/, respectively, in certain contexts.

An underlying /r/ followed by an underlying dental will merge into a the retroflex counterpart of the dental in question in central and northern Swedish dialects. This pronunciation is commonly represented in lexicon transcriptions. However, the retroflex feature tends to spread recursively rightwards if followed by any more dentals in the underlying form. The recursive retroflex assimilation rules handle this retroflex feature spreading. The rule system includes a few special rules for common suffixes (e.g. noun inflections) and word stems. Finally, there are some rules that reduce vowels in certain contexts. For example, there are rules that reduces unstressed vowels to schwa in certain contexts and rules that shortens long unstressed vowels. There are also rules that delete certain syllables. The vowel reduction and syllable elision rules use word-internal stress patterns as context.

3 METHOD

A subset of the rules in the system, judged to be especially important, were targeted and the transcriptions from a large full-form lexicon were processed by the rule system. From the set of words whose transcrip-

ISBN 1-876346-48-5

tions were affected by the targeted rules, nine words corresponding to a wide range of rule applications were selected. Each word was placed in a carrier sentence. The sentences were transcribed using the input lexicon, resulting in a set of *canonical* transcriptions. The canonical transcriptions were then processed by the rule system, resulting in a set of *reduced* transcriptions.

Both the canonical and the reduced form of each sentence was synthesised using an experimental version of the Infovox 330 diphone Swedish male MBROLA voice implemented as a plug-in to the WaveSurfer speech tool (cf. Beskow and Sjölander [13]). A diphone synthesiser was chosen over a parametric synthesiser to minimise the effects of the voice quality per se on the perceived naturalness. Further, using a diphone synthesiser, built-in speech rate dependent parametric reduction was avoided and thus only effects of phonological level reduction were investigated. Three different speech rates were produced for each transcription, *low* (the system default rate), *medium* (1.3 times the default rate) and *high* (1.7 times the default rate).

The subjects used in the study were fifteen native speakers of a central or northern Swedish dialect. The subjects were engineering students at KTH with only little experience with speech synthesis and no prior knowledge of the background for the experiment.

The sentences were presented in pairs, with each sentence of a pair having the same target word and the same speech rate, but differing with respect to reduction. The subjects' assignment was to select the most natural sounding sentence of each pair. The altogether 27 pairs (nine sentences at three rates) were presented twice to each subject, resulting at a total of 54 assessments per subject. The pairs were presented in random order, using a new randomisation for each subject.

4 RESULTS AND DISCUSSION

4.1 REDUCTION AND NATURALNESS

Figure 1 shows the distribution of answers from the assessment experiment. It can be seen that the canonical pronunciations are slightly more often perceived as most natural by the subjects when the speech rate is low and that the reduced pronunciations are judged more natural when the speech rate is medium or high. The biases are significant, $p<.01$, for the medium and high speech rate ($\chi^2(1)=40.56$ and $\chi^2(1)=13.06$, respectively). For the low speech rate, there is no clear consensus among the subjects about which is the most natural sounding of the two pronunciations. The slight preference for the canonical pronunciation at the low rate is not significant.

It can thus be seen that the reduced stimuli are generally perceived as more natural for the medium and high rates while there is no general difference in perceived naturalness between the reduced and the canonical stimuli for the low rate.

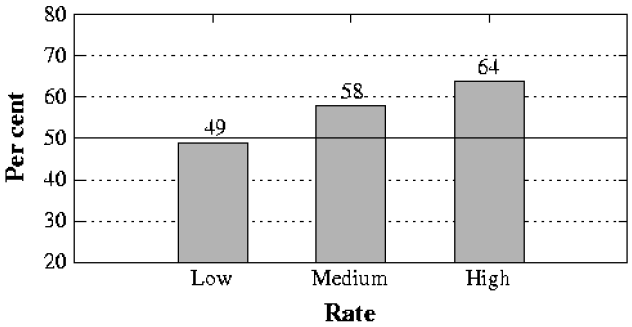

Figure 1: Percentage of pairs for which the reduced pronunciation was judged most natural at different speech rates.

4.2 REDUCTION AND SPEECH RATE

The significance of the observed difference in preferred pronunciation between rates was tested. The result shows that there is a significant dependence on speech rate ($\chi^2(2)=11.61$, $p<.01$) for perceived naturalness. However, the difference between specific rates is significant only when comparing the low and the high speech rate ($\chi^2(1)=11.46$, $p<.01$). The differences between the low and the medium rate and the medium and the high rate, respectively, were not significant.

We can thus conclude that the preference bias in favour of the reduced pronunciations increases when we go from the low to the high speech rate.

4.3 DIFFERENCES BETWEEN STIMULI

Three stimulus sentences broke the general pattern and were most often perceived as more natural in their canonical forms also at the higher rates. For two of the sentences, this reverse pattern was significant. However, there does not seem to be anything generically wrong with the reduced forms not following the pattern. In fact, one of the sentences exactly corresponds to a pronunciation generally used to exemplify reduction in Swedish. This pronunciation is also well established empirically in Swedish spontaneous speech. We are thus not dealing with a generic reduction rule error. But what is then the reason for the deviating pattern of the three stimuli?

Two of these sentences were targets for syllable deletion rules and the third for three backward assimilation rules resulting in merging, i.e. a reduction of the number of segments. All of the sentences were also subjected to vowel reduction rules. In short, these were the most heavily reduced sentences. The rule system had an equivalent impact in terms of the ratio between the number of phonemes in the reduced and in the canonical transcriptions only for one sentence not breaking the pattern. The remaining sentences

ISBN 1-876346-48-5

were more moderately reduced. Further, the target word in the heavily reduced sentence actually following the general pattern, *naturligtvis* (Eng. of course, naturally), is a common 'filler word' in spoken Swedish and checking in a large newspaper corpus, *naturligtvis* shows up approximately a hundred times as often as the heavily reduced target words breaking the pattern. Since the carrier sentence does not provide the subjects with much context information, word predictability is determined almost exclusively by this global frequency of the target word in these sentences.

This seems to implicate that a moderately reduced pronunciation is generally perceived as more natural than the canonical pronunciation, but that heavily reduced pronunciations are only perceived as more natural than their canonical counterparts for highly predictable words. This corresponds well to the results reported by Fosler-Lussier and Morgan [10]. It is reasonable to assume that the rule system produces good pronunciation variants for ASR. However, for speech synthesis simplistic segmental reduction is not enough. Syntactic constraints, word predictability and other types of information must be taken into account. We thus need more types of context and dynamic rules to be able to correctly generate truly natural sounding synthesis.

5 CONCLUSIONS

The results from an assessment experiment with naïve listeners showed that reduced synthetic pronunciations of a set of sentences were significantly more often perceived as more natural than canonical pronunciations at high and medium speech rates, while there was no significant general difference in perceived naturalness depending on reduction level when the speech rate was low. The dependence on speech rate for perceived naturalness was significant; the reduced pronunciations were judged more natural significantly more often at the highest speech rate than at the lowest speech rate. There were some differences in the perceived naturalness depending on the nature of specific stimuli. The cause for these differences may be a mismatch between level of reduction and word predictability. The reported results support the notion that a general reduction rule system can be used to improve perceived naturalness, although the system tested will have to be fine-tuned considering factors such as word predictability and syntactic context.

ACKNOWLEDGEMENTS

The research reported in this paper was carried out at the Centre for Speech Technology, a competence centre at KTH (Royal Institute of Technology), Stockholm.

REFERENCES

[1] E. Gårding, "Sandhiregler för svenska konsonanter (sandhi rules for Swedish consonants)," in *Svenskans beskrivning 8*, C. Platzack, Ed., pp. 97–106. Lund: Department of Nordic Languages, Lund University, 1974.

[2] S. Eliasson, "Sandhi in peninsular Scandinavian," in *Sandhi phenomena in the languages of Europe*, H. Andersen, Ed., pp. 271–300. Berlin: Mouton de Gruyter, 1986.

[3] R. Bannert and P. E. Czigler, *Variations in consonant clusters in standard Swedish*, Phonum 7, Reports in Phonetics. Umeå: Umeå University, 1999.

[4] G. Bruce, "Elliptical phonology," in *Papers from the Ninth Scandinavian Conference on Linguistics*, Ö. Dahl, Ed., pp. 86–95. Stockholm: Stockholm University, 1986.

[5] O. Engstrand, "Systematicity of phonetic variation in natural discourse," *Speech Communication*, vol. 11, pp. 337–346, 1992.

[6] J. M. Kessens, H. Strik and C. Cucchiarini, "A bottom-up method for obtaining information about pronunciation variation," in *Proceedings of ICSLP 2000*, 2000, pp. 274–277.

[7] A. Kipp, M.-B. Wesenick and F. Schiel, "Pronunciation modelling applied to automatic segmentation of spontaneous speech," in *Proceedings of Eurospeech'97*, 1997, pp. 1023–1026.

[8] H. Nakajima, I. Hirano, Y. Sagisaka and K. Shirai, "Pronunciation variant analysis using speaking style parallel corpus," in *Proceedings of Eurospeech 2001*, 2001, pp. 65–68.

[9] M. Pastor and F. Casacuberta, "Automatic learning of finite state automata for pronunciation modeling," in *Proceedings of Eurospeech 2001*, 2001.

[10] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word frequency on pronunciations in conversational speech," *Speech Communication*, vol. 29, pp. 137–158, 1999.

[11] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Proceedings of Eurospeech 1997*, 1997, pp. 2379–2382.

[12] J. Zheng, H. Franco and A. Stolcke, "Rate-dependent acoustic modeling for large vocabulary conversational speech recognition," in *Proceedings of the 2000 Speech Transcription Workshop*, 2000.

[13] J. Beskow and K. Sjölander, "WaveSurfer - a public domain speech tool," in *Proceedings of ICSLP 2000*, 2000, pp. 464–467.