Perceptual evaluation of rule-generated intonation contours for German interrogatives

Bernd Möbius
Institut für Kommunikationsforschung und Phonetik, Universität Bonn
Poppelsdorfer Allee 47, D-53115 Bonn, Germany

ABSTRACT

This paper presents the results of a perceptual experiment that was carried out in order to evaluate naturalness and adequacy of artificial, rule-generated intonation contours for German interrogatives. In previously reported experiments, the ratings for artificial intonation patterns of declaratives were satisfactory while those for interrogatives were not. Therefore, the present study aimed at improving the rules for echo, yes/no, and wh-questions. The integration of several additional features contributed to a more adequate generation of German interrogative intonation contours.

INTRODUCTION

In this paper, the results of a perceptual experiment are presented that was carried out in order to evaluate naturalness and adequacy of artificial, rule-generated intonation contours for German interrogatives. In previously reported perceptual experiments (Möbius and Pätzold 1992), the acceptability of rule-generated intonation patterns as well as the adequate modelling of prosodic properties were critically examined by expert and 'naive' listeners. The results suggested that rule-generated F_0 contours for declaratives were nearly as acceptable as close approximations of the original contours, while the ratings for interrogatives were significantly lower. Furthermore, detailed judgements concerning the realization of word accents and sentence mode were obtained.

The study presented here aimed at improving the rules for three types of interrogative sentences (echo, yes/no, and wh-questions) by extending the speech materials and by taking into account linguistic factors that were omitted in the previous investigations.

GENERATING F_0 CONTOURS BY RULE

General procedure

The rules that generate an artificial intonation pattern for a given utterance are based on the analysis of naturally produced F_0 contours by means of the quantitative model proposed by Fujisaki (1983, 1988). The model aims at a functional representation of the production of F_0 contours by a human speaker and has been successfully adapted to German (Möbius et al., 1991; Möbius, 1993). Using an analysis-by-synthesis procedure, the complex F_0 contour of a given utterance can be decomposed into the components of the model. This is achieved by successively optimizing the parameter values which leads to a close approximation of the original F_0 course. Thus, the model provides a parametric representation of intonation contours.

The potential sources of variation of the parameter values were explored using statistical methods. Standard values were derived on the basis of the statistically significant factors. A set of rules was formulated that control the adjustment of the parameters (see Möbius 1993 for details). The rules capture speaker-dependent as well as linguistic features such as sentence mode, sentence accent, phrase boundary signals, and word accent, and generate an artificial intonation pattern for any given target utterance. The input information needed for generating an F_0 contour by rule is the temporal position of accented syllables. At present, the rules are confined to rather short isolated utterances containing not more than two prosodic phrases.

Modifications

Taking the results of the previous experiments (Möbius and Pätzold 1992) as a starting-point, the major purpose of the study presented here was to improve the rules for generating intonation contours for interrogative sentences. This was achieved by extending the speech materials and by integrating prosodic features that were obviously omitted in the previous investigations. For instance, accentuation and signalling of interrogative sentence mode superposed and compressed on the utterance-final syllable was not sufficiently modelled by earlier versions of the rules. Furthermore, their were no rules for deaccentuation in compounds and in the case of two adjacent accented syllables. These features are now incorporated into the set of rules.

The speech material under investigation may be characterized as typical 'laboratory' speech. It covers three types of interrogative sentences, i.e. echo questions, yes/no questions, and wh-questions. Recordings were made in an anechoic chamber for two female and two male speakers who read the orthographically presented sentences aloud.

PERCEPTUAL EVALUATION

Method

The aim of the perceptual experiment was to evaluate the acceptability of rule-generated contours for German interrogatives in comparison with the acceptability of their original counterparts. More specifically, the subjects were asked to judge the melodic and stress features of the stimuli with respect to 'naturalness' and linguistic 'adequacy'. As an illustration of the two criteria, the listeners were given the following examples: a) Speech melody of an utterance may sound natural although sentence mode may not be clearly or adequately signalled; b) accented syllables may stand out clearly in the course of the utterance (linguistically 'adequate') but possibly by means that sound unnatural. For each of the two criteria, the subjects expressed their ratings on a seven-point scale ranging from -3 to +3. In individual sessions, the stimuli were presented by headphone to 19 prosodically 'naive' listeners who were payed for participating in the experiment.

Since there are several speaker dependent features that are covered by the rules, one female and one male 'voice' were used in the experiment. For each speaker, two versions of each test sentence were presented to the listeners, i.e. one version with a rule-generated intonation contour and one version with the original F_0 information as extracted by a pitch determination algorithm. In order to avoid, as far as possible, any differences in the sound quality of the stimuli that may affect listeners' judgements of prosodic properties, both kinds of stimuli were manipulated by means of the PSOLA algorithm (Moulines and Charpentier 1990).

Results

Not very surprisingly, there is a marked relationship between the ratings for the criteria 'naturalness' and 'adequacy' as expressed by the high value of the Pearson product-moment correlation coefficient ($r = 0.81$; $p < 0.001$). The value of the coefficient of determination ($r^2 = 0.66$) indicates that both variables have a considerable proportion (66%) of their respective variances in common.

The subjective ratings for 'naturalness' and 'adequacy' of the original and the rule-generated intonation contours, both for the female and the male voice, are shown in figures 1 and 2, respectively. The results are given for all utterances irrespective of sentence mode.

The ratings for the original contours are significantly higher than those for the rule-generated ones ($F_{1, 82} = 67.6$, $p < 0.001$ for 'naturalness'; $F_{1, 82} = 24.7$, $p < 0.001$ for 'adequacy'). There is an overall difference of about one point on the scale between the two types of stimuli for both 'naturalness' and 'adequacy'. In general, the female voice is rated higher than the male one, but this difference is statistically insignificant, and the relation between the two types of stimuli is consistent irrespective of speaker.

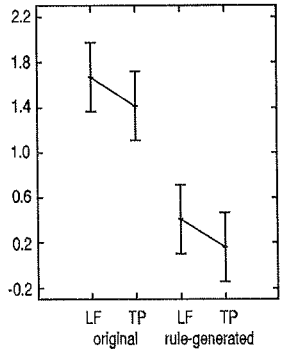

Figure 1. *Mean ratings and 95% confidence intervals for 'naturalness' of original and rule-generated intonation contours for the female (LF) and the male voice (TP).*

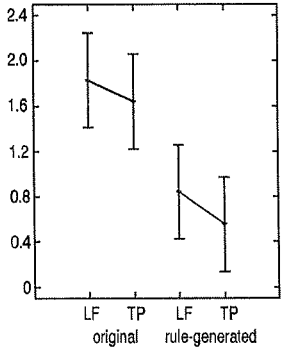

Figure 2. *Mean ratings and 95% confidence intervals for 'adequacy' of original and rule-generated intonation contours for the female (LF) and the male voice (TP).*

Discussion

The version of the intonation contour, i.e. original *vs.* rule-generated, seems to be the only factor exerting a significant influence upon the variance of the 'naturalness' and 'adequacy' judgements. Other potential factors, such as speaker or sentence mode, are statistically insignificant. The mean ratings for 'naturalness' as well as for 'adequacy' of the rule-generated contours are about one point of the seven-point scale lower than those for the original contours. This difference turns out to be statistically significant and consistent for both speakers and the three different interrogative modes.

Higher ratings for the original contours compared to the rule-generated ones meet the expectations if the procedure of developping the rules is considered. While it is true that the rules generate F_0 contours that are, in a qualitative way, representative for certain speakers or types of speakers, the standardized parameter values gained by statistical analysis can only be interpreted as averages. They generate 'prototypical' F_0 contours that will never be produced by any speaker in exactly the same way. Nevertheless, the rule-generated intonation contours seem to be not much less acceptable than the respective original ones.

CONCLUSION

The purpose of the study presented here was to improve the rules for generating intonation contours for interrogative sentences. This was mainly achieved by extending the speech materials and by providing rules for deaccentuation in compounds and in the case of two adjacent accented syllables. Furthermore, the combined effect of accentuation and signalling of sentence mode superposed in utterance-final position is now sufficiently modelled. The results of the perceptual experiment suggest that the integration of these features contributed to a more adequate generation of German interrogative intonation contours.

ACKNOWLEDGEMENTS

The author wishes to thank Ansgar Rinscheid for his pitch mark algorithm, Thomas Portele for implementation of the PSOLA algorithm and for assistance with the perceptual experiment, and Matthias Pätzold for his approximation algorithm. This study was partly supported by a grant from the German Federal Ministry of Research and Technology (BMFT).

REFERENCES

H. Fujisaki (1983), "Dynamic characteristics of voice fundamental frequency in speech and singing", in *The production of speech*, ed. by P.F. MacNeilage (Springer, New York), pp. 39-55.

H. Fujisaki (1988), "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour", in *Vocal physiology: voice production, mechanisms and functions*, ed. by O. Fujimura (Raven, New York), pp. 347-355.

B. Möbius, G. Demenko and M. Pätzold (1991), "Parametric description of German fundamental frequency contours", *Proc. 12th Internat. Cong. Phon. Sc., Aix-en-Provence, 19-24 August 1991*, Vol. 5, pp. 222-225.

B. Möbius and M. Pätzold (1992), "F_0 synthesis based on a quantitative model of German intonation", *Proc. Internat. Conf. Spoken Language Processing, Banff, Alberta, Canada, 12-16 October 1992*, Vol. 1, pp. 361-364.

B. Möbius (1993), *Ein quantitatives Modell der deutschen Intonation - Analyse und Synthese von Grundfrequenzverläufen* (Niemeyer, Tübingen).

E. Moulines and F. Charpentier (1990), "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Communication*, Vol. 9, pp. 453-467.