

PHONETIC REALISATION OF PROSODIC BOUNDARIES IN SYNTHETIC SPEECH

Angelien Sanderman

Institute for Perception Research / IPO, Eindhoven, The Netherlands

ABSTRACT

In this research different sets of prosodic boundary rules were developed and their acceptability was evaluated. The results show that the rule set distinguishing 5 levels of boundary strength produced synthetic speech that is as acceptable as its natural counterpart.

INTRODUCTION

Although synthetic speech is often quite intelligible, it sounds unnatural and is not always easy to comprehend [3]. To improve the quality of synthetic speech it is important to provide it with appropriate prosody. The research reported on here is concerned with the demarcative function of prosody at the sentence level: it studies how prosody is used to group words into phrases.

From previous research [4, 2] we know that speakers use pause, melodic marker and declination reset systematically to indicate various degrees of prosodic boundary strength between words. The perceptual relevance of these features is evident from the fact that listeners systematically assign different perceptual boundary strength-values (PBS) on a 10-point scale to word boundaries differing in such prosodic characteristics. They can even do so when the speech is made unintelligible, so that they can not rely on syntactic or semantic information. In other words, listeners use the phonetic cues used by the speaker to determine the degree of disjuncture in the flow of speech.

The results of previous experiments were used to develop boundary rules for

synthetic speech to generate well-phrased utterances. These rules describe how to realize phonetically the prosodic boundaries of three different strengths (zero, minor and major). The results of this pilot experiment showed that listeners had a significant preference for the realizations *with* the boundary rules over the realizations *without* the boundary rules [5].

The decision to distinguish three levels of boundary strength was made on the basis of practical considerations and was inspired by ideas derived from prosodic phonology [1]. In this research we want to explore whether distinguishing more than three boundary levels will lead to a further improvement of synthetic speech quality.

METHOD

Rules








To develop boundary rules, we used the results of previous experiments. From these results we know that pause, melodic marker and declination reset are important prosodic cues in the phrasing of utterances. From these studies, we can predict with a 90 % success rate what the PBS-value on the 10-point scale will be, given the prosodic cues. Conversely, when we know the PBS we know what the prosodic cues will be. This outcome was taken as the basis for developing 8 different sets of boundary marking rules, four of which are described here.

First of all, there was a version without boundary rules. This means that all the boundaries were realised with neither a pause, a declination reset nor

melodic marker.

Secondly, a set of boundary rules distinguishing three levels of boundary strength was implemented. These three levels of boundary strength do not map onto the PBS, which are expressed on a 10-point scale. Therefore, the observed PBS values were clustered into three classes, which resulted in the recipes given in Table 1. The rule applying to parenthetical clauses and to 30 % of the nonrestrictive clauses is not mentioned in the table. It is as follows: there is a downward reset, with a pause shorter than 200 ms and mostly a pitch contour of the type rise-fall-rise.

Table 1: Recipes for three levels of boundary strength (1 is the weakest boundary and 3 the strongest). The probability of occurrence of each option is given between brackets.
















	Contour type	Pause (ms)	Reset
1	no marker (1.0)	0 (1.0)	no(1.0)
2	no marker (.05)	0 (.45)	no(1.0)
	 (.07)	50 (.55)	
	 (.12)		
	 (.32)		
3	 (.09)	150 (.16)	no(.23)
	 (.10)	250 (.44)	reset
	 (.16)	350 (.27)	(.77)
	 (.65)	450 (.13)	

Thirdly, a set of boundary rules with five levels was developed. To determine the prosodic realisation it was again necessary to cluster the possibilities of the 10-point scale, this time into 5 classes. This resulted in the recipes

given in Table 2. The rule for parenthetical clauses and nonrestrictive clauses is the same as in the set of rules with three levels.

Finally, to assess the relative success of the different sets of boundary rules, a natural version was included: the pause durations and pitch patterns used by a professional speaker were copied onto the synthesized speech.

Table 2: Recipes for five levels of boundary strength (1 is the weakest and 5 is the strongest). The probability of occurrence of each option is given between brackets.

	Contour type	Pause (ms)	Reset
1	no marker (1.0)	0 (1.0)	no (1.0)
2	 (.12)	0 (1.0)	no (1.0)
	 (.19)		
	 (.34)		
	 (.35)		
3	no marker (.09)	50 (1.0)	no (1.0)
	 (.04)		
	 (.06)		
	 (.29)		
4	 (.52)		
	 (.10)	150 (.18)	no (.24)
	 (.10)	250 (.50)	reset
	 (.18)	350 (.32)	(.76)
5	 (.68)		
	 (.07)	450 (1.0)	no (.20)
	 (.07)		reset
	 (.86)		(.80)

Material and Procedure

12 Dutch sentences, ranging in length between 6 and 26 words and varying in syntactic structure, were synthesized [7]. To determine the place of accents and the position and strength of boundaries, a professional speaker was asked to read out the sentences and listeners were asked to assign boundary strengths between all pairs of words. Then, the four sets of prosodic rules were applied. It can be seen from Table 1 and 2 that there are several possibilities to realize a boundary of a certain strength. In those cases a weighted random choice was made from the set of possibilities. Therefore, the rules for three and five levels were implemented three times each on every sentence to see if different random choices from the possible prosodic realizations would affect the results (in fact, they did not). This resulted in a total of $12 \times 8 = 96$ stimuli. The total set contained actually 156 stimuli, since there were more sets of rules, not described here.

The 96 sentences were synthesized by means of diphone concatenation [7]. The phonetic realizations of the various contours, such as the size and slope of pitch movements, slope of the declination line, start and end frequencies, were based on rules given in Terken [6].

In a perceptual experiment, the 156 stimuli were presented to 18 untrained Dutch listeners, who scored the acceptability on a 10-point scale, in 3 sessions. Each session contained 2 blocks and each block contained 2 sentences with its different realisations. The combinations of sentences, the combination of blocks and the sessions were randomized.

RESULTS

Figure 1 shows the mean acceptability scores of the 12 sentences for the 4 sets of boundary rules out of 8. The mean

scores for the 8 rule sets ranged from 4.4 to 7.3 on the 10-point scale and differed significantly from each other; $F_{(7,2789)} = 103.38$ ($p < 0.0001$).

As can be seen from this figure, the natural version scored highest (7.3) followed by the rule set with 5 levels of boundary strength (6.8). These two versions did not differ significantly from each other. This means that the rule set distinguishing 5 levels of boundary strength is very acceptable.

The version with three levels of boundary strength scored 6.1 and differed significantly from the other versions plotted in Figure 1. In turn, this version scored much higher than that without boundary rules (4.4). This is in agreement with the pilot experiment [5].

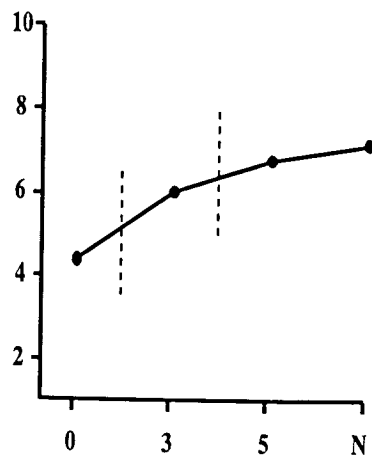


Figure 1: The mean score of acceptability for the four sets of boundary rules: 0 (version without boundary rules), 3 (version with 3 levels), 5 (version with 5 levels) and N (natural version).

DISCUSSION

From the results we can conclude that the rule set distinguishing 5 levels

of boundary strength produced speech as acceptable as the natural version. These two versions did not differ significantly.

The rule set with 5 levels is more acceptable than that with three, but both improved the synthetic speech quality in comparison to the version without boundary rules. Clearly, listeners find a well-phrased utterance much more acceptable than a poorly-phrased one. The results also confirm that the prosodic cues pause, melodic marker and declination reset are appropriate to mark boundaries.

The question remains whether listeners comprehend these well-phrased utterances more easily compared to poorly-phrased ones. Follow-up research is underway to explore this question.

ACKNOWLEDGMENTS

Many thanks are due to R. Collier, J-R de Pijper and M. Swerts for commenting upon an earlier version of this paper.

REFERENCES

- [1] Dirksen, A. (1992). Accenting and deaccenting: a declarative approach, in Proceedings of the 15th International Conference on Computational Linguistics, COLING '92, 3., pp 865-869.
- [2] Pijper, J.R. de & Sanderman, A.A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, 96 (4), 2037-2047.
- [3] Pisoni, D.B, Manous, L.M. & Dedina, M.J. (1987). Comprehension of natural and synthetic speech: effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language*, 2, 303-320.

[4] Sanderman, A.A. & Collier, R. (1994). Prosodic phrasing at the sentence level. *Festschrift for K. Harris of Physics, Modern Acoustics and Signal Processing Series*. American Institute of Physics.

[5] Sanderman, A.A. (1994). How can prosody segment the flow of (synthetic) speech? *Conference Proceedings of the second ESCA/IEEE Workshop on Speech Synthesis*, pp 147-150.

[6] Terken, J.M.B. (1993). Synthesizing natural sounding intonation for Dutch: rules and perceptual evaluation. *Computer Speech and Language*, 7, 27-48.

[7] Van Rijnsoever, P.A. (1988). A multilingual text-to-speech system. *Annual Progress Report No.23*. Institute for Perception Research, Eindhoven.