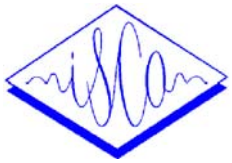

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

CHILD-DIRECTED SPEECH SYNTHESIS: EVALUATION OF PROSODIC VARIATION FOR AN EDUCATIONAL COMPUTER PROGRAM

David House, Linda Bell, Kjell Gustafson and Linn Johansson
Centre for Speech Technology, Department of Speech, Music and Hearing, KTH
Drottning Kristinas väg 31, 100 44 Stockholm, Sweden
http://www.speech.kth.se/

ABSTRACT

This paper addresses the question of how children between the ages of nine and eleven perceive and respond to prosodic variation in speech synthesis. Prosodic features were varied in samples of both concatenative and formant synthesis. Children and an adult control group were asked to compare these samples and to evaluate which were the most fun and which were the most natural. Results indicate that children perceive prosodic differences in the synthesis examples and prefer large manipulations in F0 and duration when a fun voice is intended. Even for naturalness, the children often prefer larger manipulations in F0 than are present in the default versions of the synthesis. These results can have implications for the implementation of synthesis in the context of a commercial computer program for children and more widely for child-directed speech synthesis.

Keywords: speech synthesis, prosody, child-directed speech, speaking styles, evaluation

10.21437/Eurospeech.1999-402

1. INTRODUCTION

There is currently considerable interest in examining different speaking styles for speech synthesis [1, 4]. In many new applications, naturalness and emotional variability have become increasingly important aspects. A relatively new area of study is the use of synthetic voices in applications directed specifically towards children. In this paper, we will focus on the question of how children perceive and evaluate prosodic features in synthesis designed to stimulate their learning and development.

It has been shown that there are prosodic differences between child-directed natural speech (CDS) and adult-directed natural speech (ADS). These differences often lie in increased duration and larger fundamental frequency excursions in stressed syllables of focussed words when the speech is intended for children [6, 8, 9]. Although many studies have focussed on speech directed to infants and implications for language acquisition, these prosodic differences have also been observed when parents read aloud to older children [2]. It could be useful to apply similar variation to speech synthesis for children, especially in the context of a fun and interesting educational program.

A key issue is to investigate how children react to prosodic variation which differs from default prosodic rules designed for text-to-speech applications directed to adults. The objective is to produce fun and appealing voices by determining what limits should be placed on the manipulation of duration and F0 in terms of both acoustic parameters and prosodic categories.

Also of interest is a comparison of children's reactions and evaluation of prosodic variation to an evaluation by adults of the same prosodic variation in the same context of an educational program. On the basis of investigations of child-directed speech it would be reasonable to assume that children prefer more prosodic variation than do adults.

2. METHOD

An animated character (an astronaut originally created for an educational computer game by Levande Böcker i Norden AB) was adapted to serve as an interactive test environment for speech synthesis. The astronaut in a spacesuit inside a spaceship was placed in a graphic frame in the center of the computer screen. Eight text fields in a vertical list on the right side of the frame were linked to sound files. By clicking on a text field the subjects could activate the sound file which also activated the visual animation. The animation began and ended in synchrony with the sound file. The test environment is illustrated in Figure 1.

Three sentences, appropriate for an astronaut, were synthesized using a developmental version of the Infovox 230 formant-based Swedish male voice and the Infovox 330 concatenated diphone Swedish female voice. Four prosodically different versions of each sentence and each voice were synthesized: (1) a default version, (2) a version with a doubling of the maximum F0 values in the focussed words, (3) a version with a doubling of duration in the focussed words, and (4) a combination of 2 and 3. There were thus a total of eight versions of each sentence and 24 stimuli in all. The sentences are listed below with the focussed words indicated in capitals.

(1) Vill du följa MED mig till MARS? (Do you want to come WITH me to MARS?)

(2) Idag ska jag flyga till en ANNAN planet. (Today I'm going to fly to a DIFFERENT planet.)

(3) Det tar mer än TVÅ DAGAR att åka till månen. (It takes more than TWO DAYS to get to the moon.)

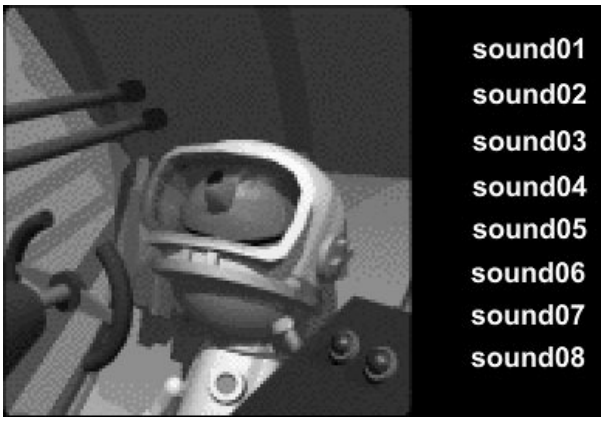

Figure 1. Illustration of the test environment

The sound files comprising the eight versions of each sentence were linked to the text fields next to the astronaut in mixed order. Each sentence was tested separately. Two different types of evaluation tasks were carried out; a ranking task and a scaling task. In the ranking task, subjects were asked to select the most natural and the most fun version of each sentence. In the scaling task, subjects were asked to rate each of the eight versions of each sentence on a scale from one to five where five is best.

A total of eight children participated in the listening tasks. The listeners were five boys and three girls, all between the ages of nine and eleven. Four of the subjects (two boys and two girls) served as a test panel for the ranking task. The other four subjects (three boys and one girl) participated in the scaling task. In both tasks, the subjects were allowed to listen to each stimulus as many times as they wished. Judgements were then presented orally to the test administrator (one of the authors) who recorded the results on a form. The ranking task was carried out with the test panel listening as a group while the subjects performed the scaling task individually. The children were paid modestly for their participation and also rewarded with soft drinks and snacks.

Four adults served as the adult comparison group and participated in the scaling task. The adults are all staff members of Levande Böcker i Norden AB and work with various aspects of educational computer programs for children. The adult subjects did not have extensive experience of speech synthesis. The adults performed the scaling task individually and recorded their results on a form.

3. RESULTS

3.1 Children

The children's results for the ranking task are presented in Table 1 for the most natural version and in Table 2 for the most fun version. The test panel seemed to prefer the diphone synthesis for naturalness as ten of the twelve votes were for this category. However, the panel also preferred higher F0 excursions or greater duration on the focussed word than in the default versions depending on the sentence, but did not consider a combination of higher F0 and greater duration as most natural.

The panel preferred the formant synthesis as the most fun, and for two of the three sentences (1 and 2) voted unanimously for the version with combined higher F0 and greater duration as can be seen in Table 2. For sentence 3, the version with greater duration only was chosen by three of the four subjects as the most fun.

Table 1. Total number of votes for the most natural version of each sentence in the ranking task (children)

Prosodic Version	Sentence 1 (Mars)		Sentence 2 (Planet)		Sentence 3 (Moon)	
	Diph	Form	Diph	Form	Diph	Form
Default		1				
F0			4			
Dur	3				3	
F0+dur						1

Table 2. Total number of votes for the most fun version of each sentence in the ranking task (children)

Prosodic Version	Sentence 1 (Mars)		Sentence 2 (Planet)		Sentence 3 (Moon)	
	Diph	Form	Diph	Form	Diph	Form
Default						
F0					1	
Dur						3
F0+dur		4		4		

Mean scores for the scaling task are presented in Tables 3 and 4. The results for naturalness in Table 3 differ from the ranking results in Table 1 in that the default versions and higher F0 only versions received higher scores in the scaling task while versions with increased duration and a combination of increased duration and higher F0 were judged as less natural.

The results on the fun scale in Table 4 are more consistent with the ranking results presented in Table 2 in that stimuli with manipulations to F0 and/or duration were given higher scores than the default stimuli. Furthermore, in the scaling experiment, no systematic differences between the diphone and formant synthesis types can be seen except for sentence 1 where the

formant synthesis was consistently rated slightly higher for both naturalness and fun.

Table 3. Mean scores on a five-point naturalness scale for each version of the test sentences in the scaling task (children)

Prosodic Version	Sentence 1 (Mars)		Sentence 2 (Planet)		Sentence 3 (Moon)	
	Diph	Form	Diph	Form	Diph	Form
Default	2.25	3.25	2.25	3.25	2.5	3.5
F0	2	3.5	3.75	4	2.75	3.75
Dur	2.25	2.5	2	2	2	2.75
F0+dur	2	2.75	2.5	3.25	2.5	1

Table 4. Mean scores on a five-point fun scale for each version of the test sentences in the scaling task (children)

Prosodic Version	Sentence 1 (Mars)		Sentence 2 (Planet)		Sentence 3 (Moon)	
	Diph	Form	Diph	Form	Diph	Form
Default	2.25	4	3.75	2.75	3.25	3
F0	3.25	4.25	4.5	4	3.5	2.5
Dur	3.5	4	3.25	3.25	3.5	3.25
F0+dur	3.25	3.75	3.5	4.25	4	4

Mean scores for each prosodic version in the scaling task are presented in Figure 2. Here it is evident that all stimuli scored higher in general on the fun scale than on the naturalness scale. For the fun category, the default version scored lowest while all three types of manipulations contributed to increased ratings in this category. For naturalness, the children clearly preferred the F0 manipulated version. The default version, however, scored higher for naturalness than the versions with increased duration and a combination of duration and F0.

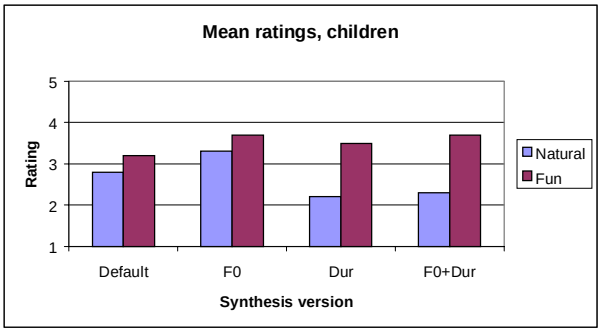

Figure 2. Results showing mean scores on a five point scale for naturalness and fun.

3.2 Adults

Mean scores for the adult listeners in the scaling task are presented in Tables 5 and 6. The results for naturalness in Table 5 show a general preference for the default version and the F0 manipulated version of the diphone-based synthesis. For the fun category (Table 6), however, the adults preferred the F0 manipulated version of both diphone-based and formant synthesis.

Table 5. Mean scores on a five-point naturalness scale for each version of the test sentences in the scaling task (adults)

Prosodic Version	Sentence 1 (Mars)		Sentence 2 (Planet)		Sentence 3 (Moon)	
	Diph	Form	Diph	Form	Diph	Form
Default	3.5	2.5	4	2.25	3	2.25
F0	4	2	3.25	2	4	2.5
Dur	2.75	1	3.5	1.75	3	1.75
F0+dur	2.25	1	3.5	1.75	2.5	2

Table 6. Mean scores on a five-point fun scale for each version of the test sentences in the scaling task (adults).

Prosodic Version	Sentence 1 (Mars)		Sentence 2 (Planet)		Sentence 3 (Moon)	
	Diph	Form	Diph	Form	Diph	Form
Default	3	2.25	2	2.5	1.75	2
F0	3.75	3.75	4.25	3.75	3.5	3.5
Dur	2	2	2	3.25	2.25	3.25
F0+dur	2	1.75	3	4.25	3.75	2.75

Mean scores for each prosodic version in the scaling task for the adult listeners are presented in Figure 3. For the adult listener group, both the default prosody and F0 manipulation received the highest rating for the naturalness condition. For the fun condition, the F0 manipulated version was preferred. These results differ from those of the children in that the adults seem more critical to versions containing manipulations to duration in the fun category.

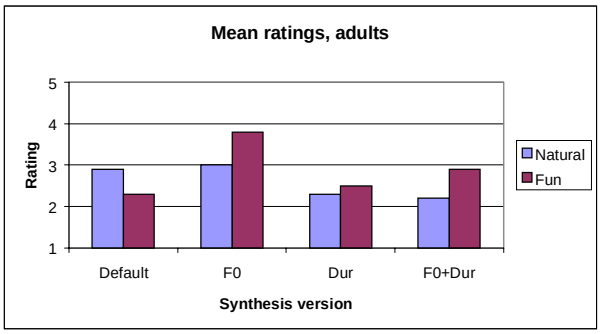

Figure 3. Results showing mean scores on a five point scale for naturalness and fun.

3.3 Synthesis type: diphone vs formant

Mean ratings for naturalness and fun grouped by synthesis type (diphone and formant) are presented for both listener groups in Figure 3. The adult listeners seem to be more sensitive to synthesis type preferring

diphone synthesis for naturalness while rating both diphone and formant as equal in the fun category. Children preferred the formant version for both naturalness and fun rating. These results conflict with the results from the children who served as the listener panel for the ranking task (Tables 1 and 2). In the ranking task, the children preferred the diphone synthesis for naturalness and the formant synthesis for the fun category.

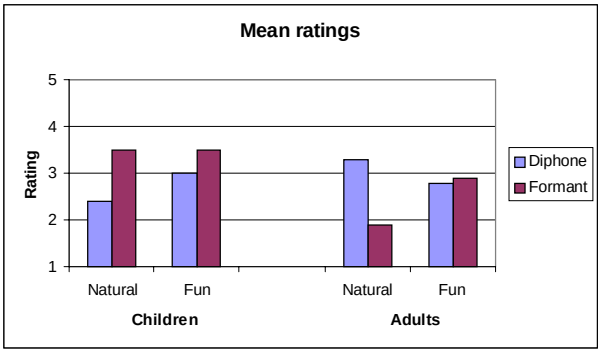

Figure 4. Results showing mean scores on a five point scale for naturalness and fun grouped by synthesis type.

4. DISCUSSION

Although this study comprises a limited number of subjects, it is clear that the children responded to prosodic differences in the synthesis examples in a fairly consistent manner, preferring large manipulations in F0 and duration when a fun voice is intended. Even for naturalness, the children often prefer larger manipulations in F0 than are present in the default versions of the synthesis which is intended largely for adult users. Differences between the children and the adult listeners were according to expectation, where children preferred greater prosodic variation, especially in duration for the fun category. Furthermore, the children also preferred the formant synthesis. In the context of this experiment the children may have had a tendency to react to formant synthesis as more appropriate for the animated character portraying an astronaut while the adults may have judged the synthesis quality from a wider perspective.

The children also responded positively to changes which involved the focussed words only, although manipulations involving the entire sentence were not tested, as this was judged to produce highly unnatural and less intelligible synthesis. Manipulations in the current synthesis also involved raising both peaks (maximum F0 values) of the focal accent 2 words. This is a departure from the default rules [3] but is consistent with production and perception data presented in Fant and Kruckenberg [5]. This strategy may be preferred when greater degrees of focal accent are intended.

Prosodic variation may be an important consideration when using speech synthesis as part of an educational computer program or when designing spoken dialogue systems for children [7]. The strategy used here of involving focussed words only puts greater demands on the synthesis system to correctly indicate focus. The use of synthesis also allows the educational games to acquire an interactive dimension where children can write their own character lines and have the characters speak these lines. If children are to enjoy using a text-to-speech application in an educational context, it is probable that more prosodic variation should be incorporated in the prosodic rule structure. Further tests where children themselves write in the utterances and have some control over prosodic parameters could prove highly useful in designing text-to-speech synthesis for children.

5. ACKNOWLEDGEMENTS

We wish to thank David Skoglund for assistance in creating the interactive test environment and the subjects who participated in the experiments.

6. REFERENCES

[1] Abe M. (1997). Speaking styles: statistical analysis and synthesis by a text-to-speech system. In van Santen, J.P.H., Sprout, R., Olive, J.P. and Hirschberg, J. (eds) *Progress in speech synthesis*, 495-510. New York, Springer-Verlag.

[2] Bredvad-Jensen A-C. (1995). Prosodic variation in parental speech in Swedish. In *Proceedings of ICPhS-95, Stockholm, Sweden* 3, 389-399.

[3] Bruce, G. and Granström, B. (1993). Prosodic modelling in Swedish speech synthesis. *Speech Communication* 13, 63-73.

[4] Carlson R., Granström B. and Nord L. (1992). Experiments with emotive speech – acted utterances and synthesized replicas. In *Proceedings of the International Conference on Spoken Language Processing. ICSLP–92, Banff, Alberta, Canada* 1, 671-674.

[5] Fant G. and Kruckenberg A. (1998). Prominence and accentuation. Acoustical correlates. In *Proceedings FONETIK 98*, Dept. of Linguistics, Stockholm University, 142-145.

[6] Kitamura C. and Burnham D. (1998). Acoustic and affective qualities of IDS in English. In *Proceedings of ICSLP 98*, Sydney, 441-444.

[7] Potamianos A. and Narayanan S. (1998). Spoken dialog systems for children. In *Proceedings of ICASSP 98*, Seattle, 197-201.

[8] Snow C.E. and Ferguson C.A. (eds) (1977). Talking to children. Language input and acquisition. Cambridge, MA Cambridge University Press

[9] Sundberg U. (1998). *Mother tongue – Phonetic aspects of infant-directed speech*. (Perilus XXI), Department of Linguistics, Stockholm University.