Fun or Boring? A Web-based Evaluation of Expressive Synthesis for Children

Kjell Gustafson and David House

Centre for Speech Technology, Department of Speech, Music and Hearing, KTH
Drottning Kristinas väg 31, 100 44 Stockholm, Sweden
{kjellg|davidh}@speech.kth.se

Abstract

Prosodic features were varied in four sentences synthesized using a developmental version of the Infovox 330 concatenated diphone Swedish male voice. The sentences were part of an interactive evaluation test carried out on a commercial website for a period of three months. 78 girls and 56 boys between the ages of 5 and 15 rated the sentences on a qualitative four-point scale. Results indicate that both girls and boys interpret large-scale F0 manipulations as representing a fun voice while longer durations are generally regarded as boring, especially by the boys. The results also confirm the feasibility of using a website for remote evaluation even with children.

1. Introduction

There is currently considerable interest in examining different speaking styles for speech synthesis [1], [2] and [3]. Naturalness, expressiveness and emotional variability are becoming increasingly important aspects of many new applications [4]. 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 aspects of expressive synthesis.

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 [5], [6] and [7]. 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 [8].

In a previous study, prosodic parameters (duration and F0) were varied in samples of both formant and concatenative synthesis [9]. Although the study comprised a limited number of subjects (eight children and four adults), 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 was intended. 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.

This paper presents results from a follow-up study carried out on a much larger scale. The goal of the study was two-fold. First of all we wanted to see if the previous results could be replicated using a commercial web-based evaluation environment which would attract considerably more subjects. Secondly, we wished to test the feasibility of using such a web-based environment for testing children.

2. Method

The study was carried out in collaboration with a leading Swedish company for learning software, *Levande Böcker i Norden AB*. The web-based environment used for this study was constructed as part of *Levande Böcker's* website for children between the ages of six and thirteen [10]. The site is freely accessible to members who have joined by supplying their name, age, email address and a password. In addition to information about *Levande Böcker's* products and game tips, the site contains a number of educational games and activities for children such as interactive science quizzes and protocols for constructing and viewing home pages, and for sending email postcards to other members. Users can also download extra games and components for *Levande Böcker's* CD-ROM software.

Subjects entered the test environment by clicking on a blimp flying over the sea on the website's main page. Inside the airship, the subjects were greeted by Professor Voxmix in his "voice lab" who asked them to help him with his experiment on computer voices. Instructions were given in text form in speech bubbles and in an instruction manual. Text was used so as not to influence the evaluation of the test utterances. The environment is illustrated in Figure 1.

Figure 1. Test environment showing Professor Voxmix, one of the four characters on the monitor, numbered buttons on the monitor for the different versions of the utterances and the evaluation buttons on the handheld computer.

Four sentences in the form of a short dialog, semantically appropriate for the Professor Voxmix context, were synthesised using a developmental version of the Infovox 330 concatenated diphone Swedish male voice. Four prosodically different versions of each sentence were synthesised: (1) a default version, (2) a version with an approximate doubling of the maximum F0 values in the focussed words, (3) a version with an approximate doubling of duration in the focussed words, and (4) a combination of 2 and 3. There were thus a total of four versions of each sentence and 16 stimuli in all. The sentences are listed below with the focussed words indicated in capitals.

(1) När jag blir vuxen vill jag bli VETENSKAPSMAN. *When I grow up I want to be a SCIENTIST.*

(2) Ja! Vetenskapsmän får ju göra så många SPÄNNANDE saker. *Yes! Scientists get to do so many EXCITING things.*

(3) Kommer MASKINERNA att kunna läsa våra TANKAR? *Will MACHINES be able to read our THOUGHTS?*

(4) Jag HOPPAS dom inte kommer att kunna läsa MINA tankar! *I HOPE they won't be able to read MY thoughts!*

Although the experimental rules were designed to generate a doubling of both F0 maxima and duration in various combinations, there is a slight deviation from this ideal in the actual realisations. This is due to the fact that there are complex rules governing how declination slope and segment durations vary with the length of the utterance, and this interaction affects the values specified in the experiments. However, as it was not the intention in this experiment to test exact F0 and duration values, but rather to test default F0 and duration against rather extreme values of the same parameters, these small deviations from the ideal were not judged to be of consequence for the results. Figure 2 shows parameter plots for the four versions of sentence 3. As can be seen from the diagrams, the manipulation was localised to the focussed word(s).

Subjects were required to supply their name, age, and gender, and then requested to evaluate each version of each sentence on a qualitative four-point scale using the words "super fun", "fun", "boring" and "totally boring" by marking the corresponding button on a virtual handheld computer (see Figure 1). The subjects could listen to each version as many times as they wanted by clicking on a numbered button on the monitor. To make the test more interesting, each sentence was coupled to a different character on the monitor. The four sentences were always presented in the dialog order as listed above, with the other variables being randomised (i.e. version presentation order and character). When the subject had evaluated all four versions of sentence 1, sentence 2 was presented, and so on until the subject completed the test.

Each subject's name, age, gender and evaluation results were logged as were the presentation order for the different versions for each sentence and which character ID was coupled to each sentence. This information was collected in a database for a period of three months.

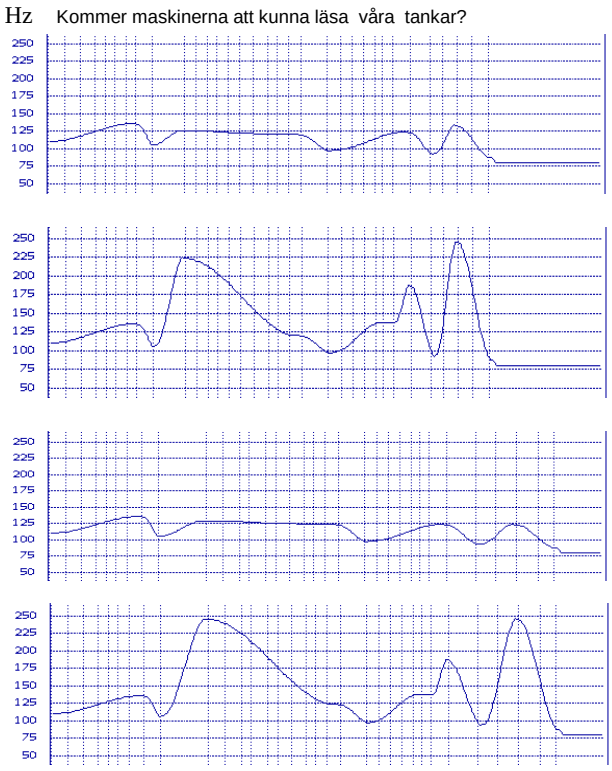

Figure 2. F0 parameter plots for sentence 3. From top to bottom: default, F0 max extended, duration extended, F0 max and duration extended.

3. Results

A total of 1748 subjects completed part of the test of which 78 girls and 56 boys between the ages of 5 and 15 completed the entire test. Only the results from these subjects are reported here. One subject who gave his age as 2 was excluded from the data as well as three adults aged 24, 25 and 32. Figure 3 shows the age distribution of the subjects.

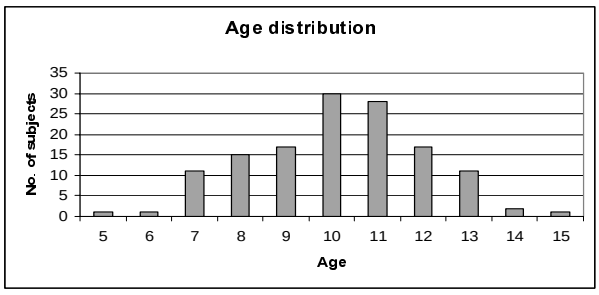

Figure 3. Age distribution of children who completed the test.

Figure 4 shows the distribution of votes for the four evaluation categories as a function of prosodic type for all children and for girls and boys separately. As can be seen, the versions with extended F0 range were evaluated as more fun than those with default F0 values, and versions with extended durations were judged as more boring than those with standard durations. The combination of extended F0 range

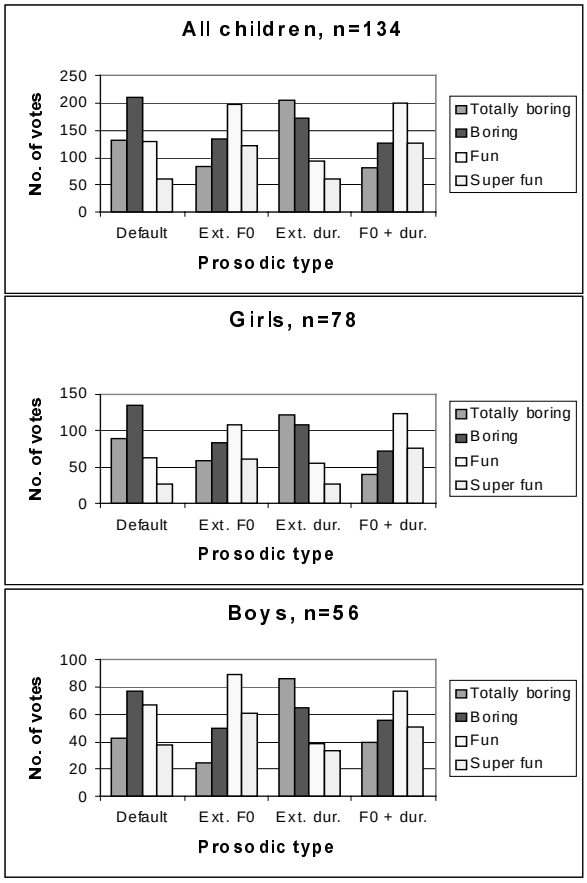

Figure 4. Results of prosodic type for all children who completed the task, and for girls and boys separately.

and extended duration was also evaluated as more fun than the default, especially by the girls. A $\chi 2$ test of independence on the distribution for all children gives $p<0.001$ where $\chi 2=30.44$ and $df=9$.

There was a general tendency to evaluate the sentences as boring or fun rather than the extremes totally boring and super fun. Only the extended duration evoked a majority of extreme votes for totally boring.

In Tables 1-3, the distributions of votes are collapsed into two categories: boring and fun. There is an overall bias for the girls to choose the boring category while the boys' votes are more evenly distributed.

In Table 1, showing distributions as a function of prosodic type, both girls and boys find extended duration as the most boring version by far. The girls, however, find the combination of F0 and duration as most fun, while the boys find F0 alone as most fun. A $\chi 2$ test of independence on the distribution in Table 1 for all children gives $p<0.001$ where $\chi 2=172.98$ and $df=3$.

Table 2 shows the distribution of votes as a function of sentence number. Only sentence 3 was evaluated as more fun than boring and only by the boys. A $\chi 2$ test of independence on the distribution in Table 2 for all children gives $p<0.05$ where $\chi 2=8.16$ and $df=3$.

Table 3 shows the distribution of votes as a function of character ID. This variable did not produce significant distribution differences: for all children $p>0.05$ where $\chi 2=1.43$ and $df=3$.

Table 1. Distribution of votes for (totally) boring and (super) fun determined by prosodic type

Prosodic type	Girls		Boys		All	
	Boring	Fun	Boring	Fun	Boring	Fun
Default	223	89	120	104	343	193
F0	142	170	74	150	216	320
Dur	229	83	151	73	380	156
F0+dur	112	200	96	128	208	328
Total	706	542	441	455	1147	997

Table 2. Distribution of votes for (totally) boring and (super) fun determined by sentence number

Sentence number	Girls		Boys		All	
	Boring	Fun	Boring	Fun	Boring	Fun
1	179	133	118	106	297	239
2	191	121	110	114	301	235
3	156	156	103	121	259	277
4	180	132	110	114	290	246
Total	706	542	441	455	1147	997

Table 3. Distribution of votes for (totally) boring and (super) fun determined by character ID

Character ID	Girls		Boys		All	
	Boring	Fun	Boring	Fun	Boring	Fun
1	178	134	119	105	297	239
2	174	138	104	120	278	258
3	181	131	107	117	288	248
4	173	139	111	113	284	252
Total	706	542	441	455	1147	997

4. Discussion

These results are consistent with the results of the earlier experiment [9] and also verify the feasibility of using the Internet for running remote evaluation experiments even with children. The age distribution shown in Figure 3 is quite in accordance with the expected distribution regarding the user profile of the *Levande Böcker* website. This helps offset concerns about the lack of control over the selection of subjects. Additionally, the number of subjects who completed the test during the three-month period is encouraging for the use of this type of evaluation in the future. Of course, access to a website which attracts many visitors is a prerequisite.

It is quite clear from these results that a boring voice can be created by extending durations on the focussed words and that a fun voice can be created by extending F0 excursions on these words. However, the interplay between these cues is also interesting. It appears that F0 is more salient in that the combination stimuli elicited more fun votes than boring and that the girls actually favoured this version as most fun.

The children responded 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 synthesis. Manipulations in the current

synthesis also involved raising both F0 peaks of the focal accent 2 words. This is a departure from the default rules [11] but is consistent with production and perception data presented in Fant and Kruckenberg [12]. This strategy may be preferred when greater degrees of focal accent are intended and can contribute to more expressiveness.

The different sentences did not have a major effect on the evaluation results with three of four sentences being evaluated as boring (Table 2). However, the fact that the boys evaluated sentence 3 as fun and that this resulted in a significant distribution difference at the 5% level merits a comment. Sentence 3 was given two foci where the final focal accent 2 concluded the sentence. According to the rule structure, this sentence contained the greatest F0 excursions of the four sentences (see Figure 2).

The use of only two prosodic parameters in this experiment demonstrates the importance of F0 and duration for expressive synthesis. However, the general bias for evaluating the voices as boring may indicate that the creation of a fun voice is the more challenging of the two. The evaluation of the extended duration stimuli as "totally boring" (the only extreme evaluation) shows that the use of this one parameter may be sufficient to create a convincingly boring voice. For a fun voice, however, other parameters may be needed in addition to a more carefully controlled relationship between duration, F0 peak height and F0 range, such as those proposed in the modelling of prosody in the context of a man-machine dialogue system (the Waxholm project) for Swedish [13]. In addition to such strictly prosodic parameters, voice quality characteristics, such as breathy and tense voice, are likely to be highly relevant to the creation of a convincing fun voice [14]. Further investigations are also needed to establish how voice quality characteristics interact with the prosodic parameters.

5. Conclusions

This study demonstrates the importance of prosodic variation for creating expressive synthesis for children for use in educational programs or spoken dialog systems [15]. If children are to enjoy using a text-to-speech application, more prosodic variation needs to be incorporated in the prosodic rule structure. Fairly simple prosodic variation can achieve considerable success when modelling both fun and boring characters. Furthermore, the interactive dimension of synthesis can be exploited in certain applications where children could write their own character lines and have the characters speak these lines. Children could also be allowed to have some control over prosodic parameters with a variety of animated characters.

6. Acknowledgements

The research reported here was carried out at the Centre for Speech Technology, a competence centre at KTH, supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations. We wish to thank Linn Tornérhielm, Thomas Wiroth and *Levande Böcker* for fruitful collaboration on the test environment. We are also grateful for having had the opportunity to expand this research within the framework of COST 258.

7. 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] 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.

[3] Cahn, J. (1998). Generating pitch accent distributions that show individual and stylistic differences. In: *Proceedings of the 3rd ESCA/COCOSDA International workshop on speech synthesis*. Jenolan Caves, Australia, 121-126.

[4] Keller, E. (Forthcoming). Towards greater naturalness: Future directions of research in speech synthesis. In: Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. (eds.) *Improvements in Speech Synthesis*. New York, NY: John Wiley & Sons.

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

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

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

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

[9] House, D., Bell, L., Gustafson, K. and Johansson, L. (1999). Child-directed speech synthesis: evaluation of prosodic variation for an educational computer program. In: *Proceedings of Eurospeech 99*, Budapest, 1843-1846.

[10] http://www.barnlandet.se/

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

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

[13] Bruce, G., Granström, B., Gustafson, K., Horne, M., House, D. and Touati, P. (1995). Towards an enhanced prosodic model adapted to dialogue applications. In: Dalsgaard P et al. (eds.), *Proceedings of ESCA Workshop on Spoken Dialogue Systems,* Vigsø, Denmark, 201-204.

[14] Gustafson, K. and House, D. (Forthcoming). Prosodic Parameters of a 'Fun' Speaking Style. In: Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. (eds.) *Improvements in Speech Synthesis*. New York, NY: John Wiley & Sons.

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