Lund University, Dept. of Linguistics
Working Papers 49 (2001), 50–53

Expressive synthesis for children, a web-based evaluation

Kjell Gustafson and David House
Centre for Speech Technology, Department of Speech, Music and Hearing, KTH
{kjellg|davidh}@speech.kth.se

Abstract

Prosodic features were varied in four sentences synthesised 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 including aspects of naturalness, expressiveness and emotional variability (Keller, forthcoming). It has been shown that there are prosodic differences between child-directed natural speech and adult-directed natural speech. These differences often lie in increased duration and larger fundamental frequency excursions in speech intended for children (Bredvad-Jensen 1995, Sundberg 1998). In a previous study comprising a limited number of subjects, children responded to prosodic differences in both formant and concatenative synthesis in a fairly consistent manner, preferring large manipulations in F0 and duration when a fun voice was intended (House et al. 1999).

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 (http://www.barnlandet.se/). 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.

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.

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. Information containing each subject's name, age, gender and evaluation results were logged and collected in a database over a period of three months.

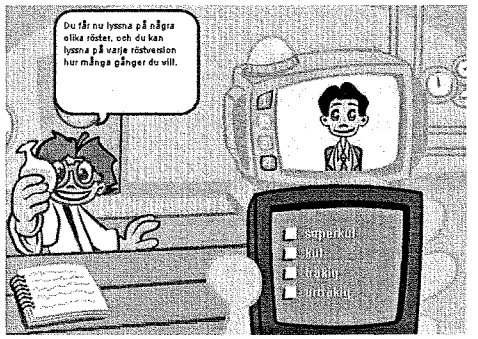

Figure 1. Test environment showing Professor Voxmix, one of the four characters on the monitor, numbered buttons on the monitor for the different versions of the utterances and the evaluation buttons on the handheld computer.

(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!

3 Results

A total of 78 girls and 56 boys between the ages of 5 and 15 completed the entire test. In Figure 2 the age distribution of the subjects is shown along with 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 and extended duration was also evaluated as more fun than the default, especially by the girls. A χ^2 test of independence on the distribution for all children gives $p<0.001$ where $\chi^2=30.44$ and df=9.

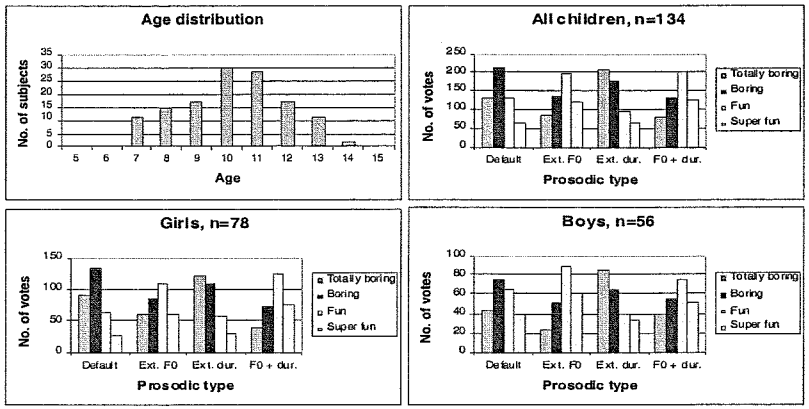

Figure 2. Age distribution of the subjects and results for prosodic type.

In Tables 1-3, the distribution of votes are collapsed into two categories: boring and fun as a function of prosodic type, sentence and character ID. A χ^2 test of independence was run on the distribution for all children in each table. Table 1 (prosodic type): χ^2=172.98, df=3, p<0.001; Table 2 (sentence): χ^2=8.16, df=3, p<0.05; Table 3 (character): χ^2=1.43, df=3, p>0.05. Differences by prosodic type were highly significant while those by character were not significant. Differences by sentence were marginally significant.

4 Discussion

These results are consistent with the results of the earlier experiment (House, et al. 1999) and also verify the feasibility of using the Internet for running remote evaluation experiments even with children. The age distribution shown in Figure 2 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 than duration in that the combination stimuli elicited more fun votes than boring and that the girls actually favoured this version as most fun.

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 Bruce et al. (1995). In addition to such strictly prosodic parameters, voice quality characteristics are likely to be highly relevant to the creation of a convincing fun voice (Gustafson & House, forthcoming). Further investigations are also needed to establish how voice quality characteristics interact with the prosodic parameters.

Table 1. Distribution of votes 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 by sentence.

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 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

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 organisations. We wish to thank Linn Tornérhielm, Thomas Wiroth and Levande Böcker for fruitful collaboration.

References

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

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

Gustafson, K. & House, D. Forthcoming. Prosodic parameters of a 'fun' speaking style. In E. Keller, G. Bailly, A. Monaghan, J. Terken, & M. Huckvale (eds.) *Improvements in Speech Synthesis*. New York, NY: John Wiley & Sons.

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

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

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