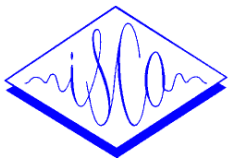

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PRESERVING NATURALNESS IN SYNTHETIC VOICES WHILE MINIMIZING VARIATION IN FORMANT FREQUENCIES AND BANDWIDTHS

Niels-Jørn Dyhr, Marianne Elmlund, and Carsten Henriksen
(names in alphabetical order)

Tele Danmark Research
Lyngsø Allé 2, DK-2970 Hørsholm, Denmark

ABSTRACT

As a preliminary to improving the naturalness of the synthetic male and female voices in a Danish text-to-speech system using a rule-driven formant synthesizer, the relative importance of the individual formant frequencies and bandwidths has been investigated. Recordings of a Danish compound word consisting entirely of voiced segments were analyzed. Based on these recordings and the analysis, a number of manipulated, synthetic stimuli were created and presented in two listening tests. The main results of these simplifications are: a) Bandwidths (B5-B8) are more sensitive to simplifications than formants (F5-F8). b) F5-F8 may be held constant throughout the utterance, and B1-B4 may be kept constant per segment without perceptible loss of naturalness. c) B5-B8 may also be held constant, though with a minor loss of naturalness. A similar approach has been tried with female synthetic voices, and preliminary results corroborate the results outlined above. Among the more comprehensive simplifications in the male voice a hierarchy of acceptability was established.

I. INTRODUCTION

For a number of years two Danish research institutions, Tele Danmark Research (formerly Telecommunications Research Laboratory) and Department of General and Applied Linguistics, University of Copenhagen, have been collaborating on the development of a text-to-speech synthesis system for Danish [1] [2] [3]. The various modules of the system are rule-driven, e.g. [4], and the synthesizer is a formant-coded terminal analog, originally with a sampling frequency of 10 kHz [5] [6]. The voice source is an implementation of the Liljencrants/Fant (LF) model [7]. The system has attained an adequate level of intelligibility, yet the perceived naturalness is somewhat unsatisfactory. Attempts have previously been made to address this shortcoming [8]. In connection with increasing the sampling frequency from 10 to 20 kHz in the synthesizer it proved necessary to gather more detailed information on formant frequencies and bandwidths from natural speech. The aim was to arrive at simplifications in the parameter values extracted from natural speech without sacrificing perceived naturalness. In this way the task of formulating phonetic rules could be considerably eased [8].

II. EXPERIMENT AND TEST PROCEDURE

The following preliminary steps were taken:

a) Recordings from a database of natural speech were analyzed for resynthesis.
b) The parameter tables from the analysis were further manipulated with respect to formant frequencies and bandwidths.
c) A listening test was prepared from the synthetic utterances for the purpose of providing an overview. A second test was used to supplement and clarify the results of the first one.

A more detailed description is given below.

a) Recordings of a Danish compound word consisting entirely of voiced segments served as basis. The speech material was selected from a database and consisted of utterances from one male speaker.
The recordings had been made to meet the requirements of inverse filtering and covered the range 22.5-10.000 Hz. The speech was recorded to DAT tape via a Bruel & Kjaer condenser microphone type 4179 and a Bruel & Kjaer microphone preamplifier type 2660, and a Bruel & Kjaer measuring amplifier type 2607. The recordings were made in an anechoic chamber at the Technical University of Denmark.
Parameter values for formant frequencies and bandwidths were extracted from the speech material with locally developed analysis software. These parameters were then used to drive a 20 kHz synthesiser with eight formants. The voice source parameters were kept constant in all cases. Both formant extraction and synthesis were pitch-synchronous. The latter synthesizer is in all respects identical to the synthesizer that exists in the TTS system, except that it accepts a different input format (notably the output from our analysis software), and allows the experimenter to override the rule base to any extent.
The resynthesis scheme provided synthetic output which sounded very similar to the natural counterpart, it was in fact possible to identify the original speaker without difficulty. One of these utterances was selected to serve as reference and basis for the manipulations.

b) The manipulation consisted in systematic

simplification of formant frequency values and bandwidth values in the parameter tables resulting from the analysis. The formant frequencies and bandwidths were divided into the following parameter groups: F1-F3 and B1-B3, F4 and B4, and finally, F5-F8 and B5-B8. This particular division was motivated partly by common phonetic considerations, partly by preliminary, informal tests. Of the above parameter groups F1-F3 were never touched, while the others were set constant, either through the segment or through the entire utterance. A constant value is here defined as the values in a parameter averaged over the segment or over the entire utterance, respectively. Below, the former kind of manipulation will be denoted by *Xs*, the latter by *Xu*.

c) An AB listening test was prepared from the synthetic utterances. In the test, all manipulated stimuli appeared in turn against the reference stimulus (i.e. the unmodified, resynthesized utterance). From the outset it was considered impossible to include all possible comparisons in the test; instead a broad set of stimuli was selected to appear in the first test. These selected comparisons were presented to groups of four listeners via two loudspeakers in a sound-treated room (loudspeakers were preferred because they are used in the normal development situation). The listeners were phonetically naive and had passed a hearing test prior to the listening task. In all, 24 subjects participated in this test.
The second test was similar to the first one, only, there were certain new stimuli and only a few comparisons against the reference. This time the principal aim was to test the relative merits of those simplifications which had turned out to be interesting. 19 listeners out of the previous group of subjects took the second test.

III. RESULTS

The responses from the listening tests were tested for significance with a chi-square test (one-tailed goodness of fit test). The required level of significance was 5 pct.

In this section the stimuli will be presented graphically as shown in figure 1.

Fig. 2 below presents an overview of the score against the reference.

3.1 Formants

From fig. 2 it appears that all manipulations that involve formants exclusively (stim. 12, 14, 15, 16) are found either non-significantly different from the reference or (in one case) significantly better than the reference (stim. 16).

3.2 Bandwidths

It appears from fig. 2 that the parameter groups B1-B3 and B4 *Xs* simplifications are not significantly different from the reference (stim. 10, 11, 13).
The parameter group B1-B3, B4, and B5-B8 cannot be subjected to *Xu* simplification as tested against the reference (stim. 2, 4, 7, 9). However, stimuli containing B1-B3 and B4 *Xs* reduction vs. B1-B3, B4 *Xs* and B5-B8 *Xu* simplification are not found significantly different. This implies that *Xu* simplification of B5-B8 is allowed in this particular context, however with some loss of naturalness.

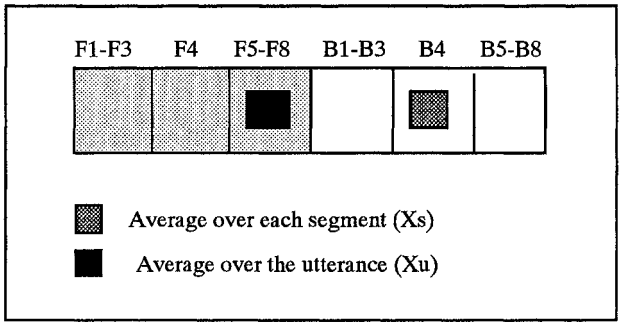

Fig. 1. A key to the graphic representation of the stimuli. The parameter bar contains six fields, one for each parameter group, the shaded fields representing the formants, unshaded fields the bandwidths. In the figures below, the bar may be oriented vertically or horizontally. The particular manipulation of the stimuli illustrated in fig. 1 is described as follows: The formant groups, F1-F3 and F4 are unmanipulated, the parameters in formant group F5-F8 are averages of these individual formants over the utterance (Xu simplification). The bandwidth groups B1-B3 and B5-B8 are unmanipulated, while B4 is the average B4 values over each segment (Xs simplification).

3.3 F4 and B4

It appears from fig. 2 (stim. 16) that it is possible to simplify F4 to an *Xu* value, but if this is done in combination with most other simplifications it leads to a significant loss of naturalness. On the other hand, there are combinations in which this F4 manipulation does not affect the perceived naturalness.
Against the reference, B4 is not amenable to *Xu* simplification (stim. 4), nor in combination with simplifications in other groups.

3.4 Formants vs. Bandwidths

In order to clarify the relative importance of formant and bandwidth manipulations we compared the stimuli exclusively involving manipulations of formants to those exclusively involving manipulations of bandwidths. The following distributions of the scores were observed:
In the formant group, 25 pct. were better than the

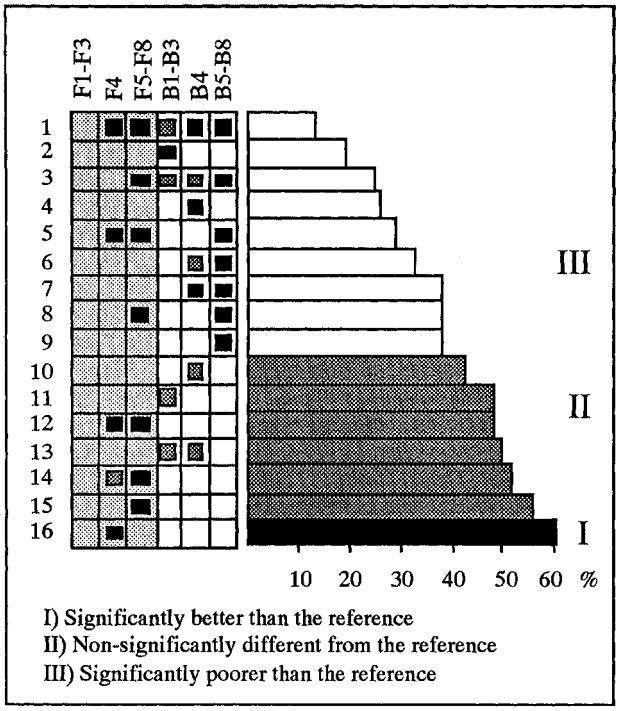

Fig. 2. Percentage score against reference stimulus. The left part of the figure shows the manipulations associated with each stimulus as explained in fig. 1. The histogram to the right shows the score of each stimulus.

reference, while 75 pct. were found non-significantly different from the reference; none was significantly poorer. In the bandwidth group, none was better than the reference, 62.5 pct. were found non-significantly different from the reference (all these were *Xs* simplifications), and 37.5 pct were significantly poorer (all these were *Xu* simplifications).

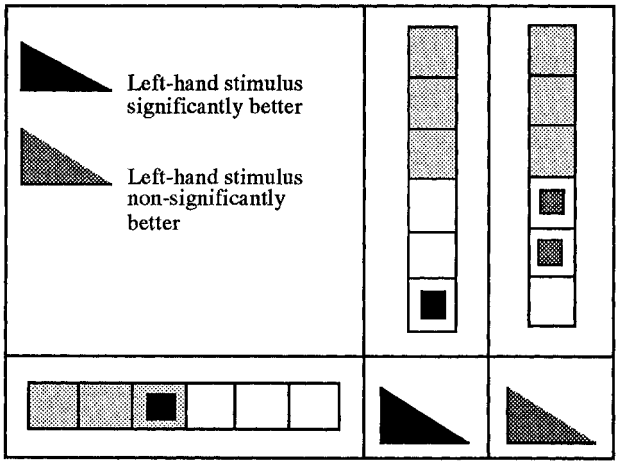

Fig. 3. Formant manipulation vs. bandwidth manipulation.

Fig. 3 illustrates the relative importance of F5-F8 *Xu* manipulations vs. bandwidth manipulations.
I appears that the F5-F8 *Xu* simplification is found significantly better than B5-B8 *Xu* simplification, while there is non-significant difference between the formant manipulation and *Xs* simplification of B1-B3 and B4.

3.5 Formant/Bandwidth Simplification in Combination

As the aim of the investigation was to economize rule writing, a number of the cruder, more comprehensive simplification strategies were selected for establishing a hierarchy. These stimuli all belonged in the group which obtained a significantly lower score than the reference (group III in fig. 2). At first sight this may seem an odd choice, but in order to gain simplicity we wanted to reduce at least F5-F8 and B5-B8 (stim. 1, 3, 5, 8). We also wanted to consider F4 and B4 (stim. 1, 5). And finally, simplification of B1-B3 (stim. 1, 3) was included. An all-against-all comparison showed that only one of these simplifications (stim. 8) stood out as being preferable to the others, which were found indistinguishable. The reductions preferred are *Xu* simplification of F5-F8 and B5-B8 (stim. 8 in fig. 2), although no difference could be shown between this stimulus and the stimulus which in addition contains B1-B3 and B4 *Xs* simplification (stim. 3).

3.6 Simplifications Applied to Female Voices

To verify if the simplifications from the male voice have any bearing on female voices, a similar set of simplifications were carried out on two different female voices (one with a high F0, the other with a relatively low F0). For both of these voices the synthetic reference signals were of a poorer quality than for the male voice, still, the original speakers were clearly identifiable.
As the work on the female voice has not progressed so far as the work on the male voice, it was only possible to obtain data from an informal listening test.
This test was taken by a small group of listeners, this time including the authors. It brought out that, generally speaking, all kinds of simplications applied to the male voice were allowed in the female voice. None of the simplifications were found poorer than the reference.
In one of the voices, however, *Xu* simplification of F5-F8, B5-B8, and *Xs* simplification of B1-B3, B4 was not allowed. Also, *Xu* simplification of F5-F8 failed, but in this case there was conflicting evidence from some of the other stimulus pairs. Obviously, some of the uncertainty can be ascribed to the rather small number of responses in the informal test.

IV. CONCLUSION

The investigation has brought out that it is possible to reduce the variation in parameter values without

significantly impairing the perceived naturalness of the synthetic speech.

In consideration of the linguistic function of the three first formants, it appears reasonable to assume that any simplification in these values would likely distort the identity of the individual segments and generally impair the synthetic speech.
As stated above, preliminary tests indicated that it would be interesting to look into the role of F4 and B4 to see if they belong to the low parameter groups, or rather belong to the high parameter groups.
The results indicate that F4 and B4 belong to the low parameter groups, in that it proved impossible to apply any sort of simplification to this group without perceptible loss of naturalness.

It turned out, a little unexpectedly, that no bandwidths can be simplified to averages of the entire utterance without loss of naturalness; all stimuli containing this particular manipulation were found significantly poorer from the reference.

The comparison between formant and bandwidth manipulations indicate that the bandwidths are more sensitive than the formants, as all the formant manipulations are non-significantly different from the reference. The majority of the bandwidth manipulations, however, are significantly poorer than the reference.
Apparently, bandwidths are generally quite sensitive and each plays a more important role than previously assumed.

The stimuli with the highest scores are also those which contain the least simplifications. The practical, everyday development situation requires more substantial simplifications, however. Therefore, a set of more far-reaching simplifications, although with a lower score, might serve the need better. These considerations led us to adopt a simplification which consists in averaging the individual parameter values of F5-F8 and the corresponding bandwidths over the entire utterance, and at the same time averaging the B1-B4 parameter values on a per-segment basis.
Thus, the task of writing phonetic rules can be eased, and the generalizations contained in the more comprehensive simplifications can still add naturalness to the synthetic speech. The F1-F3 group presents a problem in the sense that the frame-to-frame variation observed in the resynthesized reference cannot be imitated in the phonetic rules. In this parameter group stady-state frequencies and appropriate transitions substitute for the variations.

The data from the female voices suggest that they are less sensitive to the various simplifications than the male voice. Yet, this may be due to the somewhat poorer quality of the synthetic references, and there is no conclusive evidence due to the scarcity of data.
On the other hand, the fact that it is possible to apply the simplifications to the female voices with no loss of perceived naturalness suggests that this kind of simplification is a valid procedure.

PERSPECTIVES

The insights gathered from this study is presently being implemented in our TTS system.
One goal is to refine the female voices to attain a level of maturity comparable to the male voice. At that time it will be possible to replicate the listening test with the female voices, thus elaborating on the results which have been presented in this paper.

REFERENCES

[1] P. Holtse, H. Nielsen, and J. Rischel, "A Survey of the Speech Synthesis Project", *Copenhagen Working Papers in Linguistics*, Vol. 1, pp. 145-52, 1990/91.

[2] B. Bagger-Sørensen, O. Bertelsen, P. Dømler, C. Henriksen, P. Holtse, P. Molbæk Hansen, H. Nielsen, N. Reinholt Petersen, and J. Rischel, "A Text-to-Speech System for Danish", In Torres, L.; E. Masgrau and M. A. Lagunas (eds.), *Signal Processing V, Theories and Applications*, Vol. 2, Elsevier, pp. 1119-22, 1990.

[3] P. Molbæk Hansen, "The Linguistic Components of the Danish Text-to-Speech System", *Copenhagen Working Papers in Linguistics*, Vol. 1, pp. 153-62, 1990/91.

[4] C. Henriksen, and N. Reinholt Petersen, "The Phonetic Rule System", *Copenhagen Working Papers in Linguistics*, Vol. 1, pp. 163-170, 1990/91.

[5] D. H. Klatt, "Software for a Cascade/Parallel Formant Synthesizer", *J. Acoust. Soc. Am.* 67, pp. 971-95, 1980.

[6] B. Bagger-Sørensen, O. Bertelsen, and P. Dømler, "The Speech Synthesizer", *Copenhagen Working Papers in Linguistics*, Vol. 1, pp. 181-86, 1990/91.

[7] G. Fant, J. Liljencrants, and Q. Lin, "A Four-Parameter Model of Glottal Flow", *Speech Transmission Laboratory, Quarterly Progress and Status Report* 4, pp. 1-13, 1985.

[8] I. Frehr, M. Elmlund, and H. Nielsen, "Improving the Spectral Balance of Digital Speech Synthesis Applied to a Female, Synthetic Voice", *Proc. 3rd European Conference on Speech Communication and Technology*, Vol. 3, pp. 1915-18, 1993.