SYNTHESIS OF A FEMALE VOICE, A PRELIMINARY STUDY

Inger Karlsson[1]

Abstract: The ultimate aim of my work is to synthesize a female voice with the help of our OVEIII synthesizer. I have started this work by surveying the existing literature where only three articles describing experiments in this area could be found, none of which claimed to have been successful. We copied one of those experiments and also tried to copy one utterance made by a female speaker. None of these synthesized samples were acceptable as a female voice.
To gain more knowledge about the perceptual load of different parameters we ran a test where the fundamental frequency, the bandwidths, and the higher formants were varied. The listeners seemed to be aware of differences in average fundamental frequency that were greater than 10 Hz and were more sensitive to changes in the first formant than in the higher formants.

Experiments with female voice synthesis

The ultimate goal of my research is to produce a fully acceptable synthesis of a female voice. Little seems to have been done in this area. I have found only three relevant articles: U. Goldstein (1972), H. Sato (1974), T. Yasuhiro and K. Ozeki (1976). None of those claimed to have been wholly successful. Anyhow, we decided to try to copy the most promising of these studies, namely the one by Yasuhiro and Ozeki. They have used LPC technique, see B. Atal and S. Hanauer (1971), the output rate being set to 1.3 times the input rate, thus increasing the formant frequencies and bandwidths by 30%, and the fundamental

---

1) Department of Speech communication, Royal Institute of Technology (KTH), S 100 44 Stockholm 70, Sweden

frequency being set to 2.1 times the original value. A sentence from a reading made by a male speaker was used as test material. The processed utterance was passed through a comb filter with zeros at 900, 2800, 4700 and 6600 Hz to get a more female-like result, and this way, according to the authors, "...an almost satisfactory female voice ..." was achieved. We tried to copy this experiment with the R. Carlson and B. Granström (1976) rule synthesis program, the formant frequencies and bandwidths being raised by the same amount as in the Yasuhiro and Ozeki experiment, and the resulting synthesis being filtered as described above. Examples of this synthesis, both filtered and unfiltered, were played at the seminar; neither of these examples had any great resemblance to a female voice. One possible explanation of this is that the relations between male and female formant frequencies and bandwidths are nonuniform, see G. Fant (1966) and O. Fujimura and J. Lindqvist (1971), whereas in this experiment they were assumed to be uniform. We therefore synthesized the same sentence with the formant values set to typical average values for female speakers, but this did not remarkably improve the "femaleness" of the synthesized voice. We have also tried to make a synthetic copy of an utterance made by a woman. The result can be seen in Fig. 1, where spectrograms of the original and the synthesized copy are shown. These samples were also played at the seminar. Though a similarity could be heard between the two, the similarity was not so much in voice quality as in prosody, and the synthesized voice could hardly be accepted as a female voice.

In all the examples mentioned above the parameters altered to get a more female voice were formant frequencies, bandwidths, and fundamental frequency. As we could ascertain by listening to the synthesized speech, these parameters do not carry all information that is needed to unambiguously define a certain sex. Further parameters to be taken into consideration are those of the voice source, as well as variations within and between sounds of the parameters that we used in the experiments described above.

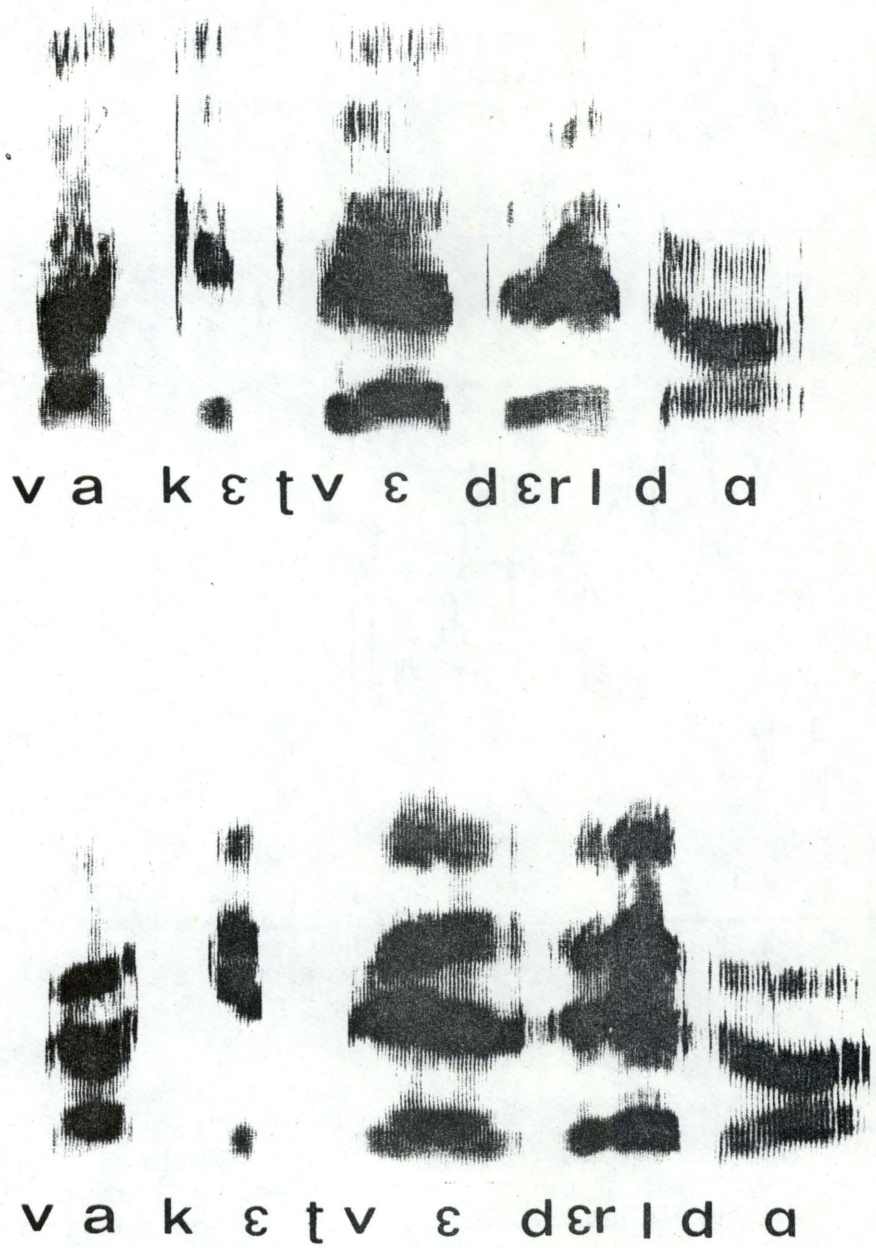

Figure 1

Upper part human voice, lower part synthetic copy of the utterance: "Vackert väder idag".

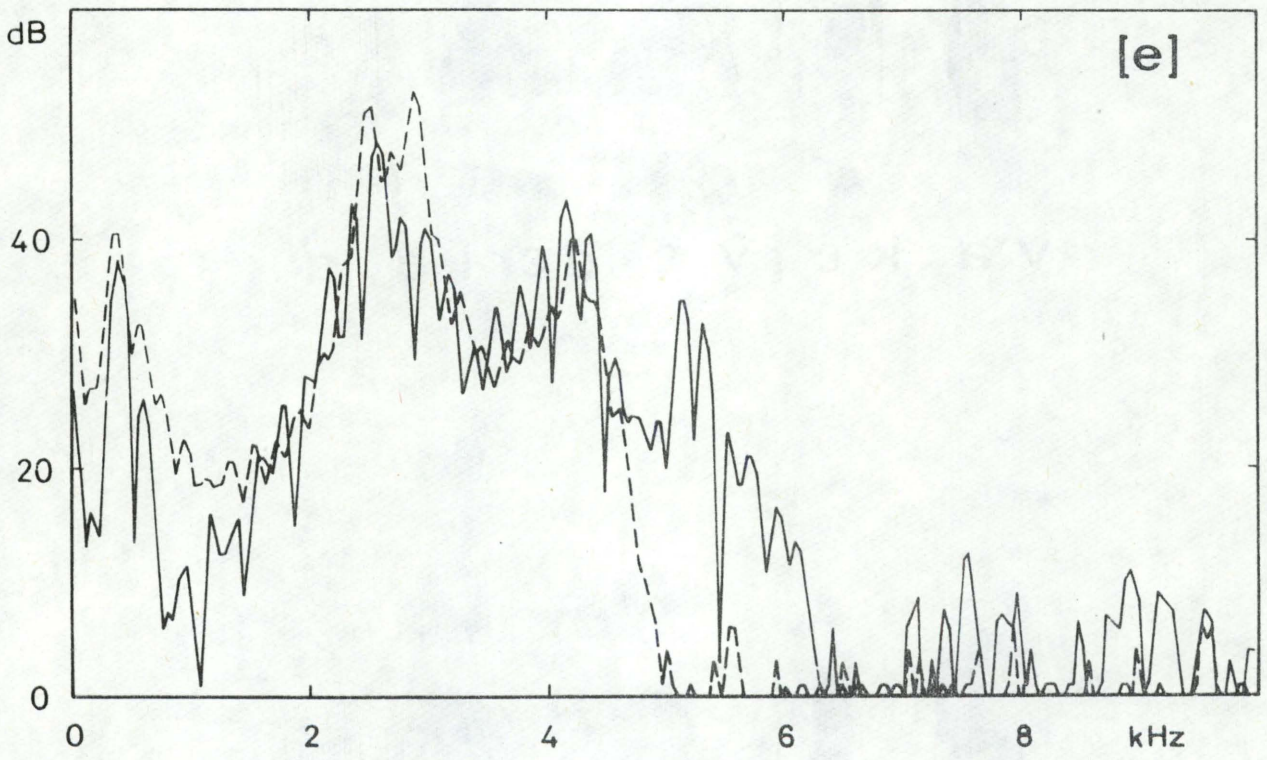

Figure 2

Spectrum sections of natural and synthetic speech. The spectrum of the natural speech is drawn with a solid line, the broken line indicates the synthetic speech.

Listening test

We have also tried to estimate the relative importance of different parameters for the perception of similarity of sounds. To do this we used a listening test where the test participants were asked to judge the similarity between two syllables using a scale ranging from 5 for full similarity to 1 for no similarity. The first syllable was the same throughout the whole test: a syllable pronounced by a woman. The second was a synthetic copy of the first where the frequency and amplitude of the fourth formant, the average value of the fundamental frequency, the bandwidths of the first and of the second + third formants were varied, one at a time. The other parameters were held at values that, as judged by spectrograms, were as similar to those of the first syllable as possible. Spectrum sections of the natural vowel and of the synthetic copy that was used as a reference for the different parameters are shown in Fig. 2.

The results of the test are given in Fig. 3. We can see that the fundamental frequency has to be altered at least 10 Hz before the similarity rating decreases. For the fourth formant an 8 dB decrease in amplitude does not influence the result, whereas a 500 Hz decrease in frequency shows only a weak tendency towards lower ratings. That a frequency increase gives such high rating values probably depends on the fact that the speaker we used in the test has a strong fifth formant. The bandwidth of the first formant seems to be an important parameter, whereas the listeners seemed to be unaware of changes in the second and third formant bandwidths. Due to the type of synthesizer we used (a terminal analog synthesizer) it is impossible to tell whether it is really the differences in bandwidths that are perceived; it may just as well be the amplitude differences. According to Kakusho et al. (1971), the ear seems to be fairly sensitive to amplitude differences, especially if these are located in the first formant. This suggests that voice quality depends much on

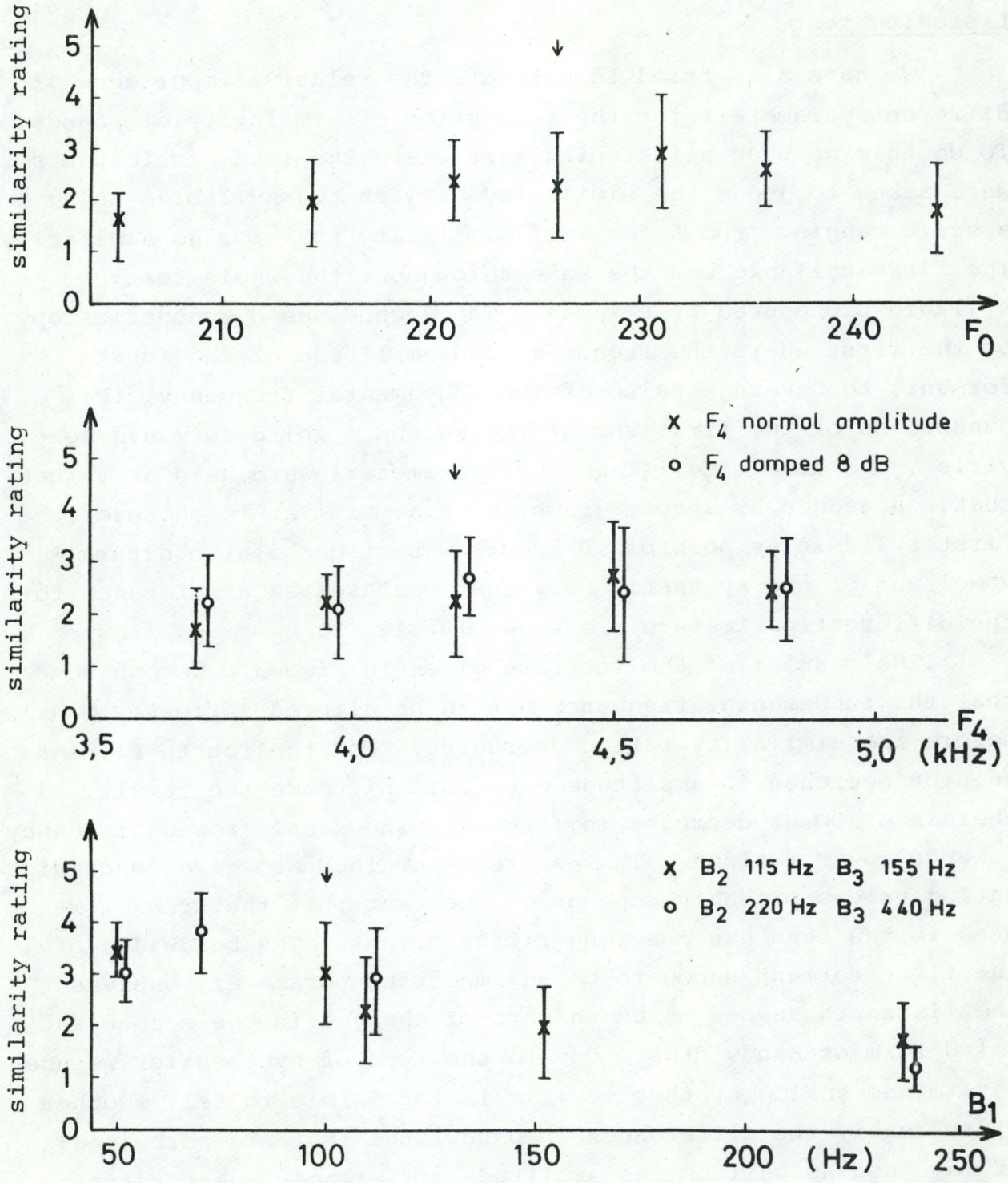

Figure 3

Results of the listening test. Mean values and standard deviations of the similarity ratings are shown. The reference sample that was judged the most similar to the natural speech sample from spectrograms is indicated with an arrow.

the lower parts of the spectrum, the fundamental frequency and the first formant region. It is interesting to compare with the K.Ågren and J.Sundberg (1976) study of tenor and alto voices: they found that the differences between those two groups can be ascribed, for the most part, to a stronger fundamental and a higher, +400 Hz, fourth formant for the alto voices. Combined with our results this would indicate that the next thing I ought to do to improve the female voice synthesis would be to raise the amplitude of the fundamental.

References:

Atal, B.S. and S.L. Hanauer 1971: "Speech analysis and synthesis by linear prediction of the speech wave", JASA 50, p. 637-655

Carlson, R. and B. Granström 1976: "A text-to-speech system based entirely on rules", Conf. Records, 1976 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Philadelphia USA, p. 686-689

Fant, G. 1966: "A note on vocal tract size factors and non-uniform F-pattern scalings", STL-QPSR 4, p. 22-30

Fujimura, O. and J. Lindqvist 1971: "Sweep-tone measurements of the vocal tract characteristics", JASA 49, p. 95-100

Goldstein, U. 1972: Comparisons of terminal-analog synthesis of male and female voices, Unpublished thesis, MIT

Kakusho, O., H. Hirato, K. Kato and T. Kobayashi 1971: "Some experiments of vowel perception by harmonic synthesizer", Acoustica 24, p. 179-190

Sato, H. 1974: "Acoustic cues of female voice quality", Electronics and Communication in Japan 57-A, p. 29-38

Yasuhiro, T. and K. Ozeki 1976: "An experiment on male to female voice conversion", J. of the Acoustical Society of Japan 32, p. 362-368

Ågren, K. and J. Sundberg 1976: "An acoustic comparison of alto and tenor voices", STL-QPSR 1, p. 12-16

My talk was followed by a discussion over the topic: Can you judge the sex from the voice? In this discussion I cited the following articles:

Ingeman, F. 1968: "Identification of the speaker's sex from voiceless fricatives", JASA 44, p. 1142-1144

Lass, N., K. Hughes, M. Bowyer L. Waters and V. Bourne 1976: "Speaker sex identification from voiced whispered and filtered isolated vowels", JASA 59, p. 675-678

Weinberg, B. and S. Bennett 1971: "Speaker sex recognition of 5- and 6-year-old children's voices", JASA 50, p. 1210-1213