First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

A FEMALE VOICE FOR A TEXT-TO-SPEECH SYSTEM.

Inger Karlsson

Department of Speech Communication and Music Acoustics,
KTH, Box 70014, S-100 44 STOCKHOLM, SWEDEN

ABSTRACT

There is a great need for better voice quality in text-to-speech systems. Today, only mechanically sounding male voices can be produced. The lack of success in producing a better sounding voice quality has been due mainly to a lack of knowledge of the voice source. We have also needed a good voice source model and an implementation of such a source in a text-to-speech system. The LF-model for the voice source has given us a tool for a description of the voice source dynamics in speech. The implementation of this source model in our text-to-speech system raises opportunities for synthesis with better voice quality and with different voices. In this paper the work on a female voice is described. The voice source dynamics in sentences and in different stress environments are studied. Acoustic parameters for a female reference speaker are compared to the male synthetic voice. These data are compiled into rules for synthesis and the results of these rules will be played at the conference.

INTRODUCTION

There is a great need for a better voice quality in text-to-speech systems. This is particularly true for systems that are used as voice prostheses, were users want a personal voice. Especially, women and children would like to sound as women and children. Earlier tries have not been very successful, with the exception of Klatt's attempts to synthesize female vowels [1]. The lack of success has been due mainly to a lack of knowledge especially of the voice source and also a lack of a good model for the source and an implementation of such a source in a text-to-speech system. The four-parameter voice source LF-model for voiced sounds proposed by Liljencrants and Fant [2] has recently been implemented in our text-to-speech system. This has given us an opportunity to improve the synthetic voice quality and to synthesize different voices. In this paper the work on a female voice for the synthesizer is described. This includes a voice source description, acoustic parameters for different speech segments and prosody rules.

The voice source model

A condensed description of the properties of the LF voice source model is given in figure 1. Here the influence on the voice source spectrum by three of the four parameters that are presently used in the text-to-speech system are shown. The fourth parameter is the amplitude of the excitation spike which controls the amplitude of the generated sound.

Rules for the text-to-speech system are presently written for modelling female speech using the new glottal source. The voice source model is also utilized in attempts to give a description of the voice source in real speech [3].

ANALYSIS OF FEMALE SPEECH

Voice source

The speech materials for the voice source description consists of a few complete sentences, some with focal stress on different words, and isolated vowels and syllables. So far one female speaker's production of the materials has been analysed. This speaker is from now on called the female reference. The

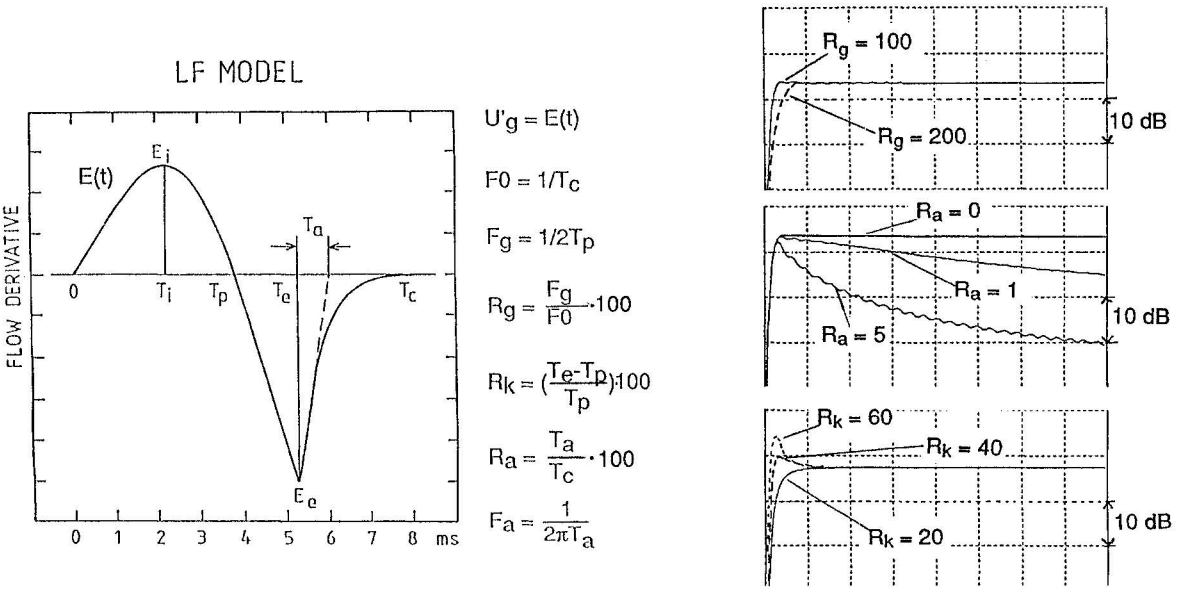

Figure 1. The LF model for the voice source pulse. To the left the flow derivative is shown together with the parameters. In the right part, the impact of variations of some parameters on the spectrum of the second derivative of the flow are shown.

results from this study on one voice have been compared to a study of two-word sentences spoken by seven women, whose voices were classified [4]. The female reference participated in this study and was judged to have a "normal, somewhat tight, sonorous" voice quality.

The speech materials were recorded with a preservation of a correct phase response even at the lowest frequencies, thereby minimizing the risk of distorting the glottal pulse shape. The speech signal was inverse filtered with an interactive filtering program. The bandwidth of the inverse filtered speech signal was 20-4000 Hz. A model source pulse was fitted to each voice pulse, also using an interactive program. This gave the parameters for the voice source model.

Speech segments

Speech materials consisting of nonsense words and syllables uttered by the female reference were used for this analysis. The materials contained both stressed and unstressed vowels and also consonants in different vowel and stress contexts. It was not the same as the materials used for the voice source study. The first four formants in vowels and voiced consonants were measured from broadband spectrograms. Spectral sections for all sounds were matched with a spectral section of a synthetic sound where the synthesis parameters were varied to give a best match. In this way both voice- and noise-exited formants and their bandwidths were obtained together with source amplitudes for nasal, fricative, aspirated and voiced segments.

Prosody

The intonation patterns for the sentences used for the voice source study were registered. This gave some examples of sentence intonation and of F0 movements and range in focally stressed words. Segment durations have not been studied, it is for this study presumed that the differences between a normal male and female speaker are negligible.

RESULTS

Voice source parameters

Dynamic variations within a sentence: In all utterances the termination is signaled by the voice source: E_e decreases, R_a is raised to 20% and R_k to 60% and R_g comes close to 100%, see figure 2. Within an utterance, two origins of dynamic variations can be detected, one is related to the segment itself, the other to coarticulation. Voiced consonants have, as a rule, an up to 10 dB lower E_e and also a slightly higher R_a, 2-3% higher, than the vowels as can be seen in figure 2. Voiced /h/ and voiced plosives tend to have higher R_k than vowels, R_k for these segments are often 50% as compared to about 35% for vowels. Voiced segments preceding a voiceless sound, an unvoiced plosive or a fricative, end with a considerably raised R_a, 20%, and R_k, 60%, and a low E_e, as can be seen in figure 4. The results for the female reference conforms well with Gobl's results for male speakers using the same speech materials [5].

In isolated vowels and syllables R_a was found to be higher for a more front/closed vowel. The higher R_a value was combined with a slightly higher F0, the resulting F_a was still higher for a more front/closed vowel, as demonstrated in figure 3. The mechanism behind this difference remains to be explained but it could have the same origin as intrinsic pitch differences.

Variation with stress: The same four-word sentence was read with focal stress either on the second, third or fourth word. The third word from the three versions were specially studied and the voice source parameters are shown in figure 4. Here a tendency towards a lower R_a can be seen in the stressed vowel and a higher R_a in the following /l/ in the focally stressed word (together with a F0 minimum in the stressed vowel). There is also a tendency towards a stronger excitation, E_e, in the focally stressed word.

Speech segments

The mean value for the third formant in the open vowels was used to decide an approximate vocal tract length for the female reference. The mean value was 14% higher for the female reference than for the synthetic male voice, which indicates that this female speaker has a vocal tract that is about 14% shorter than the synthesis reference male speaker. This is slightly lower than the average male-female difference [6].

The values of the first two formants for the female speaker are plotted together with the male synthetic voice in figure 5. The male-female differences conforms well with the differences described in [6]. The formant bandwidths in the vowels for the female reference speaker were often higher than in the synthesis, the first formant bandwidth was in many vowels 100-140 Hz as compared to 50-70 Hz for the male synthetic voice. The frequency differences between energy maxima in voiceless fricatives are smaller than 14%. This could be expected since they mainly depend on the front cavity and the main difference in vocal tract length between men and women is to be found in the pharynx.

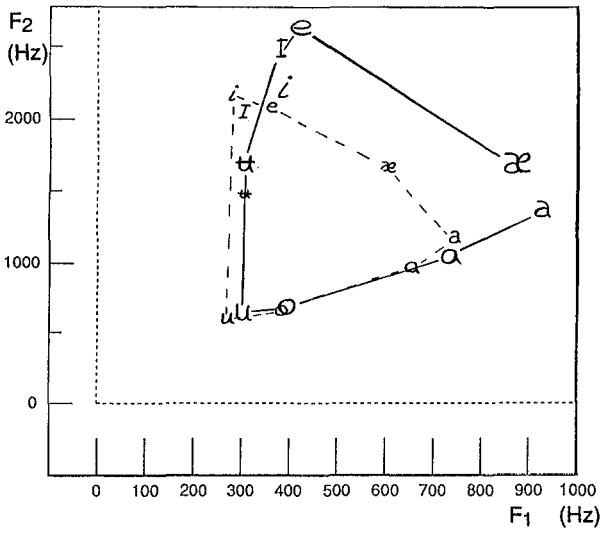

Figure 5. Vowel spaces and some examples of vowels for the female reference, solid line and large symbols, and the male synthetic voice, dashed line and small symbols.

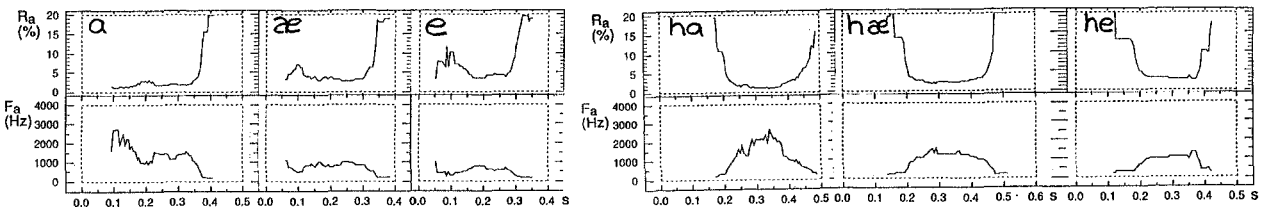

Figure 3. R_a and F_a for the vowels [e], [æ] and [ɑ].

Prosody

The fundamental frequency curve for a whole sentence is found in figure 2. In figure 4 the F0 curve for the same word in different stress positions are found. In the sentence the F0 varies between peak values of 280 Hz for stressed vowels and a lowest value of 140 Hz in the last vowel. Notable is that F0 is raised towards the very end of the sentence, the same phenomenon was found in other sentences for this speaker and also for other speakers in an earlier study [4]. The word carrying focal stress, figure 4, is by the female reference speaker designated by a low F0 in the stressed vowel, 165 Hz, and high F0 in the next vowel, 280 Hz. When focal stress is put on the preceding word F0 starts at a high value, 325 Hz, and falls through the word to 160 Hz. For the same word followed by a focally stressed word F0 is fairly constant and about 200 Hz, which is about the average F0 for the speaker. The average F0 for the female reference speaker is about 0.9 octaves higher than F0 for the male synthetic voice, which is a normal male-female difference. The F0-range in the sentence for the female reference is about one octave. The same sentence produced by the synthesis system shows a similar F0-range in octaves.

CONCLUSION

The analysis results are now being included in the text-to-speech system. Formant values and corresponding bandwidths are used for defining the sound inventory for a female voice. Rules for voice source and fundamental frequency behavior are being compiled. A first impression is that, together with the voice source parameters R_a and R_k, the formant bandwidths for the lower formants are very important factors for synthesizing a female voice. The work on analysis and rule implementation is continuing and further results will be presented at the conference.

ACKNOWLEDGEMENTS

This project has been supported in part by grants from the Swedish Board for Technical Development (STU) and Swedish Telecom

REFERENCES

[1] D Klatt, "Detailed spectral analysis of a female voice", J Acoust. Soc. Am. Suppl.1, Vol 80, S97: 1986

[2] G Fant, J Liljencrants and Q-G Lin, "A Four-parameter Model of Glottal Flow," STL-QPSR 4, pp.1-13: 1985.

[3] R Carlson, G Fant, C Gobl, B Granström, I Karlsson and Q-G Lin, "Voice source rules for text-to-speech synthesis", Proc. ICASSP-89, Vol.1 pp.223-227: 1989

[4] I Karlsson, "Glottal Waveform Parameters for Different Speaker Types," Proc. of SPEECH 88, 7th FASE Symp., Edinburgh, pp.225-231: 1988

[5] C Gobl and I Karlsson, "Male and female voice source dynamics", to be published in Proc. of Vocal Fold Physiology Conference, Stockholm: 1989

[6] G Fant, " Non-uniform vowel normalization", STL-QPSR 2-3, pp.1-19: 1975

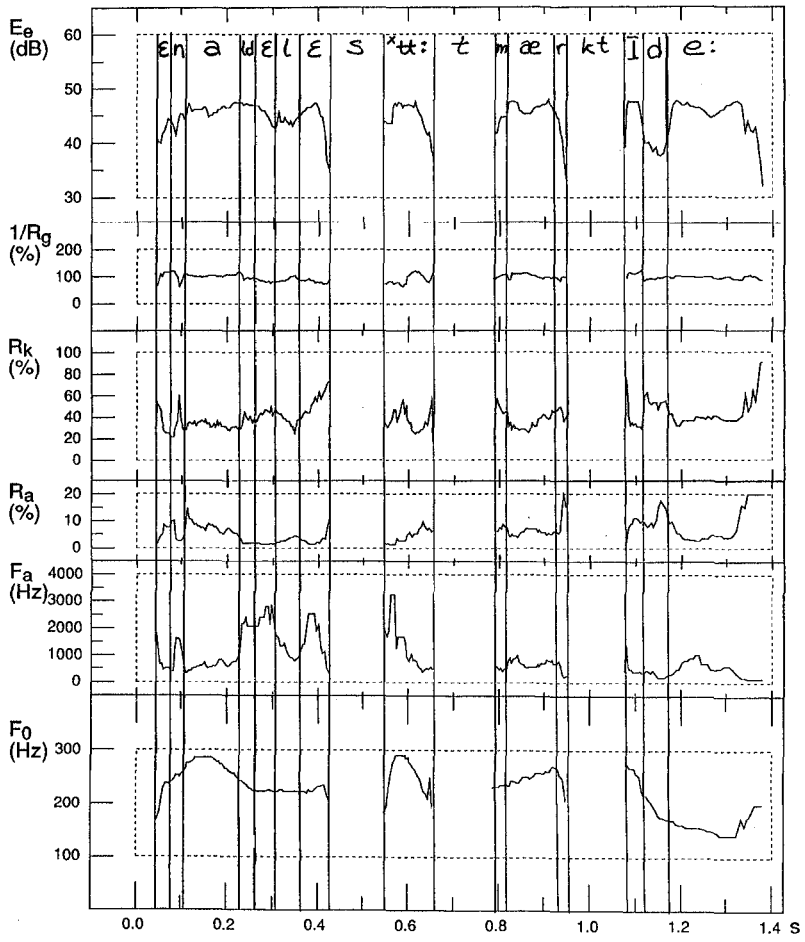

Figure 2. Voice source parameters and F0 for the sentence "En alldeles utmärkt idé." uttered by the female reference speaker.

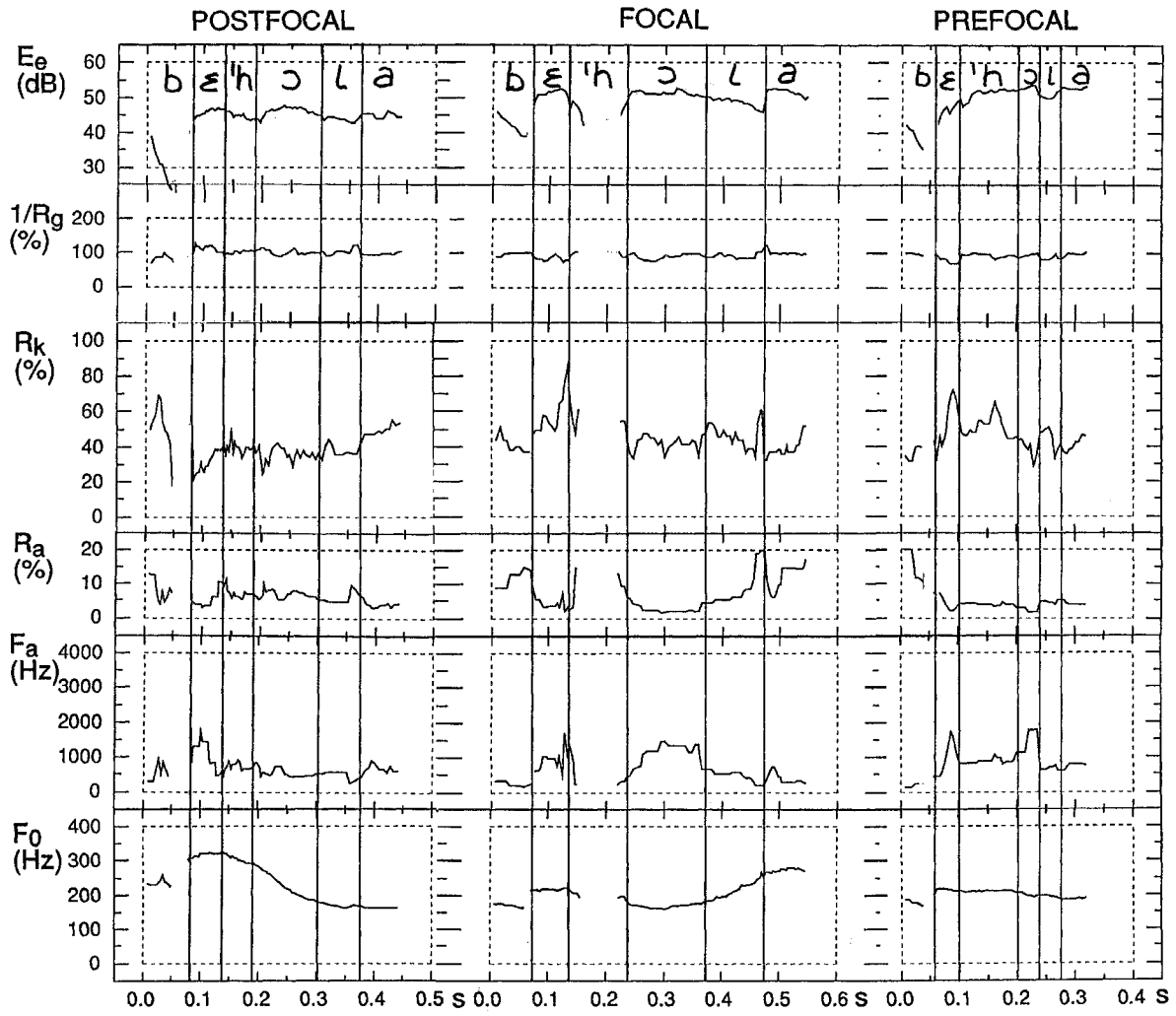

Figure 4. Voice source parameters and F0 for the word "behålla" in different sentence positions relative to the focal stress.