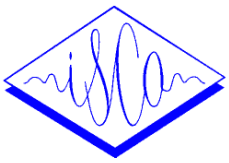

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

CONSONANTS FOR FEMALE SPEECH SYNTHESIS

Inger Karlsson

Dept of Speech Communication and Music Acoustics, KTH,
Box 70014, S-100 44 Stockholm, Sweden

ABSTRACT

This study reports on some results from an ongoing project to systematically describe consonants in female speech. All Swedish consonants and consonant allophones have been recorded in nonsense words read in a carrier phrase. The acoustic properties for the consonants are measured by inverse filtering and by matching natural speech with synthetic speech. The results are tested, using the KTH speech synthesis system. The aim of the project is to produce natural-sounding female text-to-speech synthesis.

INTRODUCTION

Text-to-speech systems using formant synthesisers can not demonstrate a convincingly female voice today. An explanation for this is the lack of methodical descriptions of the acoustics of female speech. Vowels have been fairly thoroughly investigated and also been compared to vowels uttered by male speakers, but our knowledge of the acoustics of consonants and of prosody is more superficial. This paper reports on a continuing project consisting of a systematic acoustic analysis of Swedish consonants uttered by normal female speakers. The aim of the study is to supply descriptions of consonants suitable for producing synthetic female voices in text-to-speech synthesis using a GLOVE formant synthesiser [1]. Both source and vocal tract features have been studied.

Mixed voice and noise excitations of the vocal tract have received special attention. Attempts have been made to separate the two types of excitation. Mixed excitation obviously occurs in voiced fricatives, it also occurs in voiced consonants even for languages lacking true voiced fricatives, like Swedish. It has also been claimed to be an attribute of female speech. A proper balance between voice and noise excitation has a large effect on the naturalness of a synthetic voice.

SPEECH MATERIAL

The speech material consisted of nonsense words read in a carrier phrase. The nonsense words were built on the pattern /bet'VCVt/. They were pronounced with Swedish accent 1, that is, only one syllable was stressed. The consonants always occurred after the stressed vowel. The duration of the VCV-sequence was about 250 ms. The consonant durations were 50-100 ms. The vowels in the nonsense words were either a short /a/, /i/ or /u/. In the words that have been investigated so far, both vowels were the same. All Swedish consonants and also some allophonic variations were recorded. The recordings were made in an unechoic chamber using a condenser microphone and a DAT tape recorder. The recorded digitised speech was transferred directly from the DAT tape to computer memory. All measurements were made using software developed at our department. Two female speakers have recorded the material. The two speakers were chosen because their voices were different but normal. The same speakers have served as subjects in earlier investigations as well, [2,3]. The data reported in this paper are obtained from one speaker.

THE VOICE SOURCE

A recent implementation of a more realistic voice source in the KTH text-to-speech system, an expansion of the LF-model, [4], has made it possible to synthesise different voices. The LF-model is also used for parametric descriptions of inverse filtered speech. In the LF-model the voice source pulse is defined by four parameters plus F0. The parameters RK, RG, FA and EE are chosen for the parametric description of the analysed voice pulses and for synthesis. RK corresponds to the quotient between the time from peak flow to excitation and the time from zero to peak flow. RG is the time of the glottal cycle divided by twice the time from zero to peak flow. RG and RK are expressed in percent. EE is the excitation strength in dB. FA is the frequency above which an extra -6 dB/octave is added to the spectral tilt. RG and RK influence the amplitudes of the two to three lowest harmonics. FA determines the high frequency content of the spectrum and EE controls the overall intensity. The LF-model usually gives a good approximation to the natural voice for voiced speech segments. However, when the vocal tract is excited by a mixture of harmonic and noise energy, additional parameters are needed. In the new version of the synthesis, a parameter, NA, has been introduced to control the mixing of noise into the voice source. The noise is added according to the glottal opening and is thus pitch synchronous. The NA parameter models noise generated at the glottis. The spectral properties of the new voice source and the possibility of dynamic variations of these parameters makes in easy to test the results that have been obtained in the analysis.

MEASUREMENT METHODS

Formants in consonants and formant transitions between consonants and adjacent vowels have been studied. Spectral sections of the unvoiced consonants have been matched by synthesis. This yields both source and vocal tract resonance parameters specially suited for our synthesis program. The method is described closer in [1]. The voiced consonants and the transitions between vowels and consonants were studied using inverse filtering. The inverse filtering was performed using an interactive filtering program. Preliminary formant frequencies and bandwidths were calculated automatically using the Linear Prediction auto correlation method. The formants and bandwidths were then finely tuned by hand, using both time and frequency representations of the speech wave before and after filtering. The LF-model was fitted to the inverse filtered voice pulse by hand. Traces of formants and voice source parameters for about 1 sec of the speech signal were shown on the computer screen.

We chose not to use fully automatic inverse filtering as the methods developed so far are not always successful. These methods are supposed to determine the formants during the closed phase of the glottal cycle. Accordingly, these methods work particularly well for open, long vowels uttered with a fairly low-pitched voice where the closed phase is fairly long and well defined. The closed phase of the glottal pulse can be very short for a high pitched voice, though. It is also often very short in consonants and in aspirated articulations. A very short closed phase will, in the automatic methods, result in great

10.21437/ICSLP.1992-127

uncertainty and variability in the calculated parameters. In some methods, where a minimal length of the closed phase is presumed, the parameter values will be chosen so that this hypothesis is fulfilled. The hypothesis may, accordingly, decide the length of the closed phase for segments with a very short closed phase. The aim of this study is to obtaining voice source data for synthesis on consonants uttered by female speakers. For these purposes, interactive methods have been found to be necessary. Automatic fitting of the LF-model to the inverse filtered wave has also been considered. The fitting was performed in the time domain. It was found hard to define a criterion for a good fit of the return phase. A small misfit for this parameter in the time domain will create a large difference between the model and the natural voice pulse in the frequency domain. This difference would be clearly discernible audibly. The return phase, the time between excitation and complete closure of the vocal cords, is called TA in the LF-model. It determines the value of FA, the spectral tilt parameter.

The sampling frequency for the speech signal was 16 kHz. Normally, seven formants were identified and cancelled within the bandwidth of the signal (8 kHz). To attain a good inverse filtered voice pulse, it was necessary to cancel extra pole/zero pairs for many consonants and adjoining parts of vowels. These pole/zero pairs can have different origins. In nasals and laterals they are due to the geometry of the vocal tract. Both nasal and lateral articulations contain shunting cavities in the vocal tract. Voiced fricatives contain pole/zero pairs that have their origins in the cavity behind the frication source. In aspirated articulations, the vocal cords never close. Accordingly, the cavities below the glottis are not acoustically isolated from the vocal tract. This will create pole/zero pairs similar to what occurs in leaky or breathy voices. Formant frequencies and bandwidths for each voice pulse was obtained from the inverse filtering.

RESULTS

Inverse filtering

Inverse filtering is performed in the middle of the vowels and the consonants, and also for the transition between vowels and consonants. Voice source values for some consonants are given in Table 1. In the table, only mean values are given. For most consonants, the parameter values show small variations within the segment. The consonant /h/ is the exception among the consonants shown here. For /h/, both FA and RK varied considerably from one period to the next. The mixed excitation, /h/ was strongly aspirated, is presumably the explanation for these variations. The excitation, EE, is stronger for the nasals than for the other consonants. It is about equal to EE for vowels. The EE value for the other consonants is only slightly lower.

The different vowel contexts are shown separately to check for possible context influence on the voice source parameters. For the consonant samples shown here, /l/ has a higher FA value and lower RK value in /a/-context and /j/ has a lower FA value and RK value in /i/-context. This may be only normal variations, but it will be studied further.

Some data on dynamic variations of voice source parameters are given in Table 2. Here VCV-sequences containing the consonants /n/ and /l/ are reported. Comparing the influence of the two consonants, we find considerably lower FA values for the vowels before and after the nasal consonant. This is due to the nasalization of the vowels, but the low FA values are found even when no nasal zero or formant is visible in the spectrum. The /n/ excitation, EE, is at least equal in strength to the preceding stressed vowel. EE for the first voice pulse in /l/ is often considerably lower, 3-5 dB, than EE for either the preceding vowel or the rest of the /l/.

The difference in FA for different vowels that has been reported earlier [2], is found here as well. In the same context, FA is higher for the open /a/ and lower for the close, front /i/. The EE value does not differ much between different vowels.

Table 1. Voice source parameters for consonants in different vowel contexts. FA and F0 are given in Hz, RK and RG in %. EE, the excitation strength, is given in uncalibrated dB. As the different utterances were recorded at the same session, comparisons within and between segments and utterances can be done.

Consonant	Vowel context	FA	RK	EE	F0	RG
/n/	/a/	260	35	59	255	110
	/i/	240	35	62	245	105
	/u/	260	45	63	245	115
/l/	/a/	500	35	58	260	110
	/i/	320	50	57	255	110
	/u/	300	45	59	260	110
/m/	/a/	320	44	59	245	105
	/i/	270	35	62	250	105
	/u/	250	45	65	235	100
/h/	/a/	300	50	60	255	95
	/i/	350	50	55	235	110
	/u/	400	45	56	250	105
/j/	/a/	500	50	52	230	110
	/i/	350	35	52	255	110
	/u/	500	40	57	235	105

Table 3. Frequency and bandwidth data for spectral zeros for consonants in VCV-sequences. The values are the mean taken over about 5 periods. The zeros are indicated by Z followed by a number and the corresponding bandwidth with BZ and the number. The zeros are numbered according to frequency, a lower frequency value is indicated by a lower number.

Consonant	Context	Z1	Z2	BZ1	BZ2
/h/	/a/	1090	2310	200	200
	/i/	1050	1510	200	500
	/u/	1040	3650	250	900
/j/	/a/	1040		450	
	/i/	1300		400	600
	/u/	1350	3700	300	700
/l/	/a/	980	2280	250	300
	/i/	1240		600	
	/u/	1020	1540	300	200
/m/	/a/	850	3280	250	400
	/i/	860	3030	200	500
	/u/	1060		400	
/n/	/a/	910	2610	150	500
	/i/	980		200	
	/u/	1000	2300	180	300

Spectral Zeros

Zero/pole pairs were cancelled for some consonants and adjacent vowels. Data on spectral zeros in consonants are given in Table 3 and for vowel-consonant transitions in Table 4. The zeros found in /h/ and in the vowels adjoining /s/ are due to the open glottis and the consequent coupling to the subglottal cavities. The frequency values for the zeros, about 1050 Hz and 2300 Hz, are slightly higher than the values given for female speakers by Klatt and Klatt, [5], 900 Hz and 2200 Hz. These differences are most likely due to personal variations. The zeros identified in the other consonants are depending on the articulation of these consonants. The variations between different vowel contexts can be caused by small differences in articulation. Further investigations are needed to explain the variations.

The spectral zero that occurs in the vowels preceding /s/ is visible mainly in the last 5-6 pulses. The bandwidth of this zero decreases rapidly. This is, combined with a fall in FA, an important contribution to the fall in total intensity of the vowel. The excitation energy, EE, is only slightly reduced. The nasal zeros in vowels preceding /n/ occurs about halfway into the vowel and is visible in the following vowel for about as long. As the passage through the mouth is open, the frequency of the nasal zero is different from in /n/. The frequency of the spectral zero in vowels varies considerably with vowel articulation.

Formant frequencies

Formant frequencies and bandwidths have been collected for all segments that have been inverse filtered. The dynamic variations of these data will be used for speech synthesis rules. These rules will be demonstrated by synthesis at the conference. Some data on formants in nasal consonants are included in this paper, Table 5. The lowest formant in all samples has a rather high value, especially for the /a/ context. The high energy in the lowest harmonics in nasals is modelled by voice source parameter RG in the voice source description.

Mixed excitations.

One aspect of pronunciation that it is necessary to investigate further is the noise content in voiced consonants. The noise usually have two different origins. In aspirated voiced consonants, the noise originates at the glottis. It is due to incomplete closure of the vocal cords. One example of such consonants in Swedish is an inter-vocalic /h/. Frication noise occurs in very constricted articulations, for example when the tongue is very close to the palate. In the Swedish language, only two voiced fricatives occur, /v/ and /j/. The fricative noise content in these consonants is always small in Swedish. In inter-vocalic position, as in the present investigation, /v/ and /j/ is often articulated without frication. Different measurement methods have been suggested for estimating the noise content. A comparison between the energy content in the harmonics and the energy content between the harmonics have been suggested by Kasuya and Ando [6] Lee and Childers [7] propose that a similar comparison is made for frequencies above 2 kHz. Klatt and Klatt [5] found a significant correlation between perceived breathiness and the amplitude of the first harmonic compared to the rest of the spectrum for an excerpted vowel segment. These methods are primarily aimed at classifying different voices. They have been shown to correlate well with perceived breathiness of vowels. For our purposes, a description of consonants and vowel - consonant transitions, where rapid changes in F0 can occur, other methods are needed. We are presently trying to estimate the amount of noise in a voice pulse by comparing the spectrum of the inverse filtered voice pulse with the spectrum of an FL-model pulse, Figure 1. This method will give information that is easily incorporated in our text-to-speech system. Figure 1 shows an example of frication noise in /j/. The noise energy is visible over 2.2 kHz, where the LF-model spectrum and the spectrum of the inverse filtered speech differ considerably. To ascertain that the energy above 2.2 kHz consist mainly of noise, a partly inverse filtered voice pulse, with F3 remaining was studied. As can be seen i Figure 1, F3 was excited mainly by noise.

Table 4. Frequency and bandwidth data for spectral zeros in vowels measured from the periods adjacent to a consonant. The values are the mean taken over at least 5 periods before and after the consonant. The zeros are indicated by Z followed by a number and the corresponding bandwidth with BZ and the number. The zeros are numbered according to frequency, a lower frequency value is indicated by a lower number.

Consonant	vowel	Z1	Z2	BZ1	BZ2
/s/	/a/	1100	2320	300	450
	/I/	1050		400	
	/u/	1040		300	
/n/	/a/	990	2300	250	600
	/I/	1040		400	
	/u/	1560	2300	450	500

Conclusion

The validity of the acoustic descriptions that have been achieved in this study are tested using the GLOVE synthesiser. Synthetic stimuli demonstrating the results will be played at the conference.

ACKNOWLEDGEMENT

This project has been supported by a grant from the Swedish Council for Planning and Co-ordination of Research.

REFERENCES

[1] Carlson R., Granström B. Karlsson I. (1990): "Experiments with voice modelling in speech synthesis" , *Speech Communication*, vol 10, pp. 481-490

[2] Karlsson I. (1990): "Voice source dynamics for female speakers" *Proceedings of the 1990 International Conference on Spoken Language Processing*, Kobe, pp.69-72.

[3] Karlsson I.(1991): "Dynamic voice quality variations in natural female speech", *Proc of XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, pp. 4:10-13

[4] Fant, G., Liljencrants, J & Lin, Q. (1985): "A four-parameter model of glottal flow," *STL-QPSR* 4/1985 pp.1-13.

[5] Klatt, D., and Klatt, L. (1990): "Analysis, synthesis and perception of voice quality variations among female and male talkers". *J.A.S.A.* Vol. 87, pp. 820-857

Table 5. Formant frequencies for the nasals /m/ and /n/. The data are taken from voice pulses where one spectral pole/zero-pair is cancelled. The nasal formant is included.

Vowel context	F1	F2	F3	F4
Consonant /m/				
/a/	520	1180	1800	2860
/i/	370	1270	1800	3620
/u/	270	1210	1850	3100
Consonant /n/				
/a/	405	1270	1880	3070
/i/	380	1170	2000	3000
/u/	360	1330	1810	3030

Table 2. Voice source parameters for VCV-sequences. FA and F0 are given in Hz, RK and RG in %. EE, the excitation strength, is given in uncalibrated dB. As the different utterances were recorded at the same session, comparisons within and between segments and utterances can be done.

	FA	RK	EE	F0	RG
Consonant /n/					
Vowel /a/					
vowel before 700	30	60	235	110	
end of vowel	275	40	59	240	110
consonant	260	35	59	250	110
start of vowel	450	30	59	260	90
Vowel /i/					
vowel before	460	35	60	235	110
end of vowel	250	40	60	240	110
consonant	240	35	62	245	105
vowel after	350	35	61	260	95
Vowel /u/					
Vowel before	600	52	61	240	115
End of vowel	350	45	62	245	115
consonant	260	45	63	245	115
vowel after	300	45	59	245	110

	FA	RK	EE	F0	RG
Consonant /l/					
Vowel /a/					
vowel before	1500	37	63	230	110
end of vowel	800	50	61	240	120
consonant	500	35	58	260	110
vowel after	1200	40	56	260	110
Vowel /i/					
vowel before	650	33	60	250	110
end of vowel	600	30	60	250	105
consonant	320	50	57	255	110
vowel after	600	50	58	265	125
Vowel /u/					
vowel before	800	45	61	255	115
end of vowel	500	50	60	260	120
consonant	300	45	59	260	110
vowel after	600	50	60	270	130

[6] Lee C.-K. and Childers, D. (1991): "Some acoustical, perceptual, and physiological aspects of vocal quality," in *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, ed.: J Gauffin and B Hammarberg, Singular Publishing Group Inc., San Diego, pp. 233-242

[7] Kasuya, H., and Ando, Y. (1991): "Acoustic analysis, synthesis and perception of breathy voice," in *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, ed.: J Gauffin and B Hammarberg, Singular Publishing Group Inc., San Diego, pp. 251-258

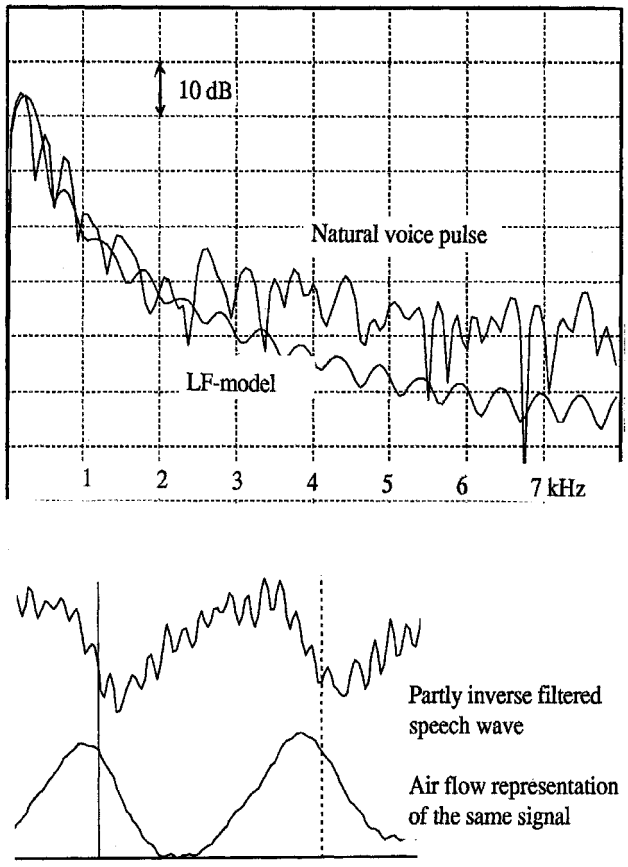

Figure 1. Above: Spetra of inverse filtered voice pulse from the consonant /j/ and an LF-model fit to the voice pulse. Note the difference in energy above about 2.2 kHz.

Below: The same voice pulse in the time domain. Here the third formant at about 3 kHz is not cancelled. This formant is mainly excited by noise.