Acoustic description of consonants in female speech; voiced consonants.

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

This paper contains some acoustic data from a study of Swedish consonants pronounced by two female speakers. The analysis has been performed using inverse filtering, which has given data on both the voice source and formants and bandwidths. In addition, the occurrence of pole-zero pairs in consonants have been investigated. The origin of these poles and zeroes are also discussed.

INTRODUCTION

This paper forms part of a study of the acoustic properties of Swedish VCV sequences in female speech. The aim of the study has been to acquire a comprehensive description of both voice source and vocal tract properties of all Swedish consonants uttered by female speakers. Some data from this study have been published earlier (Karlsson 1992). In this paper, the descriptions will be restricted to voiced consonants and transitions between vowels and voiced consonants. Very little data concerning the acoustic properties of consonants uttered by female speakers have been published so far. This is especially true for voice source data.

The descriptions of consonants and vowel-consonant transitions are intended to be used in composing rules for synthesis of natural sounding speech. The validity of the measured data have continuously been tested by using the latest versions of the KTH text-to-speech system (Carlson, Granström & Karlsson 1991, Karlsson & Neovius 1994)

SPEECH MATERIAL

The speech material consists of nonsense words read in a carrier phrase. The nonsense words are composed according to the pattern /bet'VCVt/. They were pronounced with Swedish accent 1. The vowels in the nonsense words were either a short /a/, /i/ or /u/. The studied 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 material has been read by two female speakers, one with a normal and the other with a leaky voice quality. The recordings were made in an anechoic chamber, using a condenser microphone and equipment that did not effect the phase of the signal. The recorded signal was sampled with 16 kHz sampling frequency. Resulting bandwidth after filtering was 25-4000 Hz.

INVERSE FILTERING

Data on voiced speech segments were achieved by inverse filtering of the speech wave. The inverse filtering was performed using an interactive filtering program. Preliminary formant frequencies and bandwidths were calculated automatically using the LPC autocorrelation method. The formants and bandwidths were then finely tuned by hand for one fundamental period at a time, using both time and frequency representations of the speech wave before and after filtering. The LF voice source model (Fant, Liljencrants

and Lin 1985) was fitted to the inverse filtered voice pulse by hand. This procedure resulted in a parametric description of the voice source and also gave formant frequencies and bandwidths. The analysis was carried out in the middle of the consonants and vowels and during the transition between vowel and consonant with a few exceptions. Sometimes very rapid formant transitions occur between adjacent phonemes. These transitional segments can not be properly analysed by inverse filtering as the formant movements are large even within a voice pulse

The inverse filtering technique presupposes that it is possible to completely separate the voice source and the influence on the speech wave from the vocal tract resonances. This is not true, but for the present purposes it was judge to be a good enough approximation. In real speech, there are always some interactions between the voice source and the sub- and supra-glottal cavities. These effects can not be completely removed. In the present study pole-zero pairs were sometimes introduced to take care of some of these effects.

The two female speakers that were analysed were both rather tall. As the female vocal tract is considerably shorter, for the present speakers it was judged to be about 15% shorter, than the male vocal tract, only 7 formants were estimated within the frequency range for the inverse filtering, 25 to 8000 Hz. The inverse formant frequencies and bandwidths were decided to get a smooth closed phase. As a second criterion a smooth spectrum for the whole voice pulse was used. This second criterion was used especially in aspirated or 'leaky' speech segments, where a closed phase does not exist.

RESULTS

For many consonants and adjoining parts of vowels it was necessary to introduce extra pole/zero pairs as well as formants to attain a good inverse filtered voice pulse. These pole/zero pairs can have different origins. In nasals and laterals they are due to the geometry of the vocal tract (Nord 1976, Chafcouloff 1985). In Table 1, formant and pole-zero data for the lateral /l/ in three vowel contexts are given. The lowest zero, Z1, has a frequency near the lowest zero in /h/, see Table 3. This indicates that this zero has its origin in the subglottal cavities. The second zero, Z2, is much lower in /u/-context than in /i/- or /a/-context which presumably is due to a more retracted place of articulation for /l/ in /u/-context. Data on the nasal /n/ are given in Table 2. The values for the lowest pole-zero pair as well as for the first formant are very hard to decide using the inverse filtering as they are all very close together. The measurement accuracy is not high for these values and they need to be checked by some other method, for example sweep tone measurements.

In aspirated articulations, the vocal cords never close completely. 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 (Cranen and

Table 1: *Formants, F, poles, FP, and zeros, Z, and their bandwidths, B, BP, BZ, for /l/ in three different vowel contexts. All values are given in Hz.*

	F2	F3	FP1	FP2	B2	B3	BP1	BP2	Z1	Z2	BZ1	BZ2
context												
a	1190	2650	1760	3470	160	100	170	290	960	2250	210	370
i	2490		1680	2910	130		230	210	1130	2310	510	520
u	1280		1850	2880	110		120	200	1040	1570	390	400

Table 2: *Formants, F, poles, FP, and zeros, Z, and their bandwidths, B, BP, BZ, for /n/ in the sequence /ana/. All values are given in Hz.*

FP1	F1	FP2	F2	FP3	BP1	B1	BP2	B2	BP3
300	580	1290	1760	2760	170	110	70	130	700

Z1	Z2	Z3	BZ1	BZ2	BZ3
420	880	2700	810	180	380

Table 3: *Poles, FP, and zeros, Z, and their bandwidths, BP, BZ, for /h/ in three different vowel contexts. All values are given in Hz.*

	FP1	FP2	FP3	BP1	BP2	BP3	Z1	Z2	Z3	BZ1	BZ2	BZ3
context												
a	1820	2920	3700	180	310	850	1120	2310	3340	220	260	160
i	1580			250			1240			260		
u	1650	2910	3860	210	350	190	1100	2150	3310	310	530	720

Boves 1987, Klatt and Klatt 1990). In Table 3, poles and zeros for voiced, aspirated /h/ in different vowel contexts are given. They all originate in the subglottal cavities, as /h/ is articulated with an open glottis. There are only very small differences between the different vowel contexts.

The LF-model approximates the derivative of the glottal flow pulse using four parameters. The parameter EE decides the excitation strength, that is the amplitude. RG decides the balance between the first and second harmonic, and is normally higher in a more tense speaking mode. RK is lower in a tense voice than in a soft voice. FA is typically low for a soft voice and high for a tense voice.

As an example, LF-parameter values for the consonant /v/ and adjacent vowels are given in Table 4. The differences in LF parameter values show a trend that is fairly typical for all voiced consonants investigated so far. The excitation strength, EE, is lower in most consonants as it is in /v/. Notable exception are nasals, where EE is about equal to the preceding vowel. The consonant /l/ shows a characteristic pattern. During the first few pulses, there was an additional 1-3 dB dip in the EE level.

The variation in the RG parameter is always small and no relationship between articulation and RG has been found so far.

The parameter RK was generally higher for the consonants compared to adjacent vowels. The RK values given for /v/ in Table 4 are typical for all consonants. A high RK value means that the first harmonic is comparatively strong.

The parameter FA was lower, that is the return time was longer, in all the consonants investigated as it is for /v/, than in the surrounding vowels. This implies that the voice source in consonants contained relatively less high frequency energy. The lowest FA values were found in the nasals and the voiced plosives. FA was also low in nasalised parts of vowels. There is also a consistent difference between vowels, high front vowels always have lower FA values than open vowels, as can be seen in Table 4. There is a tendency for FA to be lower in unstressed vowels, as can also be seen when comparing the stressed vowel before, with the following unstressed vowel in Table 4.

The data given so far are mean values over a few pulses. For a complete description of vowel-consonant transitions, the timing of the different parameter changes is important. For example, the nasal zeros in vowels preceding /n/ occur about halfway into

Table 4: *LF-parameter values for the consonant /v/ and adjoining vowels. EE is given in uncalibrated dB, FA in Hz, RG and RK in percent*

Vowel /a/				Vowel /i/				Vowel /u/			
Vowel before											
FA	RK	EE	RG	FA	RK	EE	RG	FA	RK	EE	RG
1600	43	64	108	450	38	61	114	850	46	62	115
End of vowel											
FA	RK	EE	RG	FA	RK	EE	RG	FA	RK	EE	RG
900	45	62	110	450	42	60	115	450	45	60	108
Consonant											
FA	RK	EE	RG	FA	RK	EE	RG	FA	RK	EE	RG
450	50	55	115	350	55	55	109	280	55	58	108
Following vowel											
FA	RK	EE	RG	FA	RK	EE	RG	FA	RK	EE	RG
1200	35	62	100	600	45	60	127	400	50	58	105

the vowel and are visible in the following vowel for about as long. When the passage through the mouth is closed, the frequencies of the nasal zeros change very rapidly from vowel to nasal value. The spectral pole/zero pair that occur in vowels preceding fricatives and voiceless stops, and that is due to the opening of the glottis, is visible mainly in the last 5-6 voice pulses. Their bandwidths decrease rapidly towards the end of the vowel. This is, combined with a lowering of FA, typical for most vowel-to-unvoiced consonant transitions in Swedish. It explains most of the fall in the total intensity that occur in the last part of the vowel, the excitation energy, EE, is only slightly reduced. The timing of the parameter changes in the vowel-to-consonant transitions will be discussed further in future reports.

ACKNOWLEDGEMENTS

This work has been supported by FRN and the European CEC-ESPRIT project Speech Maps.

REFERENCES

Carlson R., Granström B., Karlsson, I. (1991): "Experiments with voice modelling in speech synthesis", *Speech Communication* 10, 481-490

Chafcouloff, M., (1985): "The spectral characteristics of the lateral /l/ in French", *Travaux de l'Institut de Phonétique d'Aix*, Vol. 10, 63-98

Cranen, B. and Boves, L. (1987): "On subglottal formant analysis"; *JASA* 81, 734-746

Fant, G., Liljencrants, J., Lin, Q. (1985): "A four-parameter model of glottal flow.", *STL-QPSR* 4/1985, 1-14.

Karlsson, I. (1992): "Consonants for female speech synthesis", *Proc. of 1992 Int. Conf. on Spoken Language Processing*, Vol. 1, 491-494

Karlsson, I., and Neovius, L. (1994): "VCV-sequences in a preliminary text-to-speech system for female speech ", *STL-QPSR* 1/1994 (forthcoming).

Klatt, D., and Klatt, L. (1990): "Analysis, synthesis and perception of voice quality variations among female and male talkers". *JASA* 87, 820-857

Nord, L. (1976): "Experiments with nasal synthesis", *STL-QPSR* 2-3/1976, 14-19