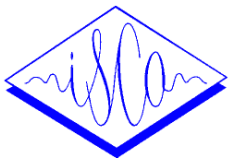

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

# VOICE SOURCE DYNAMICS FOR FEMALE SPEAKERS.

Inger Karlsson

Department of Speech Communication and Music Acoustics,
Royal Institute of Technology (KTH),
Box 70014, S-100 44 Stockholm, Sweden

## ABSTRACT

Dynamic variations of the voice source in ordinary speech have been studied by means of inverse filtering of the sound pressure wave. The inverse filtered voice pulses have been matched by the LF-model to achieve a parametric description. Special attention was given to the spectrum of the pulses.

The speech material consisted of sentences and vowels uttered by three female speakers judged to differ in voice quality. The voices were judged to be normal and their quality ranged from somewhat tight and sonorous to thin.

The variation of the voice source parameters with place and manner of articulation will be discussed. The length of the closing time in the voice pulse is shown to vary with vowel height; a more open vowel shows a shorter closing time. This tendency seem to be true for the three different voices studied even though the range can vary with voice type. Consonants have been studied, especially concerning the influence of spectral zeros and the transition between consonants and vowels.

Variations of the different glottal parameters within sentences have also been studied in relation to voice type.

## INTRODUCTION

Voice source dynamics in normal speech have been studied for three female speakers. The results from these studies are summarized to give a description of voice source behaviour for normal speakers with different voice qualities. So far, the investigated speech consists of isolated vowels and sentences read aloud from a list. The speech signal have then been inverse filtered and a model of the voice source, the LF-model [1], have been fitted to the inverse filtered speech. This gives a parametric description of the voice source which facilitates comparisons between speakers and between different segments of speech. One application of the results is the production of a more human voice that can speak with different voice qualities in our speech synthesis system [2,3].

## SPEAKER DESCRIPTIONS

Three female speakers have been investigated so far. They have earlier been classified by a speech therapist according to voice quality and also investigated to decide their degree of constant leakage, [4].

The degree of leakage in the voices were decided using a Rothenberg mask to record the oral air flow, [5] . Examples of the syllable /pɑ:/ have been inverse filtered for each speaker to obtain the glottal air flow. After inverse filtering the bandwidth of the signal was 0-1200 Hz. The peak flow and the constant leakage flow in the most closed period of the glottal cycle, called dc flow, have been measured from the inverse filtered signal, see Table 1. The mean over four consecutive fundamental periods has been calculated. Care was taken not to measure during the beginning of the vowel where articulatory movements can cause air flow.

| Speaker | dc flow | peak flow | dc/peak flow |
|---|---|---|---|
| W1 | 60 | 200 | 0.30 |
| W2 | 100 | 220 | 0.45 |
| W3 | 130 | 330 | 0.41 |

Table 1. Glottal air flow in ml/sec for three female speakers. The context, aspirated /p/, partly explains the large amount of dc-flow.

The voice qualities of the three women have been judged by a speech therapist. A reading of an excerpt from a novel, about 1 minute long, was used for the voice classification. All three speakers were classified as normal speakers without voice problems, so the given judgements should be read as if preceded by "a tendency towards". The results were:

Speaker W1: Normal, somewhat tight, sonorous.
Speaker W2: Thin, not breathy, lacking sonority, young, high pitch.
Speaker W3: Dark, slightly coarse, sonorous, swollen vocal cords.

## DESCRIPTION OF THE VOICE SOURCE MODEL

The LF-model is a four parameter model of the voice source pulse, a full description of the model can be found in [1]. The four parameters chosen for the source matching in this study are RK, RG, EE, and FA, see Figure 1. RK corresponds to the quotient between the time from peak flow to excitation and the time from zero to peak flow, (in the differentiated flow in Figure 1 this corresponds to the time from the zero-crossing to the negative peak divided by the time from the start to the zero-crossing). RG is the time of the glottal cycle divided by twice the time from zero to peak flow. RG and RK influence the amplitudes of the two to three lowest harmonics and are expressed in percent. EE is the excitation strength in dB and FA the frequency above which in effect an extra -6dB/octave is added to the spectral tilt. Figure 1 explains the function of these parameters. In addition, the fundamental frequency, F0, is measured.

The LF-model usually gives a good approximation to the natural voice pulse for voiced speech segments. However, when the vocal tract is excited by a mixture of harmonic and noise energy additional parameters would be needed. Effects of voice source - vocal tract interactions as described in [7] are not taken care of either by the model.

## PROCEDURE

The speech material consisted of sentences and isolated vowels uttered by three female speakers. The speech was recorded in an unechoic chamber. The band-width of the recorded signal was 10 to 16 000 Hz and the recording was phase-true. After digitizing and inverse filtering the upper frequency limit was 4 000 Hz. The inverse filtering was done

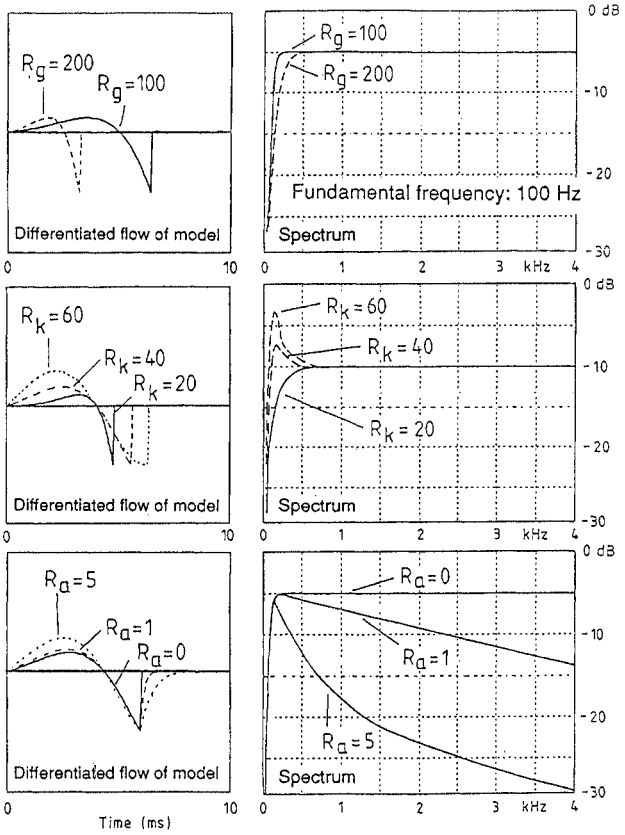


Figure 1. The influence of the parameters RG, RK and FA on the differentiated glottal flow pulse shape and spectrum. From [6]

by hand using an interactive filtering program on an Apollo work station. In the program it was possible to simultaneously study the time wave and the frequency spectrum thereof before and after filtering. The time window for the frequency spectrum was set to one voice period. The same program was used to match a voice pulse created by the LF-model to the inverse filtered voice pulse. Particular care was taken to get a good spectral fit and consequently a good perceptual similarity. An example of inverse filtering and voice source model fitting is given in Figure 2.

## RESULTS

### Voice source in sentences

So far three complete sentences have been analyzed for the speakers W1 and W2. For speaker W3, one of these three sentences plus a shorter version of another have been analysed. The sentence that have been analysed for all three speakers consisted of the two Swedish words, "ja ajö" pronounced /'ja: a'jø:/, an utterance consisting of fully voiced segments. The phoneme /j/ contains next to no noise in this position. Voice parameter values for this sentence will be used to illustrate differences between speakers, see Figure 3. Speaker W2 here shows considerably lower FA-values than the two other speakers, which means that her voice is relatively weaker in the higher frequencies. W2 also have much less variation in FA.

For the two speakers W1 and W2 two further, longer sentences have been analysed. For one of these sentences the FA differences are the same as discussed above, while for the third sentence W2 displays higher FA values and larger variations in FA compared to her other utterances. However, her highest FA values are still lower than W1's. These variations of FA seem accordingly to be one of the differences between a thin and a more sonorous voice quality. For speaker W2, the low FA seems to constitute part of her speech habits. There also is a slight tendency towards lower RK values for W1 in all sentences, this is presumably related to the impression of a slight tightness in her voice.

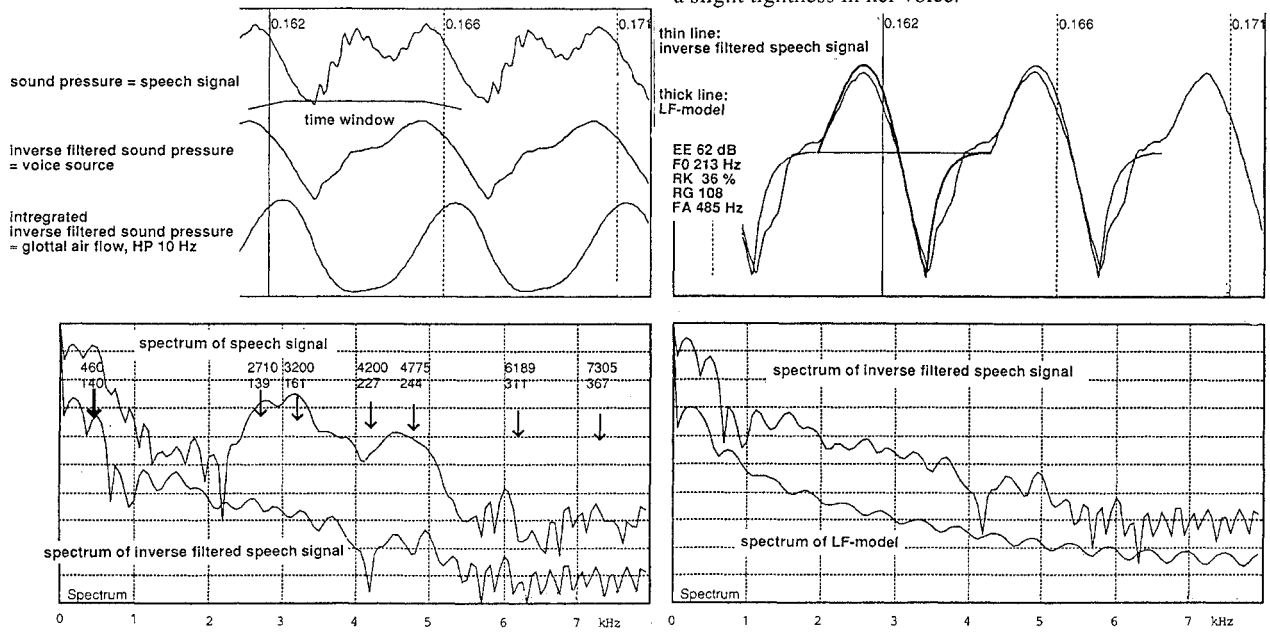


Figure 2. Inverse filtering of a vowel. In the left half, the speech signal before and after filtering is shown. The anti-formant filters are indicated by arrows and the frequency and bandwidths are given in Hz above the arrow. In the right half, the inverse filtered signal is shown together with the LF-model. The LF parameters are given in the upper part.

### Voice source differences in vowels within and between speakers

Special interest have been placed on the variation of the glottal pulse shape with place of articulation for vowels. Nine long Swedish vowels were read in isolation by speakers W1 and W2. As the vowels were read from a list, they all acquired equal stress except the last and first items. The first and the last three vowels from the list were therefore not analyzed. One example of each of the nine vowels for each speaker was inverse filtered and matched by the model. Some measured parameter values, the mean over five voice pulses about one third into the vowel, are displayed in Table 2. For both speakers FA is higher for more open, low vowels. A probable explanation for this phenomenon is the difference in interaction between vocal tract and voice source, the interaction is larger for a more constricted articulation. The vocal tract inertia increases as the minimum area of the vocal tract decreases. A combination of these two effects can act as a low pass filter of the flow, [7].

For the other voice source parameters, no clear tendencies can be detected, even though the more open vowels seem to have somewhat lower RK-values and slightly shorter open quotients. The differences are very small, though. No correlations could be found between any of the different voice source parameters.

Some speaker differences can be detected. As is found in the sentence material, FA varies more for speaker W1 who also shows higher values for this parameter. Speaker W2 shows a tendency towards higher FA-values for rounded vowels than for their unrounded counterpart which is contrary to what is expected. This might be explained by W2 using less extreme articulations for the rounded vowels but may also be a personal speech habit. The RK-parameter show, if anything, a reverse pattern compared to the sentence data. RK is slightly higher for speaker W1 than for W2.

The tendency for FA to be higher for more open vowels has been investigated in connected speech as well. The speech material used for this was the Swedish sentence "Pia odlar blå violer, gula liljor och röda dahlior" uttered by speakers W1 and W2 and the shorter Swedish sentence "Pia odlar blå violer" uttered by speaker W3. The vowels carrying word stress were investigated and the voice source parameter values are given in Table 3. The vowels are given in the order they appear in the sentence. The same tendency in FA can be seen also here. The close front vowels /i/ and /I/ show the lowest FA values for all three speakers while the back more open vowel /o/ show the highest. The rounded vowels all show high FA values. The fairly low FA-value for the open /ɑ/-vowel for speakers W1 and W2 may be due to its position in the last word in a long utterance.

| Speaker W1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Vowel | i | y | e | ø | æ | ɑ | o | u | ʉ |
| FA | 384 | 390 | 491 | 601 | 840 | 1063 | 467 | 372 | 359 |
| RK | 45 | 45 | 37 | 38 | 38 | 40 | 58 | 53 | 62 |
| F0 | 188 | 184 | 180 | 189 | 171 | 171 | 186 | 190 | 186 |
| OQ | 73 | 66 | 70 | 66 | 56 | 52 | 76 | 68 | 68 |
| EE | 53 | 54 | 49 | 55 | 54 | 53 | 54 | 54 | 53 |

| Speaker W2 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Vowel | i | y | e | ø | æ | ɑ | o | u | ʉ |
| FA | 370 | 510 | 360 | 617 | 602 | 731 | 538 | 521 | 515 |
| RK | 34 | 43 | 41 | 49 | 39 | 38 | 39 | 45 | 40 |
| F0 | 218 | 203 | 204 | 212 | 212 | 212 | 213 | 216 | 224 |
| OQ | 65 | 64 | 71 | 62 | 65 | 63 | 63 | 69 | 64 |
| EE | 60 | 58 | 58 | 61 | 62 | 63 | 59 | 60 | 60 |

Table 2. Voice source parameter values for vowels in isolation. FA and F0 are given in Hz, RK in % and OQ, the part of the voice pulse when the vocal cords are open, in % of the total voice pulse length. EE, in dB, is not calibrated and should accordingly only be used for comparisons within one speaker.

Normally, while inverse filtering vowels, only anti-formant filters corresponding to the vocal tract resonances were used. For one of the three speakers, W3, an additional pole/zero pair had to be cancelled to achieve a good fit to the LF-model. The origin of this pole/zero pair is presumably the subglottal system. Speaker W3 showed a large amount of air leakage during the most closed phase of the voice pulse, see Table 1. This implies a fairly large opening between the vocal cords even in the most closed phase of the voice pulse and a good coupling between the sub- and the supraglottal cavities. The frequency values of the pole/zero pair, a zero at about 800 Hz and a pole at about 1500 Hz compares well with known values for subglottal poles and zeros for women, [8].

The unstressed vowels followed the same pattern as the stressed vowels; the open /a/ vowels consistently showed high FA values while the closed /I/ vowels showed much lower FA values. The differences in FA between stressed and unstressed vowels were found to be much smaller than between vowels with different manner of articulation.

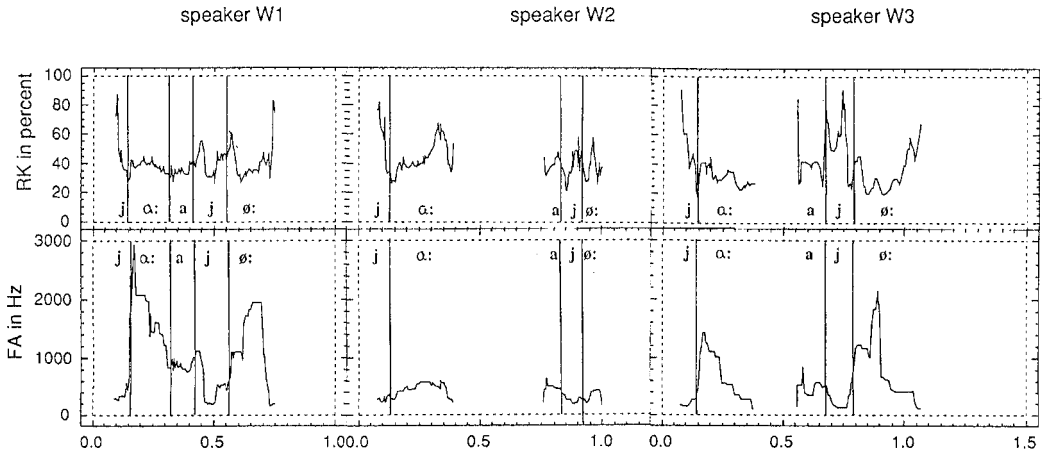


Figure 3. Variation of the voice source parameters FA and RK in a short utterance for three female speakers.

| Vowel | i | u | o | u | ʉ | I | ø | ɑ |
|---|---|---|---|---|---|---|---|---|
| Speaker W1 | | | | | | | | |
| FA | 521 | 889 | 1015 | 907 | 925 | 503 | 503 | 768 |
| RK | 36 | 53 | 44 | 47 | 36 | 43 | 38 | 43 |
| F0 | 224 | 236 | 204 | 178 | 221 | 204 | 202 | 153 |
| OQ | 60 | 55 | 61 | 66 | 61 | 62 | 66 | 65 |
| Speaker W2 | | | | | | | | |
| FA | 422 | 729 | 938 | 809 | 796 | 305 | 877 | 766 |
| RK | 22 | 43 | 46 | 42 | 23 | 40 | 45 | 38 |
| F0 | 260 | 276 | 261 | 213 | 242 | 243 | 236 | 196 |
| OQ | 76 | 63 | 62 | 61 | 56 | 60 | 57 | 73 |
| Speaker W3 | | | | | | | | |
| FA | 375 | 823 | 1275 | 1182 | | | | |
| RK | 32 | 44 | 27 | 50 | | | | |
| F0 | 237 | 219 | 166 | 202 | | | | |
| OQ | 66 | 59 | 58 | 69 | | | | |

Table 3. Stressed vowels in a sentence. For speakers W1 and W2 the sentence is "Pia odlar blå violer, gula liljor och röda dahlior" / ˣpi:a ˣu:dlar 'blo: vI'u:ler ˣgʉ:la ˣlIljUr ɔ ˣrø:da 'dɑ:lIUr/, for speaker W3 "Pia odlar blå violer" / ˣpi:a ˣu:dlar 'blo: vI'u:ler/. FA and F0 are given in Hz, RK in % and OQ, the part of the voice pulse when the vocal cords are open, in % of the voice pulse.

### Voice source in voiced consonants

The investigated sentences also contained voiced consonants: the plosives /b, d, g/, the voiced fricatives /j, v/ that between sonorants both contain very little noise in Swedish, the nasal /n/, the sonorants /l, r/ and voiced /h/. These consonants were also inverse filtered when possible. It was often impossible to inverse filter the stops as these were too weak compared to the background noise. The /r/-phoneme was realized as a vowel-like segment in the studied sentences and had the same voice source characteristics as an unstressed vowel. To get a good fit between the LF-model voice source and the inverse filtered wave-form for the remaining consonants, it was often necessary to cancel an extra pole/zero pair, especially in /n/, /l/ and /v/. As compared to the vowels in the same utterance the consonants showed higher values for RK, that is more energy in the lowest harmonics. The excitation amplitude, EE, was only marginally lower for /r, j, n, h/ than for vowels. For /v/ and the plosives, EE was at least 10 dB weaker. FA showed considerably lower values for all consonants with the exception of /r/, FA was normally found to be only slightly higher than F0. This implies that the voice source contains less high frequency energy for consonants than for vowels.

Attempts have been made to describe the voice source behaviour during transitions between consonants and vowels. In these cases the inverse filter time window was moved forward in time only in very small steps. Typically, the time steps were one fourth of the total voice pulse length. An extra excitation of the vocal tract often occurred during the open phase in these transitional segments. The voice source - vocal tract interaction was also pronounced and the formants sometimes moved very rapidly. It was accordingly almost impossible to achieve a good inverse filtering with the present method.

## CONCLUSIONS

Some results concerning voice source dynamics for three different female voices have been discussed in this paper. The voice source in vowels is shown to vary with degree of opening in such a way that the voice source for the most closed vowels contain less high frequency energy, have a lower FA, than more open vowels. The contextual effects seem to be much smaller. The consonants show even lower FA values in combination with a higher RK value than for vowels. This results in a voice source that contains relatively more energy in the lowest harmonics for consonants than for vowels. Voice quality difference can also be related to voice source behaviour; the speaker that was judged to have a thin voice showed lower FA values than the sonorous voices and she also showed less variations in FA.

The results discussed in this paper have been tested in our speech synthesis system and have in informal listening tests been found to improve the quality of the female synthetic voice considerably.

## ACKNOWLEDGEMENTS

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

References

[1] Fant, G., Liljencrants, J & Lin, Q. (1985): "A four-parameter model of glottal flow", Speech transmission laboratory, Quarterly progress and status report 4/1985, Stockholm, pp.1-13

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

[3] Carlson, R., Granström, B.& Karlsson, I. (1990): "Experiments with voice modelling in speech synthesis." Proceedings of the tutorial and research workshop on speaker characterization in speech technology, Edinburgh 26-28 June 1990, pp. 28-39

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

[5] Lin Q (1987): "Nonlinear interaction in voice production." Speech transmission laboratory, Quarterly progress and status report 1/1987, Stockholm, pp. 1-12

[6] Karlsson, I. (1988): "Glottal waveform parameters for different speaker types." Proceedings of Speech '88, 7th FASE Symposium, Edinburgh 1988 Vol.1, 225-231

[7] Rothenberg M. (1973): "A new inverse-filtering technique for deriving the glottal air flow waveform during voicing.", J. Acoust. Soc. Am. 53, 1632-1645.

[8] Klatt, D. & Klatt, L. (1990): "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am., vol. 87(2), pp. 820-857.