Cries and whispers: acoustic effects of variations of vocal effort

Andersson, A. and Eriksson, A. and Traunmüller, H.

journal: TMH-QPSR
volume: 37
number: 2
year: 1996
pages: 127-130

http://www.speech.kth.se/qpsr
Cries and whispers: Acoustic effects of variations in vocal effort

Anita Andersson, Anders Eriksson, Hartmut Traunmüller
Institutionen för lingvistik, Stockholms universitet

Abstract
This is a preliminary report of an ongoing investigation examining the acoustic effects of the adjustment in vocal effort that is required when the distance between the speaker and the listener is varied over a large range (0.3 m – 187.5 m). Five kinds of characteristics have been studied in the speech of men and women: segment durations, \( F_0 \), formant frequencies, sound pressure level, and spectral emphasis. When the distance increased, the following effects were observed: The durations of vowel like segments increased successively, as distinct from that of most consonants. The mean value of \( F_0 \) increased by a factor close to two, while its SD in semitones remained roughly constant. The levels of voiced segments increased by about 30 dB, but less for unvoiced segments. The increase in level was much more pronounced in the upper than in the lower part of the spectrum.

Introduction
The acoustic properties of speech sounds vary not only because of linguistic factors, but also as a function of the speaker's age, sex, vocal effort, and speech mode (phonated vs. whispered speech). Among these factors, the effects of variations in vocal effort have been least studied. Variation in sound pressure level and in \( F_0 \) have been looked at by several investigators, e.g. Rostolland (1982). It is known that the increase in sound pressure entails an emphasis of the higher frequency components of the spectrum. It has also been reported that formant frequencies increase with increasing vocal effort (cf. Traunmüller, 1988). The duration of vowels was observed to increase with vocal effort and that of consonants to decrease somewhat (Fönagy and Fönagy, 1966; Rostolland, 1982; Bonnot and Chevré-Muller, 1991).

In most previous studies, the subjects were simply asked to adjust their vocal effort. The purpose of the present study is to examine the acoustic effects of variation in vocal effort when this variation is caused by communicative needs as the distance between the speaker and the listener is varied over a large range. Detailed knowledge of the acoustic effects of paralinguistic variation is crucial for increasing the generality of theories of speech production and perception, and it is necessary in order to be able to simulate human behaviour in speech synthesis and automatic recognition of speech, if true tolerance for such variation is desired.

Five kinds of characteristics will be considered: segment durations, fundamental frequency, formant frequencies, sound pressure level, and spectral emphasis.

Method
Subjects
The subjects were six male and six female students and staff members from the department, all speaking Stockholm Swedish. They had no known speech or hearing disorders. They were not singers nor had they been subject to any other voice training.

Speech material
The speech material consisted of a single sentence designed so as to provide good data for measurements of duration, fundamental frequencies, formant frequencies and spectral slope, and level. The sentence used was "Jag tog ett violett, åtta svarta och sex vita, 'I took one violet, eight black and six white'."

Procedure
The recordings took place in an open field in a reasonably quiet area outside Stockholm, using a high quality microphone and a portable DAT-recorder. The microphone was kept at a distance of about 5 cm from the speaker's lips. Experimenter A gave the subjects instructions as to whether the utterance was to be phonated or whispered. Experimenter B asked each subject "Hur många kort tog du av varje färg? 'How many cards of each colour did you take?'", standing at five previously measured distances.
from the subject (see Table I). The subjects were instructed to produce each utterance with the vocal effort considered necessary for the specific distance. For the utterance to pass, experimenter B had to accept it as clearly intelligible.

Due to a mistake, one utterance is missing from one of the recordings, leaving 83 utterances to analyse.

**Table 1. Distances between speaker and listener.** The full range was used for phonated speech. Whispered speech was only used at the two shortest distances.

<table>
<thead>
<tr>
<th>Version</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distance (m)</td>
<td>0.3</td>
<td>1.5</td>
<td>7.5</td>
<td>37.5</td>
<td>187.5</td>
</tr>
</tbody>
</table>

### Acoustic analyses

For the purpose of acoustic analysis, the utterances were digitised at 16 kHz, using 16 bits per sample. In order to remove some of the low-frequency noise that was present in most of the recordings, caused by the wind, the signals were high pass filtered at 70 Hz. For filtering and level measurements, specially designed programs were used. All other acoustic analyses were made using the ESPS WAVES program package.

### Durations

Segment boundaries were marked using broadband spectrograms and oscillograms. Stop occlusions were considered to begin where the formant structure of the preceding vowel could no longer be seen, and to end where the burst began. Vowels were considered to begin at the first positive zero crossing after a consonant or a pause and to end at the beginning of a following stop occlusion.

### Formant frequencies

The analysis of formant frequencies is still at a preliminary stage. In the louder versions, \( F_0 \) was often very high. This will introduce serious measurement errors using any of the conventional formant tracking algorithms. We are at present working on a method which may substantially reduce this error and we will, therefore, not report any results on formant frequency values until measurement accuracy has been considerably improved.

### Fundamental frequency

\( F_0 \) data from each speaker were edited using the probability of voicing parameter so that only data from voiced segments were considered. Sections with creaky voice were also excluded. Mean values and standard deviations were calculated for each individual speaker and for male and female speakers taken together.

### Levels

The overall sound pressure levels of the signals were measured as well as those of selected segment types. The spectral energy distribution was analysed by measuring the signal levels in 10 frequency bands, each with a width corresponding to 3 Bark, shifted upwards in steps of 2 Bark.

### Results and discussion

#### Durations

The mean values obtained for the total duration of the utterances, excluding pauses, were 2693, 2648, 2817, 3158, and 4015 ms for the phonated versions, and 2950 and 3054 ms for the whispered versions. The durations of pauses, vowel-like segments (14 vowels, 1 [j], and 3 [v]), and consonantal segments (8 stop occlusions + VOT, 3 [s] and 1 [l]) are shown in Figure 1.

While the duration of the vowel-like segments increased progressively with increasing distance, the consonants behaved in a different way. Their duration decreased slightly up to a distance of 7.5 m, whereafter they increased slightly.

The increase in the durations of vowels agrees qualitatively with previous observations (Fónagy and Fónagy, 1966; Rostolland, 1982; Bonnot and Chevrie-Muller, 1991). There is no such agreement for the consonants. It may be that subjects asked to increase their vocal effort without any communicative aim tend to vary their vocal effort mainly on the vowels, because the consonants make only a minor contribution to it.

The durations of all segments tended to be slightly longer in the whispered versions than in the corresponding phonated ones.

#### Fundamental frequency

Figure 2 shows \( F_0 \) to increase progressively with increasing vocal effort. The increase in mean \( F_0 \) from the softest to the loudest versions was 15.3 and 11.6 semitones for men and women, respectively. SD, however, remained roughly constant for both men and women, 2.6 and 2.0 semitones, respectively. This is different from variations in liveliness, involvement, and type of discourse, which have been observed to cause the SD of \( F_0 \) to vary (cf. Traunmüller et al., 1989).
Figure 1. Mean value of the total duration of different types of segments in the two whispered and the five phonated versions, as produced by male and female speakers (pooled). Thick lines (+): vowel-like segments. Dashed lines (x): consonantal segments. Thin lines (#): pauses.

Figure 2. Mean values of \( F_0 \) as a function of vocal effort. Thick lines represent male speakers and thin lines female speakers. Dotted lines represent a deviation from the mean values by one standard deviation calculated as the mean of individual standard deviations.

Levels

Sound pressure levels are summarised in Figure 3. The increase in levels as a function of distance was more pronounced for voiced than for unvoiced segments like the three [s]. This can be attributed to the fact that, in unvoiced segments, the level increases as a power function of pulmonic pressure with an exponent of approximately 1.3 (Stevens, 1971). In voiced segments, the amplitude of the volume velocity will increase in proportion to the pulmonic pressure, but there are substantial additional contributions to sound pressure by the increase in \( F_0 \), as well as that in \( F_1 \), and in the spectral emphasis.

The mean levels obtained for the voiced segments in the speech of men and women were almost the same, and this holds for any distance. The various effects on SPL of the physiological differences between men and women thus compensate each other precisely.

There was however, a difference between men and women in their apparent capacity for increasing the SPL in unvoiced sounds such as the three [s]. For these, SPL increased consistently by a smaller amount in the speech of women than in that of men. This may be due to a greater capacity to increase pulmonic pressure in male speakers. This would, then, also explain why \( F_0 \) increased more in male speakers.

The increase in levels did not quite reach the 14 dB SPL that would be required to compensate for the fivefold increase in the distance between speaker and listener.

Spectral emphasis

A preliminary analysis of the emphasis of the spectral components in the upper frequency region revealed substantial differences between voiced and unvoiced segments. The emphasis is much more pronounced for the voiced segments. Since this analysis has not yet been completed, we show only an example of the emphasis observed in voiced segments produced by a single female speaker.

Conclusions

Most of the results agreed with what could be expected. In the light of previous research, it is perhaps surprising that when the distance was increased, (1) there was no general decrease in the duration of consonants, (2) the SD of \( F_0 \) (in semitones) did not increase, and (3) in contrast with the voiced segments, the level of unvoiced fricatives increased less in the speech of women than in that of men.
In addition to the analysis of the relationship between distance and selected acoustic properties, we plan also to analyse the quantitative relationships between all these properties. This will make it possible to simulate variation in vocal effort and to attempt automatic recognition with true tolerance for variation in vocal effort.

Acknowledgements

This research has been supported, in part, by a grant from HSFR, the Swedish Council for Research in the Humanities and Social Sciences, within the frame of the Swedish language technology programme.

References


