Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

A comparative study of the male and female whispered and phonated versions of the long vowels of Swedish

Eklund, I. and Traunmüller, H.

journal: TMH-QPSR
volume: 37
number: 2
year: 1996
pages: 131-134

http://www.speech.kth.se/qpsr
A comparative study of male and female whispered and phonated versions of the long vowels of Swedish

Ingegerd Eklund and Hartmut Traunmüller
Institutionen för lingvistik, Stockholms universitet

Abstract

Confusions in vowel quality and in speaker sex among whispered and phonated versions of the long vowels of Swedish have been analysed. The recognition rate was higher than in other studies. The recognition of vowel quality was observed to interact with that of speaker sex in the whispered versions, but not in the phonated ones. The paper also reports on $F_0$, $F_1$, $F_2$ and $F_3$, and their dynamics, and on the overall spectral shape of the vowels. The observed upward shift of the lower formants in whispering as well as the spectral level differences agree with those found in other languages. (This is a summary of a paper with the same title, submitted to Phoneica)

Introduction

There are large differences in $F_0$ associated with between-speaker differences in age and sex, and with paralinguistic and prosodic information. Listeners presented with whispered speech are, nevertheless, able to perceive most of the distinctions that are normally cued mainly by $F_0$, albeit with reduced precision.

There have been only a few studies in which the properties of whispered vowels have been compared with the phonated counterparts. The published formant frequency data show the frequency positions of the first three formants to be higher in whispered vowels than in the phonated versions (Peterson, 1961; Kallai and Emanuel, 1984a, b). It has been suggested that whispered speech is articulated with a more open configuration of the vocal tract (Traunmüller, 1988), as observed in shouted speech (Schulman, 1989).

Schwartz and Rhine (1968) have shown that listeners are able to discriminate speaker sex in isolated whispered presentations of [i] and [u]. Although the voice fundamental, $F_0$, has been shown to be the dominant cue for this distinction (Lass et al., 1976), the frequency positions of the formants constitute a cue of similar perceptual weight in whispering.

The present study includes an acoustic analysis of the whispered and phonated versions of the nine long vowels of Swedish, as produced by adult male and female speakers. In addition, the perceptual confusions have been studied in these four sets of vowels, with respect to both phonetic quality and speaker sex.

Method

The speech material was produced by 6 male and 6 female speakers of Stockholm Swedish. It consisted of the names of the nine letters which are used to represent the vowels in orthography. They were read three times with phonation and three times whispering in the following order: i j i j, y j y j, u h u h, o h o h, e l e l, o h o h, a l a l, a h a h. The last vowel $/a/$ was added as a dummy in order to avoid any phrase final phenomena. The duration of the vowels was about 0.5 s.

The recordings were made in an anechoic chamber. The speech signal was low pass filtered at 6.3 kHz before being digitised at 16 kHz using 16 bit/sample. Each token was stored in a separate file.

The productions of one male and one female speaker were discarded since they had produced some of their ‘whispered’ vowels with vocal fold vibration. There remained in total 540 tokens (3 productions of 9 vowels in 2 modes by 10 speakers) to be identified by a panel of listeners and to be subjected to acoustic analysis.

The panel of listeners consisted of 10 subjects, 5 male and 5 female. All of them had grown up in the Stockholm area, and reported having normal hearing.

The identification experiment was run individually using a specially designed response collection program. The stimuli were presented in randomised order through headphones. The subjects had to identify the vowel and the sex of
the speaker and to mark one of the response alternatives shown on a screen.

Results and Discussion

Identification experiment

The phonated vowels were identified correctly in 95.9% and 94.7% of all cases, for male and female speakers, respectively. For whispered vowels, the corresponding figures were 86.9% and 89.1%.

The numbers of correct identifications obtained with the phonated and with the whispered vowels are listed in Tables 1 and 2, which give an overview of the recognition of vowel identity as well as speaker sex. The tables also give a rough overview of the confusions.

Table 1. Summary of the confusion analysis, male speakers.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Sex</th>
<th>Phonated</th>
<th>Whispered</th>
</tr>
</thead>
<tbody>
<tr>
<td>too high</td>
<td>wrong</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>right</td>
<td>wrong</td>
<td>17</td>
<td>104</td>
</tr>
<tr>
<td>too low</td>
<td>wrong</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>too high</td>
<td>right</td>
<td>38</td>
<td>55</td>
</tr>
<tr>
<td>right</td>
<td>right</td>
<td>1278</td>
<td>1069</td>
</tr>
<tr>
<td>too low</td>
<td>right</td>
<td>17</td>
<td>98</td>
</tr>
</tbody>
</table>

Table 2. Summary of the confusion analysis, female speakers.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Sex</th>
<th>Phonated</th>
<th>Whispered</th>
</tr>
</thead>
<tbody>
<tr>
<td>too high</td>
<td>right</td>
<td>30</td>
<td>83</td>
</tr>
<tr>
<td>right</td>
<td>right</td>
<td>1257</td>
<td>1124</td>
</tr>
<tr>
<td>too low</td>
<td>right</td>
<td>42</td>
<td>50</td>
</tr>
<tr>
<td>too high</td>
<td>wrong</td>
<td>0</td>
<td>31</td>
</tr>
<tr>
<td>right</td>
<td>wrong</td>
<td>21</td>
<td>87</td>
</tr>
<tr>
<td>too low</td>
<td>wrong</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

As expected, confusions occurred most frequently between vowels that are close to each other in the vowel space defined by F1, F2 and F3.

As can be seen in Table 3, the performance in vowel recognition was substantially better than in the studies by Kallail and Emanuel (1985) who used isolated, sustained English vowels and by Tartter (1991) who used [hVd] syllables.

Our results demonstrate that isolated vowels can be identified very well when they happen to be genuine words. Only words occur naturally in isolation. The hypothesis that the recognition of vowels is facilitated by the presence of consonants (Shankweiler et al., 1977) finds no support.

Table 3. Recognition rates (in percent) for vowel quality in three investigations including whispered versions.

<table>
<thead>
<tr>
<th>Investigation</th>
<th>Phonated</th>
<th>Whispered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kallail and Emanuel</td>
<td>82.5</td>
<td>64.0</td>
</tr>
<tr>
<td>Tartter</td>
<td>92.0</td>
<td>81.6</td>
</tr>
<tr>
<td>Present</td>
<td>95.3</td>
<td>88.0</td>
</tr>
</tbody>
</table>

The performance for speaker sex was also better (98.6% and 91.0%) than observed previously by Lass et al. (1976) in a similar experiment (96.0 and 75%).

Our results show that whispered vowels have a better chance of being identified correctly when speaker sex is recognised correctly, and vice versa (see Table 4). The vowel confusions occurring in cases in which speaker sex was misperceived were biased in the expected direction.

Table 4. Interactions between sex and vowel confusions in whispered vowels.

<table>
<thead>
<tr>
<th>Condition</th>
<th>n</th>
<th>Confused sex (%)</th>
<th>Confused vowel (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex right</td>
<td>2456</td>
<td>5.2</td>
<td></td>
</tr>
<tr>
<td>sex wrong</td>
<td>244</td>
<td>25.0</td>
<td></td>
</tr>
<tr>
<td>vowel right</td>
<td>2512</td>
<td>7.3</td>
<td></td>
</tr>
<tr>
<td>vowel wrong</td>
<td>188</td>
<td>32.4</td>
<td></td>
</tr>
</tbody>
</table>

However, in phonated vowels, misperception of speaker sex did not increase the risk of misperceiving vowel quality, nor vice versa.

The results tell us that listeners are largely successful in extracting the phonetic information from the speech signal, normalising (Rosner & Pickering, 1994) or demodulating (Traunmüller, 1994) it in an appropriate way. However, the contribution of F0 was not so large in this process. It has been suggested that F0 might contribute indirectly to the perception of vowel quality in that it provides a cue to speaker sex (Johnson, 1990). The absence in our data of any vowel confusions in cases in which speaker sex was confused is an argument against this hypothesis.

Acoustic analyses

Before the acoustic analysis was performed, all stimuli were high-pass filtered at 70 Hz in order to remove some low level low frequency noise. Vowels that had not been identified correctly by a majority of listeners were considered as misproduced rather than as misperceived, and they were not included into the acoustic analysis.
**Frequency measurements**

The digitised speech signals were subjected to LPC-analysis, using 18 reflection coefficients. A Hamming window with a width of 100 ms was centred at a point 5% into the duration of the vowel and it was subsequently moved forward in steps of 10% of the total duration of each vowel. The formant frequencies obtained were edited manually when necessary.

The average values obtained for F3 were 109 Hz for men and 206 Hz for women. This is similar to Fant (1959), but in distinction from that study, we obtained clear intrinsic pitch differences similar to those observed in other languages (Whalen and Levitt, 1995). The formant frequency values obtained for phonated speech were more similar to those measured by Karlsson and Stålhammar (1972) than to those by Fant (1959).

Figure 1 shows tracings of the first three formants. For both sexes, the formant frequencies tended to be higher in the whispered vowels than in their phonated equivalents.

Figure 2 shows the average difference in log(f) (base 10) of F1, F2, and F3, between the whispered and the phonated versions as a function of log(F0). There is a highly significant negative correlation between Δ log(F0) and log(F3). When the data obtained by Kallail and Emanuel (1984a, b) are plotted in this way, the result looks very similar. We obtained no significant difference between men and women in this matter. An analogous comparison of female and male data shows no substantial variation in Δ log(F0) as a function of log(F0).

**Level measurements**

The spectral energy distribution was analysed by measuring the signal levels in 10 overlapping frequency bands, each with a width of 3 Bark.

Figure 3 shows the results. Averaging over all vowels appeared justified since there was no consistent vowel specific peculiarity in the level differences. There was, however, a small difference between men and women. An analysis of the level differences obtained by Kallail & Emanuel for male speakers (1984 b) gave a similar result.

**Acknowledgements**

This research has been supported, in part, by a grant from HSFR, within the frame of the Swedish language technology programme. We are grateful to Johan Stark for the realisation of the response collection program.

**References**

Fig. 2. Whole vowel average increase in $F_1$, $F_2$, and $F_3$ in whispering as compared with phonating, plotted against positions in the phonated versions. Co-ordinates scaled logarithmically. Formant number indicated by figures, larger sized: male, smaller sized: female productions.

Fig. 3. Spectral power level density per frequency unit (dB/Hz), in relation to an arbitrary reference level, as measured for frequency bands with a width of 3 critical bands, for phonated and whispered versions of the same vowel types produced by male and female speakers, shown as a function of the logarithm of the centre frequency of each band. Continuous: phonated; dashed: whispered; thick: male; thin: female.


