Most of the phonetic knowledge that is currently available on spoken Finnish is based on clearly pronounced speech: either read-aloud text or other formal speaking styles, such as newscasts or interviews on the radio or on TV. However, linguistic differences are known to exist between written and spoken Finnish. It is therefore important to find out what kind of phonetic differences there are between informal, spontaneous Finnish speech and reading aloud a written text. In this chapter, some segmental properties of spontaneous Finnish dialogues and read-aloud Finnish speech will be described and compared.

Material

Speakers

For the analysis of spontaneous Finnish, informal Finnish dialogues were recorded from ten young adults (aged 20-30 years, five females), referred to as Group 1 (G1), and ten middle-aged adults (aged 45-55 years, five females), referred to as Group 2 (G2). All speakers were monolingual with Finnish as their mother tongue. These speakers had lived in the capital city area of Finland (Helsinki, Espoo, Vantaa) for most of their lives, and all of them were either university students or university graduates.

This study will focus on the data for G1. However, the transcripts of the dialogue speech for G2 were used to support the analysis of the lexical and phonemic distributions in the spontaneous speaking style.

Recordings of conversational speech

The recordings of the younger speakers of G1 were performed in an anechoic room, and the middle-aged speakers of G2 were recorded in a sound-treated professional recording studio. The speakers participating in each dialogue knew each other well, and they were allowed to chat freely and unmonitored for 40 to 60 minutes. Each speaker's voice was recorded with a head-mounted microphone (AKG HSC 200 SR) and at a sample rate of 44.1 kHz on a separate
channel of either a Tascam DA-P1 DAT recorder (G1) or directly to the hard disc of a computer system (G2) running ProTools. From the DAT, the dialogues were transferred to a computer. Next, the two channels of each stereo signal were separated, resulting in one high-quality speech signal per speaker. Thus, the two audio files created for each dialogue were of equal duration and remained time-synchronous. In addition, all sound files were finally resampled to the rate of 22.05 kHz. Only minor crosstalk was found in each signal file, and almost all of the speech material was technically appropriate for acoustic analyses.

Recordings of read-aloud speech

At least one week after the first recording session of each dialogue, the speakers individually participated in another session where they were asked to read aloud the quasi-orthographic transcript of their own speech in the dialogue. Since the syntactic, lexical and morphological properties of spoken and written Finnish are somewhat different, the speakers did not find this task to be either easy or trivial. Whereas the transcript did not contain any punctuation, breaks were inserted at utterance boundaries where the speaker had originally stopped speaking for some reason. Owing to these breaks, the speakers were instructed to pause for breathing at line breaks and preferably not to pause in the middle of utterances.

In addition, a shorter version of the text was produced for reading. In this text, only a small number of the original utterances were represented, and the transcripts were edited in order to adhere to the written language syntax and punctuation. However, the editing process was found to require significant changes in both lexical choices and word order, and the modified text was not considered to be sufficiently comparable to the content in the original unscripted speech. Consequently, the present analysis does not include the recordings for reading aloud the modified texts.

The recording quality affects the kinds of acoustic and auditory analyses that can be reliably applied on a speech corpus. More importantly, however, the usefulness of any speech corpus is defined by the kinds of annotation that are available for it.

Annotation

Annotation refers to the attaching of symbolic descriptions to certain intervals, parts, or points of a text or a signal. This means that annotations can be used as landmarks or search keys for both the
researcher and the automatic analysis tools. Since rich and systematic annotation can significantly reduce the researcher’s need to manually browse through the data during analysis, it is an essential phase in preparing a speech corpus. Unfortunately, most of the annotation work must be performed manually, which is very tedious and time-consuming. In the present study, the Praat program (Boersma & Weenink, 2006) was used for both the annotation work and the acoustic analysis of the speech material.

All the utterances of each individual speaker (G1 and G2) were first transliterated by following the Finnish orthographic conventions. Utterance boundaries were marked both in the dialogues and in the read-aloud material, and the corresponding orthographic transcripts were used as labels for the utterance intervals.

For five female (F1-F4 and F6) and five male speakers (M1-M4 and M6) in G1, fragments of the recorded material were more richly annotated by using several layers of information: the boundaries for phones, syllables, words and other units were marked.

The manual segmentation and labeling of phones (i.e., individual speech sounds) was the most time-consuming part of the annotation process. In order to provide a starting point for this work, a preliminary segmentation was created with an automatic segmentation tool for most of the sound files. The human labelers were then instructed to insert or edit the phone segment boundaries using both auditory and acoustic criteria.

The number of phones was allowed to be different from the number of phonemes that could be expected on the basis of the orthographic transcript. Each phone segment was labeled with the phonetic symbol that best corresponded to the labeler’s perceptual judgment of segmental quality, allowing diacritic marks. Nevertheless, the phonetic transcriptions were to be selected by listening to segments as part of their context, and not by listening to isolated segments, since this was known to be perceptually unreliable.

The boundaries of each vowel phone were marked so that the exact quality of the adjacent consonants could no longer be perceived. Thus, the most prominent transition phases were included in the appropriate consonant segment. Diphthongs were annotated as two separate vowel segments. However, in the

---

1 A tentative phoneme sequence was required as input for the automatic segmentation tool. Fortunately, since Finnish orthography corresponds rather well to phonemic structure, such an initial phoneme sequence was easily obtained for each sound file by applying simple transformation rules on the orthographic transcript.
automatic analysis, those vowel sequences that occurred within the boundaries of the same syllable were treated as one diphthong segment.

Since ASCII encoded symbols were technically easier to process than were the character sequences mapped to the IPA symbols that can be displayed within Praat, the phonetic transcriptions were entered using the Worldbet alphabet (an ASCII version of the International Phonetic Alphabet; see Hieronymus, 1993).

The manual phonetic segmentation and labeling was only completed and checked for a small part of the material, i.e., for a net speaking time of 1 to 3 minutes per speaker and per speaking style (spontaneous speech vs. reading aloud).

Word and syllable level segmentations were first generated automatically on the basis of the utterance transcripts. The word and syllable boundaries were manually corrected for all of the material for G1. Their boundaries were also aligned with the corresponding phone segmentation that had been manually checked. Figure 1 shows how the geminate consonants that always cross a syllable boundary were segmented. In this special condition,

![Spectrogram example](image)

**Figure 1.** Annotation example of a long stop consonant /k:/ in the word *luokka* 'class' with two syllables. For practical reasons, the syllable boundary was marked roughly halfway the phone segment [k]. In reality, no phonetic boundary occurs at this location.
the syllable boundaries were placed in the middle of the corresponding phone segment. During the automatic analysis, it was then possible to treat the geminate consonant as either two segments belonging to different syllables, or as a single segment extending over a syllable boundary. This decision was part of a more general aim to create annotation that could be used and analysed in as many different theoretical frameworks and angles as possible.

The boundaries for intonation units or other prosodic entities were not annotated, since this was found to be too time-consuming. Moreover, the boundaries of intonation units are known to be both subjective and difficult to determine (on the prosodic phrasing of Finnish, see, e.g., Aho & Yli-Luukko, 2005). Thus, their validation would have required many annotators or controlled perceptual experiments.

All annotations are somewhat subjective. In continuous speech, no clear-cut boundaries exist for any of the units that were annotated for the present study. Moreover, different researchers will always more or less disagree on their selection of a classification or a transcript for the units that are being investigated. This should be borne in mind when interpreting the results reported in the following sections.

**Analysing the corpus**

With Praat, an annotated speech corpus can be automatically analysed using scripts. In Praat scripts, it is possible to automatically search all the corpus files and to query, e.g., the label and the start and end points of a given interval along with the corresponding information for an interval that occurs simultaneously in another tier. The corresponding sound waveform can be accessed and analysed at a given time or within a specific temporal region. Additional calculations can then be performed on the temporal, acoustic and symbolic information that is extracted. In this way, Praat scripts can be designed to automatically collect and save different kinds of information associated with, e.g., all the individual phones in the corpus. The data can then be further processed by using any statistical analysis tool.

In the present study, Praat scripts were used to produce several large tables that contained information for, e.g., phone segments, phonemes, syllables, and words in the whole speech corpus. Statistical analyses were then performed either directly by using the
Praat program or by using the R statistical programming environment (R Development Core Team, 2004).

It was decided that a separate phonemic annotation tier would not be created, since there are many theoretical and practical problems in defining the boundaries for phonemes in a speech signal. Such problems arise primarily because phonemes are abstract linguistic units and they do not necessarily have direct counterparts in speech.

Moreover, since the Finnish orthography has a nearly one-to-one correspondence with its phonemic structure, it was possible to automatically derive a (quasi-)phonemic representation from the orthographic transcripts for Finnish syllables, and this representation could then be automatically mapped to units in the phone tier. This mapping was performed by first dividing the orthographic transcript for each syllable into structural parts: the nucleus, consisting of the vowel phonemes (either a short vowel, a long vowel when a double character was found, or a diphthong if two different vowel characters were present), the onset, consisting of zero or more consonant phonemes preceding the vocalic nucleus, and the coda, consisting of zero or more consonant phonemes after the nucleus. As the boundaries for each syllable interval were time-aligned with the corresponding phone intervals, it was relatively straightforward to automatically associate the nucleus, onset and coda with their corresponding phone segment(s): all consecutive vowel phones within the syllable boundaries were considered to represent the vocalic nucleus (i.e., the vowel “phoneme”), any consonants preceding it were mapped to the onset consonants, etc. When a syllable transcription ended in the same consonant as the next syllable started with, these symbols were considered as one long consonant phoneme.

In the present study, the term phoneme thus refers to the structural units derived from the orthographic transcription of a syllable. The acoustic counterpart for each phoneme is the phone segment (or the sequence of contiguous phone segments) that fits in the same structural part of the syllable, considering all the phone segments that occur within the boundaries of the same syllable.

The total numbers of phonemes analysed for each speaker are shown in table 1.
Table 1. Number of phonemes analysed for ten Finnish speakers in spontaneous and read-aloud speech

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F6</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spont. Vowel</td>
<td>1351</td>
<td>414</td>
<td>747</td>
<td>382</td>
<td>853</td>
<td>611</td>
<td>476</td>
<td>690</td>
<td>560</td>
<td>292</td>
</tr>
<tr>
<td>Cons.</td>
<td>1614</td>
<td>509</td>
<td>936</td>
<td>462</td>
<td>1055</td>
<td>752</td>
<td>577</td>
<td>838</td>
<td>684</td>
<td>343</td>
</tr>
<tr>
<td>Total</td>
<td>2965</td>
<td>923</td>
<td>1683</td>
<td>844</td>
<td>1908</td>
<td>1363</td>
<td>1053</td>
<td>1528</td>
<td>1244</td>
<td>635</td>
</tr>
<tr>
<td>Read- aloud Vowel</td>
<td>1446</td>
<td>509</td>
<td>2030</td>
<td>407</td>
<td>380</td>
<td>565</td>
<td>462</td>
<td>474</td>
<td>396</td>
<td>331</td>
</tr>
<tr>
<td>Cons.</td>
<td>1781</td>
<td>636</td>
<td>2475</td>
<td>488</td>
<td>468</td>
<td>683</td>
<td>547</td>
<td>593</td>
<td>476</td>
<td>391</td>
</tr>
<tr>
<td>Total</td>
<td>3227</td>
<td>1145</td>
<td>4505</td>
<td>895</td>
<td>848</td>
<td>1248</td>
<td>1009</td>
<td>1067</td>
<td>872</td>
<td>722</td>
</tr>
</tbody>
</table>

Word frequencies

In written text or speech, all words and word forms are not equally probable. The most frequent word forms in all languages usually represent function words, e.g., particles, pronouns, and auxiliary verbs. The most common words are also generally shorter than rare words. The most frequent Finnish word forms are mono- or bisyllabic, but some forms may have many more syllables, owing to the rich inflection system and to the extensive use of compound words in Finnish.

In order to determine how words are distributed in casual conversational Finnish, a frequency dictionary of 7651 word forms was created from a total of 41534 word tokens in the five dialogues. In the present study, a word token will be used to refer to an individual occurrence or instance of any word in speech, whereas a word form will refer to the group of word tokens having identical quasi-orthographic transcripts. However, it is to be noted that two word tokens that have an identical orthographic form may be ambiguous, i.e., they may represent different words and meanings.

The dictionary obtained from the dialogue corpus is naturally far too small to represent all the lexical properties of spoken Finnish. For instance, most of the statistical language models that are currently used in language technology are based on at least 10 million words of running text. Even though the small frequency dictionary described here is not lexically representative, it can be used as a tentative measure of word frequency.

Figure 2 shows the distribution of the word form frequencies in all the dialogue transcripts. One can observe that only a few very frequent word forms cover most of the word tokens in the material, whereas a great number of extremely rare words occur only once. This kind of distribution can be described with a nearly logarithmic function, which is well known as Zipf’s law (Manning & Schütze, 1999).
Figure 2. Distribution of the frequencies of orthographically different word forms in five Finnish spontaneous dialogues. The word form frequencies were plotted in logarithmic scale.

The high frequency of occurrence for a word form tends to increase the probability of segmental reduction within the tokens of that particular word in speech (e.g., Fidelholz, 1975; Hooper, 1976). Word frequency has been found to partly correlate with segmental durations as well as with the acoustic reduction of vowels measured as distributional features of the formant frequencies (e.g., van Son, Bolotova, Lennes & Pols, 2004).

**Phoneme frequencies**

Since Finnish has a nearly one-to-one correspondence between graphemes and phonemes, it was possible to roughly calculate the distribution of phonemes in the spontaneous material for G1 on the basis of the written transcripts. The upper panel in figure 3 shows the densities for the different phonemes in “running speech”, i.e., in all word tokens in the five dialogues. The lower panel displays the distribution as calculated from orthographically unique words in
Figure 3. Phoneme distributions in five spontaneous Finnish dialogues calculated as densities from running transcripts (the upper figure) vs. a “dictionary” consisting of all the orthographically unique words in the transcripts (the lower figure). The character N refers to the velar nasal consonant /ŋ/, and the letters ä and ö refer to the front vowels /æ/ and /ø/. Long vowels and consonants are separately indicated using semicolons (;). The horizontal line refers to the relative frequency level of 5%.
the transcripts. Since some words occur much more often than others in running text or speech, the phoneme distributions are also slightly different from the “dictionary” counts, which tend to exaggerate the probability for certain phonemes (e.g., /a/, /l/, /r/, /ö/) and underestimate others (e.g., /s/, /n/, /o/, /i:/).

**Phone durations**

All the different consonants and vowels of the Finnish language may occur phonologically as either long or short (cf. Iivonen, this volume). In writing, this abstract **length** contrast is usually indicated by single and double characters, respectively. However, this length distinction is not directly reflected in the measurable **durations** of the phone segments in spoken language. In clearly pronounced speech, long vowels and consonants have been found to be approximately twice as long in duration than their short counterparts (e.g., Wiik, 1965; Lehtonen, 1970; Kukkonen, 1990; for more information, see Iivonen, this volume). The long/short duration ratio is known to be smaller in fast speech. However, speech rhythm, rate, accentuation and syllable structure do have complex effects on segmental durations, and thus the duration difference between long and short phonemes is not absolute.

For the present corpus of Finnish, the durations of 29684 phone segments were analysed: 14146 segments for spontaneous speech and 15538 segments for read-aloud speech. The utterance-initial stop consonants, phones in utterance-final syllables and phones within compound words were all excluded from the duration analysis.

The phones corresponding to the long and short phonemes were analysed as separate groups. The mean and median durations and the corresponding long/short ratios are shown in table 1. The segmental durations for both long and short phonemes are noticeably smaller than those reported for clear laboratory speech in earlier studies (for references, see livonen, in this volume).

As seen in table 1, long consonants and vowels tend to be longer in duration than short consonants and vowels. This duration ratio between long and short phonemes is slightly smaller in spontaneous speech than in reading aloud. Moreover, the duration distributions shown in figure 4 confirm that the phone durations are greater for reading aloud than in the spontaneous speaking style.

The distributions of vowel durations are shown in figure 5. In general, these durations are smaller than those obtained from read-
Table 1. Durations of the phonetic counterparts for the long and short phonemes in spontaneous (N=14146) and read-aloud (N=15538) Finnish speech

<table>
<thead>
<tr>
<th>Speaking style</th>
<th>Long (ms)</th>
<th>Short (ms)</th>
<th>Ratio long/short</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vowels</strong></td>
<td>mean</td>
<td>median</td>
<td>stdev</td>
</tr>
<tr>
<td>Spontaneous speech</td>
<td>87.2</td>
<td>76.9</td>
<td>52.3</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
<td>stdev</td>
</tr>
<tr>
<td>Read-aloud speech</td>
<td>94.3</td>
<td>86.5</td>
<td>25.7</td>
</tr>
<tr>
<td><strong>Consonants</strong></td>
<td>mean</td>
<td>median</td>
<td>stdev</td>
</tr>
<tr>
<td>Spontaneous speech</td>
<td>110.2</td>
<td>109.2</td>
<td>41.5</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>median</td>
<td>stdev</td>
</tr>
<tr>
<td>Read-aloud speech</td>
<td>120.5</td>
<td>115.4</td>
<td>53.9</td>
</tr>
</tbody>
</table>

aloud speech in previous studies (for a review, see livonen, this volume). Nevertheless, the relative durations of the different vowels are rather well in accord with the results obtained by Lehtonen (1970) and Wiik (1965). The rare vowels /ö/ and /ö:/ do stand out, which may be due to the small number of tokens within the material.

Voiced consonants have been reported to be shorter than voiceless consonants in Finnish (e.g., Lehtonen, 1970). In addition, the mean duration in spontaneous speech was only 60 ms for the short voiced consonants and 82 for the long voiced, whereas the means for unvoiced consonants were 71 ms and 133 ms, respectively.

Very small negative correlations were found between the relative word frequency and the duration of segments in the word-initial syllable. The Spearman’s rho for short vowels in word-initial syllables was -0.33 and for short consonants -0.15.

In many previous studies of read-aloud Finnish, it has been demonstrated that the duration of segments tends to decrease along with an increasing number of segments or syllables in the utterance. In spontaneous speech within the present material, a small negative correlation between the segmental duration and number of syllables in the word was observed only for those long consonants on the
Figure 4. Distribution of phone segment durations in Finnish informal conversational speech (N=14146) and in reading aloud (N=15538) for long and short phonemes and diphthongs. The durations for some very long segments are not shown in this figure.

Figure 5. Duration distributions for short and long vowel phonemes in word-initial vs. non-initial syllables in spontaneous Finnish speech (N=5663, diphthongs excluded).
boundary of first and second syllable in the word \( \text{rho}=-0.10 \). Instead, short consonants \( \text{rho}=0.14 \) and long vowels \( \text{rho}=0.16 \) within word-initial syllables exhibited a small positive correlation. As for the number of words in an utterance, similar negative but extremely small correlations were only found for vowels. For other segments, in the word-initial syllables or elsewhere, no correlations of this kind were observed.

According to the current interpretation, each Finnish word has (abstract) lexical stress on the first syllable. However, not every word carries accent in spoken utterances (Iivonen, this volume, pp. 57-58). In the present material, the phone duration correlates very slightly with this word stress: for short phones, the correlation coefficient is 0.25 for spontaneous and 0.27 for read speech. For long phonemes, the correlation is even smaller, the coefficients being 0.11 for spontaneous dialogue and 0.18 for reading aloud. It is to be noted that this calculation did not take the secondary stressed syllables into account, which may distort the result.

The phonetic correlates for sentence accent or for the perceived prominence within an utterance are usually found within the initial syllable of the prominent word. Thus, the domain where sentence accent can be realized overlaps with the domain for word-internal stress. On the other hand, a word may also be completely unaccented. In such a case, there may not be phonetic markers for the word-internal stress pattern. Another essential point is that the sentence accent was not systematically annotated in the present corpus, and as a consequence, it was not possible to separately study the relationship between sentence accent and segmental duration. However, accent is known to be associated with slightly longer segmental durations in comparison to unaccented positions in Finnish (Laurosela, 1922; Suomi et al., 2003).

**Allophonic variability**

Spoken Finnish is usually not written down. On the other hand, since the orthographic system is rather loyal to phonemic structure, it is possible to use an orthographic transcript for creating a tentative phonemic transcript of Finnish speech. Thus, even when using the Finnish version of the Latin alphabet, any Finnish transcriber needs to choose between the various degrees of standard orthography and the more phonetic and impressionistic representations of what was said. Furthermore, the use of the word forms and the preferred sentence structure often differ between standard written Finnish and casual spoken Finnish. Therefore, the
“phonemic” representations derived from the transcripts must be treated with caution, since they result from the transcriber’s subjective interpretations.

For the aforementioned reason, it was considered impossible to study, e.g., segmental elisions or insertions. For instance, a transcriber may sometimes have chosen to transcribe a word form with a final \( n \) when he/she has heard a final [\( n \)] (e.g., sen, the genitive form of se ‘it’). In other cases, however, the transcriber may have written the same word form without the final \( n \). In speech, the produced form se may still represent a genitive form in that particular context. In short, independently of the quasi-orthographic transcript, a phone segment [\( n \)] may or may not have been labeled in the phone tier in each case.

Moreover, it was found that the use of different phonetic symbols varied individually among the labelers. As a result, it was not possible to make valid comparisons on the allophonic variability across the spontaneous and read-aloud speaking styles. However, the number of different transcriptions was usually greater for the vowel phonemes: each of the short vowel phonemes was phonetically represented by approximately 10 to 20 different phonetic symbols or symbol combinations. In many cases, the diacritic for vowel centralisation had also been used.

The most frequent consonant phonemes /\( s \)/, /\( t \)/, /\( k \)/ and /\( n \)/ were assigned many different transcription variants. The phoneme /\( s \)/ was very often at least partly voiced intervocally (an example is shown in figure 6), and it also tended to be undergo degrees of place assimilation with the neighbouring vowels, ending up as various coronal fricatives (cf. figure 7). The short /\( t \)/ and /\( k \)/ phonemes were sometimes produced as the corresponding homorganic fricatives or approximants. For instance, /\( k \)/ could be produced as [\( x \)], [\( y \)] or even [\( u \)], but also as a voiced [\( g \)], or even as a stop at a different place of articulation, e.g., [\( q \)]. The Finnish /\( t \)/ is most often produced as a dental or prealveolar [\( t \)], but in some cases, it may be produced as an alveolar approximant.

The short /\( n \)/ is usually assimilated to the place of articulation of the following consonant. In addition to the alveolar [\( n \)], the variants [\( m \)], [\( n \)] and even the labiodental [\( m \)] were found to occur. As sporadic cases, [\( l \)], [\( d \)], [\( h \)], some unclear central vowels, and a glottal stop [\( ? \)] were also discovered among the phonetic transcriptions for /\( n \)/. However, some of these cases were
Figure 6. Example of the phoneme /s/ in the word *se* produced as a voiced [z] by the female speaker F1 in spontaneous conversation.

Figure 7. Examples of [s] produced between two /u/ vowels (left, the word *ruusu* ‘rose’) vs. between two /i/ vowels (right, in the word *pakollisi* ‘obligatory’, partitive plural) by the female speaker F2 in spontaneous conversation. Note the slightly different spectral properties of the [s] that are due to coarticulatory effects in the two contexts.
somewhat unclear. The phoneme /n/ is sometimes dropped word-finally, or it may only occur as nasalization on the preceding vowel.

The phonetic counterparts of the long consonant phonemes were labeled almost invariably with the basic or “expected” phonetic symbol. Similarly, long vowel phonemes had fewer transcription variants than the short vowels. Although phonetic transcriptions are not a fully reliable source for qualitative information, the smaller number of transcription variants may indicate that the articulatory and acoustic qualities of the phonetic counterparts of long phonemes tend to vary less across the different contexts than the qualities of short phonemes.

**Vowel quality**

The Finnish language uses eight different vowel qualities /a e i o u y æ ø/ for marking phonological contrasts (see Iivonen, this volume). All of these vowels may occur phonologically as either long or short (or, depending on the phonological interpretation, as ‘single’ or ‘double’), or they may be combined into diphthongs. However, there are phonotactic restrictions on the occurrence of the different
vowels (e.g., vowel harmony), and not all vowel types are equally common in all positions. Due to the lexical differences between the written standard and spoken Finnish, the distribution of the vowel types may also be slightly different in these two language variants.

Finnish vowel quality in continuous speech is affected by phonological length along with many other factors. The vowel segments occurring in the accented positions of utterances or in the stressed syllables of words tend to be pronounced with more articulatory effort or precision, i.e., they are phonetically less “reduced” than unaccented vowels. Vowel reduction generally refers to the loss of a vowel’s characteristic quality with respect to an ideal reference pronunciation. Here, vowel reduction is considered as an acoustic term, i.e., more “reduced” vowels would be more affected by coarticulation and they would thus exhibit more acoustic variability (cf. van Bergem, 1995).

Perhaps the most common method for visualizing acoustic vowel quality is the F1/F2 chart, displaying the values of the two lowest formants for each particular vowel. Since the estimated centre frequencies of the first two formants have an indirect relationship with the corresponding tongue height and vowel frontness, an F1/F2 chart can be very illuminating. Nevertheless, automatically estimated formant frequencies must be interpreted with some caution.

Formant analysis

The frequencies of the two lowest formants (F1 and F2) were calculated at the temporal midpoint of each vowel nucleus using the Praat program (Burg algorithm; parameters were adjusted for male and female speakers accordingly). Those vowels that did not yield acceptable values for both F1 and F2 were discarded.

The average formant values for vowels in the word-initial syllables in spontaneous and read-aloud speech are shown in figure 9 for female speakers and in figure 10 for male speakers. The overall mean values for each speaker are indicated with black dots. The data for the vocalic nuclei containing diphthongs were excluded from these figures, since during diphthongs, the formant frequencies constantly glide, and the formant frequencies at their temporal midpoints would not be comparable to those measured from other vowel segments. Moreover, some of the charts contain fewer than eight vowels due to insufficient or missing data. Nevertheless, general observations can be made.
Figure 9. Formant charts of short and long vowels within word-initial syllables in spontaneous and read-aloud speech for five female speakers of Finnish. The formant frequencies were measured at the temporal midpoints of the segmented vowels. Each letter indicates the mean F1 and F2 frequencies on the Bark scale for the corresponding vowel category. The black dots indicate the overall mean formant frequencies for each speaker.
Figure 10. Formant charts of short and long vowels within word-initial syllables in spontaneous and read-aloud speech for five male speakers of Finnish. The formant frequencies were measured at the temporal midpoints of the segmented vowels. Each letter indicates the mean F1 and F2 frequencies on the Bark scale for the corresponding vowel category. The black dots indicate the overall mean formant frequencies for each speaker.
Vowel reduction is often associated with a “centralisation” effect on the F1/F2 formant chart. In figures 9 and 10, there is indeed a slight but visible tendency for the mean formant values to be more centralised on the F1/F2 charts for spontaneous speech. Furthermore, in both spontaneous and read-aloud speech, short vowels tend to be more centralised. This centralisation does not, however, necessarily concern the speaker's articulatory target but is caused by the averaging over vowel measurements: the more variable the formant values are, the closer their mean value is to the centre of the chart (see, e.g., van Bergem, 1995). It may thus be concluded that there is more variability in vowel quality, i.e., vowels are produced “less clearly” during spontaneous speech than during reading aloud.

**Segmental pitch**

In order to compare the pitch range the speakers used in their spontaneous conversation and read-aloud speech, the standard pitch algorithm available in Praat was used to measure the pitch at the temporal mid points of all vowel segments for all ten speakers in G1. A preliminary pitch analysis was first performed in order to define the suitable maximum and minimum pitch parameters for each speaker. For all speakers, a separate cluster of low pitch points was observed. This cluster was found to be associated with creaky voice, which is quite typical for Finnish speakers, especially in the final portions of their utterances. As a consequence, the minimum pitch analysis parameter was selected to exclude the low pitch values for creak.

The mean and median pitch values in semitones for each speaker and for the two speaking styles are shown in table 2. Due to the positively skewed distribution of pitch values, it is to be noted that the mean pitch values are less robust than the median in describing a speaker’s overall pitch level. The female speakers had a median pitch at approximately 9-12 semitones above 100 Hz, corresponding to a frequency of 170 to 200 Hz. Three male speakers (M1-M3) had a median pitch of about 0 semitones (100 Hz) and M6 around 2 ST. Speaker M4 had a slightly higher voice than the other males (median ca. 6-7 ST).

Since the minimum and maximum pitch parameters were determined manually, and since the highest pitch values may contain errors (referred to as “octave jumps”), the minimum and maximum pitch values would not be good indicators of the pitch
Table 2. Mean and median pitch as measured from the temporal midpoints of vowel segments for ten Finnish speakers in spontaneous conversation and read-aloud speech. Values are reported in semitones (0 ST = 100 Hz).

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F4</th>
<th>F6</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
<th>M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spont</td>
<td>Mean</td>
<td>10.7</td>
<td>9.6</td>
<td>11.2</td>
<td>13.2</td>
<td>11.0</td>
<td>0.6</td>
<td>1.7</td>
<td>1.5</td>
<td>6.3</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>10.3</td>
<td>9.1</td>
<td>10.7</td>
<td>12.6</td>
<td>10.5</td>
<td>-0.0</td>
<td>1.2</td>
<td>0.6</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>Stdev</td>
<td>2.6</td>
<td>2.4</td>
<td>2.2</td>
<td>3.6</td>
<td>3.3</td>
<td>3.2</td>
<td>3.0</td>
<td>3.4</td>
<td>2.9</td>
</tr>
<tr>
<td>Read</td>
<td>Mean</td>
<td>9.6</td>
<td>9.8</td>
<td>11.2</td>
<td>12.4</td>
<td>10.5</td>
<td>0.0</td>
<td>0.4</td>
<td>0.9</td>
<td>7.0</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>9.2</td>
<td>9.4</td>
<td>10.7</td>
<td>12.1</td>
<td>10.0</td>
<td>-0.2</td>
<td>0.3</td>
<td>0.7</td>
<td>6.9</td>
</tr>
<tr>
<td></td>
<td>Stdev</td>
<td>1.9</td>
<td>2.0</td>
<td>2.2</td>
<td>1.9</td>
<td>2.4</td>
<td>3.1</td>
<td>2.3</td>
<td>1.9</td>
<td>2.1</td>
</tr>
</tbody>
</table>

Figure 11. Distribution of pitch values measured at the temporal midpoints of vowel segments for ten speakers (five females) in spontaneous dialogues (white boxes) vs. read-aloud speech (grey boxes).

The overall distributions of the pitch values for each speaker are shown as a boxplot in figure 11. Indeed, it can be observed for all speakers that 75% of the pitch values in the read-aloud speech cover a narrower range around the median than the same proportion of pitch values in spontaneous speech, i.e., the speakers tend to vary less in their pitch when they read aloud.
Conclusions

The results from this study suggest that Finnish speakers use different phonetic properties in spontaneous conversation and reading aloud. The speakers’ segmental durations are typically longer when they read aloud, indicating that their general speech rate is also slower. The mean phone durations found in the spontaneous speech of the speakers in this study were much smaller than the ones measured in previous studies for clearly pronounced speech. Furthermore, the long/short ratio was also diminished for speakers in the present study. Concerning vowel quality, there is a tendency for the vowels in word-initial syllables to be acoustically less variable both for long vowels in comparison to short vowels and for read-aloud speech in comparison to spontaneous conversation. Seven out of the ten speakers had a lower median pitch when they read aloud than when they engaged in spontaneous dialogue. Speakers also tended to use a narrower pitch range when reading aloud.

Acknowledgments

This study was partly funded by the Academy of Finland (projects 53623 and 107606) and INTAS (project 00-00915). I am extremely grateful for the help extended by all the students and colleagues at the various stages of the annotation of the speech corpus.

References


